



THE NEW INSTITUTIONALISM IN STRATEGIC MANAGEMENT

Volume 19

Paul Ingram &
Brian S. Silverman

THE NEW INSTITUTIONALISM IN STRATEGIC MANAGEMENT

ADVANCES IN STRATEGIC MANAGEMENT

Series Editor: Joel A. C. Baum

- Volume 15: Disciplinary Roots of Strategic Management
 Research
- Volume 16: Population-Level Learning and Industry Change
- Volume 17: Economics Meets Sociology in Strategic
 Management
- Volume 18: Multiunit Organization and Multimarket
 Strategy

ADVANCES IN STRATEGIC MANAGEMENT VOLUME 19

THE NEW INSTITUTIONALISM IN STRATEGIC MANAGEMENT

EDITED BY

PAUL INGRAM

Columbia Business School, Columbia University, USA

BRIAN S. SILVERMAN

Rotman School of Management, University of Toronto, Canada

2002



JAI

An Imprint of Elsevier Science

Amsterdam – Boston – London – New York – Oxford – Paris
San Diego – San Francisco – Singapore – Sydney – Tokyo

ELSEVIER SCIENCE Ltd
The Boulevard, Langford Lane
Kidlington, Oxford OX5 1GB, UK

© 2002 Elsevier Science Ltd. All rights reserved.

This work is protected under copyright by Elsevier Science, and the following terms and conditions apply to its use:

Photocopying

Single photocopies of single chapters may be made for personal use as allowed by national copyright laws. Permission of the Publisher and payment of a fee is required for all other photocopying, including multiple or systematic copying, copying for advertising or promotional purposes, resale, and all forms of document delivery. Special rates are available for educational institutions that wish to make photocopies for non-profit educational classroom use.

Permissions may be sought directly from Elsevier Science Global Rights Department, PO Box 800, Oxford OX5 1DX, UK; phone: (+44) 1865 843830, fax: (+44) 1865 853333, e-mail: permissions@elsevier.co.uk. You may also contact Global Rights directly through Elsevier's home page (<http://www.elsevier.com>), by selecting 'Obtaining Permissions'.

In the USA, users may clear permissions and make payments through the Copyright Clearance Center, Inc., 222 Rosewood Drive, Danvers, MA 01923, USA; phone: (+1) (978) 7508400, fax: (+1) (978) 7504744, and in the UK through the Copyright Licensing Agency Rapid Clearance Service (CLARCS), 90 Tottenham Court Road, London W1P 0LP, UK; phone: (+44) 207 631 5555; fax: (+44) 207 631 5500. Other countries may have a local reprographic rights agency for payments.

Derivative Works

Tables of contents may be reproduced for internal circulation, but permission of Elsevier Science is required for external resale or distribution of such material.

Permission of the Publisher is required for all other derivative works, including compilations and translations.

Electronic Storage or Usage

Permission of the Publisher is required to store or use electronically any material contained in this work, including any chapter or part of a chapter.

Except as outlined above, no part of this work may be reproduced, stored in a retrieval system or transmitted in any form or by any means, electronic, mechanical, photocopying, recording or otherwise, without prior written permission of the Publisher.

Address permissions requests to: Elsevier Science Global Rights Department, at the mail, fax and e-mail addresses noted above.

Notice

No responsibility is assumed by the Publisher for any injury and/or damage to persons or property as a matter of products liability, negligence or otherwise, or from any use or operation of any methods, products, instructions or ideas contained in the material herein. Because of rapid advances in the medical sciences, in particular, independent verification of diagnoses and drug dosages should be made.

First edition 2002

Library of Congress Cataloging in Publication Data

A catalog record from the Library of Congress has been applied for.

British Library Cataloguing in Publication Data

A catalogue record from the British Library has been applied for.

ISBN: 0-7623-0903-2

ISSN: 0742-3322 (Series)

☺ The paper used in this publication meets the requirements of ANSI/NISO Z39.48-1992 (Permanence of Paper).

Printed in The Netherlands.

CONTENTS

LIST OF CONTRIBUTORS	vii

INTRODUCTION: THE NEW INSTITUTIONALISM IN STRATEGIC MANAGEMENT <i>Paul Ingram and Brian S. Silverman</i>	1

PART 1: HOW DO FIRMS BEHAVE?

POLICY AND PROCESS: A GAME-THEORETIC FRAMEWORK FOR THE DESIGN OF NON-MARKET STRATEGY <i>Guy L. F. Holburn and Richard G. Vanden Bergh</i>	33

MANAGERIAL DECISION MAKING IN NON-MARKET ENVIRONMENTS: A SURVEY EXPERIMENT <i>John de Figueiredo and Rui J. P. de Figueiredo, Jr.</i>	67

PRETTY PICTURES AND UGLY SCENES: POLITICAL AND TECHNOLOGICAL MANEUVERS IN HIGH DEFINITION TELEVISION <i>Glen Dowell, Anand Swaminathan and James Wade</i>	97

THE EVOLUTION OF UNIVERSITY PATENTING AND LICENSING PROCEDURES: AN EMPIRICAL STUDY OF INSTITUTIONAL CHANGE <i>Bhaven N. Sampat and Richard R. Nelson</i>	135

PART 2: WHY ARE FIRMS DIFFERENT?

COMPETITION, CONTINGENCY, AND THE EXTERNAL STRUCTURE OF MARKETS <i>Ronald S. Burt, Miguel Guilarte, Holly J. Raider and Yuki Yasuda</i>	167

INSTITUTIONAL CHANGE IN REAL-TIME: THE
DEVELOPMENT OF EMPLOYEE STOCK OPTIONS
FOR GERMAN VENTURE CAPITAL

Jonathan Jaffee and John Freeman

219

INSTITUTIONAL BARRIERS TO ELECTRONIC
COMMERCE: AN HISTORICAL PERSPECTIVE

Karen Clay and Robert P. Strauss

247

PART 3: WHAT LIMITS THE SCOPE OF THE FIRM?

INFORMAL AND FORMAL ORGANIZATION IN
NEW INSTITUTIONAL ECONOMICS

Todd R. Zenger, Sergio G. Lazzarini and Laura Poppo

277

‘TESTS TELL’: CONSTITUTIVE LEGITIMACY
AND CONSUMER ACCEPTANCE OF THE
AUTOMOBILE: 1895–1912

Hayagreeva Rao

307

**PART 4: WHAT DETERMINES SUCCESS AND
FAILURE IN INTERNATIONAL COMPETITION?**

LEARNING ABOUT THE INSTITUTIONAL
ENVIRONMENT

Witold J. Henisz and Andrew Delios

339

INSTITUTIONS AND THE VICIOUS CIRCLE OF
DISTRUST IN THE RUSSIAN HOUSEHOLD
DEPOSIT MARKET: 1992–1999

Andrew Spicer and William Pyle

373

LIST OF CONTRIBUTORS

<i>Ronald S. Burt</i>	Graduate School of Business University of Chicago, USA
<i>Karen Clay</i>	Heinz School of Public Policy and Management Carnegie Mellon University, USA
<i>John M. de Figueiredo</i>	Sloan School of Management Massachusetts Institute of Technology, USA
<i>Rui J. P. de Figueiredo, Jr.</i>	Haas School of Business University of California, Berkeley, USA
<i>Andrew Delios</i>	Department of Business Policy National University of Singapore, Singapore
<i>Glen Dowell</i>	Mendoza College of Business University of Notre Dame, USA
<i>John Freeman</i>	Haas School of Business University of California, Berkeley, USA
<i>Miguel Guilarte</i>	Fielding Graduate Institute, USA
<i>Witold J. Henisz</i>	Department of Management The Wharton School, USA
<i>Guy L. F. Holburn</i>	Richard Ivey School of Business University of Western Ontario, Canada
<i>Paul Ingram</i>	Columbia Business School Columbia University, USA

<i>Jonathan Jaffee</i>	Graduate School of Industrial Administration Carnegie Mellon University, USA
<i>Sergio G. Lazzarini</i>	Olin School of Business Washington University, USA
<i>Richard R. Nelson</i>	School of International and Public Affairs Columbia University, USA
<i>William Pyle</i>	Department of Economics Middlebury College, USA
<i>Laura Poppo</i>	Pamplin School of Business Virginia Tech, USA
<i>Holly J. Raider</i>	INSEAD, France
<i>Hayagreeva Rao</i>	Goizueta Business School Emory University, USA
<i>Bhaven N. Sampat</i>	School of Public Policy Georgia Institute of Technology, USA
<i>Brian Silverman</i>	Rotman School of Management University of Toronto, Canada
<i>Andrew Spicer</i>	Anderson Graduate School of Management University of California, Riverside, USA
<i>Robert Strauss</i>	Heinz School of Public Policy and Management Carnegie Mellon University, USA
<i>Anand Swaminathan</i>	Graduate School of Management University of California, Davis, USA
<i>Richard G. Vanden Bergh</i>	School of Business Administration University of Vermont, USA

James Wade

School of Business
University of Wisconsin, USA

Yuki Yasuda

Rikkyo University, Japan

Todd R. Zenger

Olin School of Business
Washington University, USA

INTRODUCTION: THE NEW INSTITUTIONALISM IN STRATEGIC MANAGEMENT

Paul Ingram and Brian S. Silverman

INTRODUCTION

The recent collapse of the huge energy-trading company Enron has prompted a cry rarely heard in the American economy: to increase regulation. All the more rare, this call is coming from both consumers and employees, who fear the power of large corporations, as well as from the corporations themselves, who fear that an erosion in the trust and confidence of employees, investors, customers and suppliers will cripple their capacity to do business. The effect of the Enron shock is to remind us of something that strategists, managers, and designers of organizations frequently ignore – that the economy rests on an institutional bedrock. Particularly in the United States, where fundamental institutions have been so effective and so stable for so long, it is easy to forget that the state, along with the various organizations and social norms that promote trust and confidence in economic transactions, have a critical influence on which organizations and strategies will succeed.

This volume examines new-institutional theory, which takes as its explicit focus the influences that were hidden, or taken for granted, in the Enron debacle. The core claim of this theory is that actors pursue their interests within institutional constraints, such as the regulations that constrained (or were presumed to constrain) Enron. This idea is the basis of a growing, pan-disciplinary

literature that seeks to explain the conduct and performance of individuals, organizations, and states. As the foundational theory about the nature and operation of institutions has been established, and as evidence on the operation and inter-relationship of alternative institutional forms has grown, the tools available for constructing new institutional explanations have been established. The accumulated research has reached a critical mass that creates numerous theoretical and empirical opportunities.

We begin this introductory chapter by addressing the question “why now?” for the new institutionalism in strategy. We identify a number of economic developments and scholarly advances that have helped to make clear the significance of institutions for strategic management. We then explain just what we mean by institutions, with a quick survey of the most relevant literature. In this survey we give explicit attention to the fact that there are a number of variants of new-institutional theory, which tend to emphasize different types of institutions. Our response to this variety is to present a classification of institutional forms, and identify the literatures that have focused on each. Our own view is that a complete theory of institutions must be comprehensive in its definition of institutions, but also explicit about differences among institutional forms and interdependencies between them. In the third section of this introduction we identify a number of pressing questions for new-institutional theory. By applying the chapters of this volume to those questions, we show that research in strategic management can play an important role in the development of new-institutional theory, particularly by helping to explain how organizations affect institutions of other types. In the fourth and final section, we make the literature on strategic management the focus, and describe how new-institutional theory can help solve pressing questions in that literature.

WHY THE GROWING INTEREST IN THE NEW INSTITUTIONALISM IN STRATEGY?

Until recently, the idea of explicitly considering institutions to explain the content and effectiveness of organizational strategies would have seemed a little like building theories of strategy based on the fact that the human subjects of the organizations we study breathe air. Institutions, like air, probably make a difference, but why, given their constancy and pervasiveness, should we invest in understanding that difference? The naiveté of this position has been made clear by a number of recent developments in the international economy, and in the scholarly field of strategy.

Transition from State Socialism.

The most significant of these developments is the transition from state socialism among countries of the former Soviet-bloc and China. The theme of the somewhat crude analysis of the early days of this transition was that capitalist economies were more productive than state-socialist ones; the prescription was for the latter economies to adopt features of the former. The result of these early changes was a range of unintended consequences (Murphy, Shleifer & Vishny, 1992; Nee, 1992; Stark, 1996). These surprise outcomes of piecemeal changes indicated something under-appreciated about the economies in question – laws, organizations, and norms operated together in a complex fashion. Economies that were made to be similar on a subset of these might yet exhibit very different performance outcomes.

As Spicer and Pyle (this volume) show, attempts to create western-style markets in formerly state-socialist economies have been crippled by malfeasance on the part of some organizations, and the corresponding distrust that developed among the public. Simply, these problems arose because the designers of new institutions and organizations paid insufficient attention to the complex interdependence of the institutions that facilitate exchange in “free” markets. For example, Spicer and Pyle document the failure of the market for household savings in Russia. On the surface, that market looks something like markets in the United States – indeed some of the most prominent American financial organizations tried to extend their operations to Russia. However, important institutional constraints that in the U.S. are provided by the state (regulation), organizations (auditing and rating of investments) and civil society (public awareness of the operation and risks of financial markets) were missing in Russia. The patchwork institutional framework that resulted enabled some organizational strategies and frustrated others.

Internationalization of Business.

Another development that points to the strategic significance of institutions is the recent increase in international trade, as well as the multinational operations of specific organizations. To be fair, international business is the one area where there is a long intellectual history of considering the impact of institutions on strategy. For example, concerns surrounding differences in national cultures (Hofstede, 1980), and the risk of expropriation of capital by a host nation (Teece, 1981), enter into prescriptions for operating a business internationally. Even in this area, however, there is rapid growth in attention to institutions. The significance of the international context for highlighting the role of institutions is that cross-national comparisons highlight institutional differences that may be taken for granted within a country.

Henisz and Delios (this volume) consider how multinational organizations can learn, and exploit their knowledge of institutions in the countries in which they operate. This is an integration of institutional forms that have been considered separately. There is a substantial literature that measures the legal and social environment for doing business in a given country (e.g. Henisz, 2000). And multinational organization has been characterized as an institutional mechanism for overcoming business risks, or other institutional shortcomings that exist in some countries. The learning theory that Henisz and Delios present forges a strategic link between those two literatures. Their arguments suggest patterns of design and expansion that can help multinationals succeed across a range of institutional environments.

Technological Development.

The interdependence between institutions and technology has been prominent in the new institutionalism. For example, North (1993) claims that technological advance by organizations forms the impetus for changing institutions – as new technologies develop, new institutions are required to effectively exploit them. Sometimes, changing technology highlights the significance of institutions because it exposes gaps in the institutional structure. For example, advances in medical transplant technology have created a market for the exchange of human organs, and exposed the unpreparedness of the law, the medical profession, and even the value-system of our culture, to govern that market (Healy, 2002).

In other instances, technological development makes institutions salient by setting off an episode of institutional creation or change. This is illustrated by Dowell, Swaminathan and Wade (this volume). They examine the role of social movements to create institutions around the technology of high definition television (HDTV). The development of HDTV created an interest in changing institutions – for example the rights over the spectrum related to television broadcasts, and the standard that HDTV would follow in the U.S. Further, organizations that had previously been unsuccessful in their campaign to influence spectrum allocation were able to harness the interest associated with HDTV to affect institutional change in their favor. Dowell, Swaminathan and Wade's account of this campaign highlights the strategic use of cultural concepts (framing) by the champions of various HDTV schemes.

Compromise and Manipulation of Existing Institutions.

The attention generated by the Enron collapse is not the result of new technology, internationalization, or the shift of political regimes. Instead, it emerges because institutions which were assumed to be stable and reliable were undermined, or otherwise shown to be lacking. For example, the failure of

Enron's auditors, Arthur Andersen, to identify questionable financial reporting practices is a violation of the role that auditors are expected to play in the institutional framework of western capitalism. Audit firms are in the business of selling surety. They vouch for the compliance by the auditee to familiar and accepted accounting principles, and thus allow investors to more reliably value the company (Strange, 1996). The objectivity of the auditor is key to providing this surety, and many of the professional policies of accountants – which include rules against accepting gifts from clients, and rules prohibiting over-reliance by the auditor on the fees of any single client – are designed to maintain objectivity. The very necessity of these rules points to a weak spot in the institutional framework. If an auditor could be influenced by the auditee, and convinced or deceived into inappropriately validating improper accounting practices, then stakeholders of many types – banks, shareholders, employees, customers – might be convinced to over-invest in the auditee.

The point of this is not to criticize a particular corporation, or auditor, but to illustrate that sometimes institutions, even those that are as venerable as auditing, are malleable and at risk of influence by their subject organizations. This malleability creates strategic opportunities for organizations. And although collusion with or deception of an auditor is illegitimate (although potentially profitable), other forms of institutional influence are more acceptable. Specifically, there is a growing attention to the possibility that organizations may influence for good or ill the institutions provided by the state, such as laws and regulations (Baron, 2001; Murphy, Shleifer & Vishny, 1993). In this volume, Holburn and Vanden Bergh, as well as De Figueiredo and de Figueiredo, build on this idea, using strikingly different methodologies. Holburn and Vanden Bergh present a formal model that helps to identify where organizations should aim their lobbying efforts. De Figueiredo and de Figueiredo use an experimental methodology to examine the question “how good are strategists at recognizing the opportunity to influence the state?” The clear significance of these questions, and the diversity of the research methods employed, point to the great opportunities for the organizations that make institutions the focus of their strategies, and the scholars that study them.

Collapse of the Tyranny of the Here and Now.

The new institutionalism has always emphasized historical research. Classics in the field examine economic and organizational outcomes from the deep past (North & Weingast, 1989; Greif, 1994). The willingness to embrace history derives from two core precepts of new-institutional theory. The first is that core elements of the theory are timeless. The behavioral assumptions of the new institutionalism amount to bounded rationality. Institutions, as we will describe,

are simply the rules that constrain the interest-seeking behavior of actors. The manner in which institutions operate is basically the same, whether the institutions are the self-policing policies of 11th century traders on the Mediterranean, or 21st century tax laws in Munich. Therefore, old institutions are understood to be as valuable as new ones for understanding economic performance – perhaps even more valuable, because historical institutions can sometimes be studied with fuller information and more objectivity. The second element of new-institutional theory that leads to historical research is the path dependence of institutions. We don't have complete theories of institutional change, but one thing that seems certain is that the options for new institutions derive in a large way from pre-existing institutions (North, 1990). So, even forward looking analysts must understand old institutions, as they are the roots of future institutions.

The treatment of history in the new institutionalism stands in sharp contrast to the normal practice in research on business strategy. Strategy often suffers from a tyranny of the here and now, a desire to celebrate contemporary phenomena and slight historical ones. This ahistoricism is one reason why research in strategy struggles for social-scientific legitimacy. By reveling in current affairs, and de-emphasizing their underpinnings in the past, strategy scholarship often undermines its own claims to develop explanations that transcend their contemporary context. In other words, the field of strategy struggles to develop good theory, because it downplays temporal transitivity and generalizability.

Clay and Strauss (this volume) illustrate these arguments. They provide a new institutional analysis of the phenomenon of Internet commerce. The treatment of this phenomenon in strategic management is a classic example of the cost of ahistoricism. The fashion (fortunately, not universal) in strategy has been to approach Internet commerce as “new” – it was a new economy, operated by new organizational forms, requiring new strategies.¹ This approach has yielded a set of thoroughly forgettable scholarship, which has not endured even the recent downturn in the technology sector, let alone provided a theoretical basis to guide organizations that attempt to transact in evolving internet-related markets. In contrast, Clay and Strauss begin by identifying the historical precedents for the contemporary struggles of those who transact over the Internet. They identify an analog to the challenges of Internet commerce in Richard Sears' attempts to conduct business by mail in the late nineteenth century. Sears and his customers had the problem of building trust with strangers – for their transaction to be successful the customers needed to be confident that Sears would deliver high-quality goods, and Sears had to be confident that the customers would pay what they owed. Similar challenges arose for credit card providers and users in the

mid-twentieth century. These problems are strikingly similar to those faced by sellers and buyers over the internet. As Clay and Strauss argue, potential solutions for contemporary internet businesses are similar to those employed by nineteenth century mail-order businesses and 1950s credit card issuers – institutions can modify the transaction so that both parties can approach it with confidence.

AN INTRODUCTION TO THE NEW INSTITUTIONALISM

There are a number of variants of “new-institutional theory” (Fligstein, 1997), so it is important to be clear just what we mean by institutions, and how we understand them to operate. We’ll begin with this statement, which for us explains action in the new institutionalism: Actors pursue their interests by making choices within institutional constraints. This simple statement contains three elements that must be explained: who are the actors, how do they make choices, and what are the institutional constraints?

The Actors

The actors in the new institutionalism are individuals, organizations, or states. Each of these classes contains components that pursue interests, are subject to institutional constraints, and which supply institutional constraints that affect other actors. For researchers of strategic management, the idea that an organization can be, like an individual, an actor that pursues an interest, is uncontroversial. Without trivializing the fact that organizations have numerous and diverse stakeholders, the field of strategic management has shown repeatedly that there is utility in the simplification of organizations as actors. For example, in game theoretic treatments of entry deterrence, or models of competitive dynamics, an organization is an actor that can make “moves.” Likewise, it is central to the field that organizations have interests (otherwise, to what end would they have strategies?). New institutional arguments tend to be agnostic as to what the interests of individuals or organizations are, so the familiar candidates from strategic management – profit, market share, growth of employment, survival – are all feasible.

It may be more of a stretch to think of states as actors. This position reflects a recent trend in political science to characterize states as sets of organizations (ministries and agencies), which are like other organizations in that they are populated with individuals (bureaucrats and politicians) who use the state to achieve their goals (a paycheck, re-election; control over many subordinates,

furtherance of an ideological value). This is not to say that the state is just another organization – it has capabilities that other organizations can't match, such as the legitimate right to employ violence. The key, however, is that states represent interests, and take action to achieve those interests. We'll employ the typical definition of the state in the new institutionalism, as an organizational actor which, at a minimum, attempts to maintain its authority by exchanging justice and order for revenue and power (North, 1981; Skocpol, 1985).

Each class of actors produces its own form of institutional constraint: individuals produce norms (private-decentralized institutions in the classification we present below), organizations produce their rules (private-centralized institutions), and states produce laws and regulations (public-centralized institutions). In this sense, it can be said that actors lead a double life in the new institutionalism, pursuing their own interests within constraints, while producing constraints for other actors. The interplay between the actors can best be understood as a three-layered hierarchy, with states superordinate to organizations, which are superordinate to individuals (Williamson, 1994; Nee & Ingram, 1998). States constrain organizations and individuals that are their subjects, and organizations constrain the individuals that are their participants. There is also upward influence in the hierarchy, as actors try to affect the institutions that constrain them.

Finally, we introduce a fourth relevant class, which we'll call civil society. It would be incorrect to call civil society an actor – it doesn't have identifiable interests, and it is incapable of forming or pursuing a strategy. Yet, in the catholic version of new institutionalism that we are developing, civil society has a role as the source of a fourth type of institution – culture (public-decentralized). As we describe below, culture constrains action in a manner that is comparable to other institutional forms. Culture also influences those forms, as when it is used to create favor for one legal option or governance form over another (Dowell, Swaminathan & Wade, this volume; Henisz & Delios, this volume).

Choice: Bounded Rationality

The new institutionalism treats actors as rational in the basic sense of making choices that further their interests, but distinguishes itself from neo-classical assumptions of rationality by attending to "cognitive costs" of decision making. The pursuit of benefits is limited by individuals' capacity to retain and process information; in other words, individuals are boundedly rational (Coase, 1937; Simon, 1957). Further, information is often costly (Barzel, 1989). These two factors create transactions costs – the costs of writing and enforcing contracts

– because individuals cannot foresee at the time of writing all contingencies that might be relevant nor can they observe all of the actions of their partners. And transaction costs give rise to the possibility of opportunism (Williamson, 1975, 1985).

In the new institutionalism, a key implication of opportunism is the problem of credible commitment. It is illustrated by the dilemma faced by a kidnap victim whose kidnapper has a change of heart and decides to set her free (Schelling, 1960). The victim gladly promises not to reveal the kidnapper to the authorities in exchange for her freedom. However, the kidnapper realizes that once the victim is free she will have no incentive to keep her promise, and reluctantly decides the victim must be killed. More generally, the problem of credible commitment is faced by any party to an exchange that wants to promise in the present to do something in the future that may not be in their interests to do when the future actually arrives. The problem is endemic because in almost every exchange there is at least a moment where one of the parties has control over all or most of the goods, and must decide whether to follow through on the agreed upon bargain, or make a grab for more. It is clearly present, for example, in Richard Sears' attempt to get farmers to send him money for goods that he promised to subsequently send to them (Clay & Strauss, this volume).

The problem of credible commitment illustrates the positive role that institutions can play to smooth exchange (and by extension, to resolve all sorts of collective-action problems). Ideally, an institution can re-arrange the incentives of the parties of an exchange to allow them to make credible commitments. For example, what if it was possible for the kidnapper's victim to somehow post a bond that she would forfeit if she revealed the kidnapper's identity? And what if the farmer's friends maintained a norm to punish any vendor that mistreated any one of them, such that it was in Sears' interest to follow through on the bargain once he had the farmer's money? As we'll see, these examples do not describe all of the ways that institutions can affect economic performance of individuals, organizations and states. However, solving problems of credible commitment is one of the most positive functions of institutions, and one of the most significant for business strategy.

Institutional Forms

We employ an extension of the classification system for institutions that was first introduced in Ingram and Clay (2000). That system classifies institutions based on their scope (public or private), and how are they made and enforced (in centralized or decentralized fashion). The scope dimension defines which actors are subject to the institution. Public institutions apply without discrimination to all actors of a

	De-Centralized	Centralized
Private	Archetypal form: norms Chief actor: social groups Representative theorists: Homans, 1950; Granovetter, 1985 Levers for strategy: human resource policy; corporate-culture building; inter-organizational networks	Archetypal form: rules Chief actor: organizations Representative theorists: Williamson, 1975; Greif, 1994 Levers for strategy: conventional strategy and structure tools; business groups
Public	Archetypal form: culture Chief actor: civil society Representative theorists: Meyer & Rowan, 1977, DiMaggio & Powell, 1983 Levers for strategy: partnerships with mobilized social groups outside the firm; framing	Archetypal form: laws Chief actor: states Representative theorists: North, 1990; North & Weingast, 1989 Levers for strategy: non-market strategy; business political activity

Fig. 1. A Typology of Institutional Forms.

type. It is impossible to opt in or out of a public institution. Private institutions apply only to actors that are part of some group or organization, so actors have some influence over the institutions that affect them as long as they can choose which associations they are part of. The centralized-decentralized dimension refers to whether or not there are designated functionaries charged with creating and enforcing the institution. Centralized institutions rely on such functionaries, for example, laws may be made by legislatures and enforced by the police. The legislature and police are “third parties” in that they make and enforce the laws, even if they are not directly affected by their violation. Decentralized institutions, on the other hand, emerge from unorganized social interaction, and really on diffuse individuals (often those directly affected) to punish institutional violations. These two dimensions create four institutional forms. We describe each form, as well as research on the form that is relevant for strategic management. Figure 1 portrays these basic institutional forms, and summarizes key information about each.

Public-Centralized Institutions

There are at least five ways that the public institutions provided by the state can be understood to affect its choices, and those of organizations and

individuals. The first is particularly relevant to strategic management. The state may smooth exchange between its subjects by providing institutions that allow them to make credible commitments. This can be achieved if the state provides a legal system to protect property rights, decrease transaction costs, and enforce contracts (North, 1990). This function is particularly vital in modern economies, in which specialization and the division of labor give rise to the need for sustaining complex exchanges over time, across space, and between strangers, creating the need for trust between disconnected actors. An effective institutional framework facilitates this trust by penalizing actors who break the rules of exchange, for example, by applying legal sanctions to actors who violate contracts.

There is quantitative evidence of the role of public-centralized institutions for enabling credible commitments. Some studies exploit changes in laws governing specific industries to show that increased legal constraint on organizations causes them to flourish. Studies of populations as diverse as U.S. health maintenance organizations and telephone companies, Toronto day-care centers, Niagara Falls hotels, and Singapore banks have demonstrated that their failure is reduced by increasing government involvement in monitoring, certifying, authorizing and endorsing their activities (Wholey et al., 1992; Barnett & Carroll, 1993; Baum & Oliver, 1992; Ingram & Inman, 1996; Carroll & Teo, 1998). Such government involvement can also affect the pattern of competition between incumbent firms and potential entrants, as demonstrated in Calabrese et al.'s (2000) study of the Canadian biotechnology industry. In the human therapeutics/diagnostics sectors, where FDA regulation is most strict, new products take a decade to come to market and short technological leads can become entrenched as regulators demand evidence of superior efficacy for later-to-market drugs. Incumbent firms' innovative activity suppresses new entry significantly more in human subsectors than in subsectors characterized by less onerous regulatory scrutiny. The effects of broader changes in public institutions are seen in Ingram and Simons' (2000) analysis of the effect of the formation of the Israeli State on the failure rates of workers' cooperatives in many industries. The transition from the weak British Mandate for Palestine to the strong Israeli State caused a radical improvement in the institutional support for credible commitment, and a corresponding sixty-percent decrease in organizational failure rates.

The second key feature of public-centralized institutions is whether or not the state can credibly commit to not subsidize subject organizations when they struggle. The recent transitions from state socialism have demonstrated that absent such a commitment, entrepreneurs will direct their energies towards "holding up" the state treasury rather than to producing economic value. As Stark and Bruszt (1998, p. 119) put it, when the state hears organizations' "siren

cry, 'Give me a hand, give me your hand,' it must be bound to respond not simply that it should not, or that it will not, but that it cannot."

The third key feature of public institutions is an outgrowth of the first two. A state strong enough to guarantee the property rights of its subjects, and to resist their calls for subsidies, is also strong enough to appropriate their wealth. Unless the state can credibly commit against such appropriations, its subjects' incentives for productive economic activity will be greatly curtailed. Evans (1995) uses the term "predatory" to describe states that exploit their subjects for short-term gain. He cites Zaire of the Mobutu regime (1965 to the present) as an archetype. Mobutu and his state cronies "systematically looted Zaire's vast deposits of copper, cobalt, and diamonds, extracting vast personal fortunes . . . In return for their taxes, Zairians could not even count on their government to provide minimal infrastructure (p. 43)." The gains from this strategy to the state, and those who dominate it, are, however, short lived. Predation on the part of the state has the effect of discouraging productive activities by organizations of all types – why invest capital or labor if the state is likely to appropriate the rewards of this activity? This effect is apparent in the deceleration of the Zairian economy – GNP per capita declined two percent per year over the first twenty-five years of Mobutu's rule. Eventually, there will be little left to plunder.

A classic illustration of the cost of a predatory state, and the institutional solution that eliminated that cost, is North and Weingast's (1989) account of the Stuarts' impact on the economy of 17th century England. After coming to the Crown in 1612, the Stuarts exploited their subjects in numerous ways: they sold monopolies (at the expense of industry incumbents and potential entrants), they sold special dispensations from laws, and even committed outright theft, as in 1640 when they seized £130,000 that private merchants had placed in the Tower of London for safekeeping. These abuses led eventually to the Glorious Revolution of 1688, which resulted in numerous institutional changes to reduce the Crown's capacity to act independently of Parliament and the courts. This loss of Crown autonomy had, however, positive implications in that it enabled the Crown to make a credible commitment not to appropriate subjects' wealth. The value of this commitment can be seen, for example, in the dramatic increase in the Stuarts' capacity to borrow funds. More generally, a national constitution, with its delineation of enduring limits to government power, may be interpreted as an attempt by a state to commit not to become predatory over time (Weingast, 1993).

Public-centralized institutions provided by the state are not always part of a grand effort to facilitate the credible commitments of actors. Sometimes they influence distributional battles over zero-sum interests, which is their fourth role (Knight, 1992). These may be the battles between suppliers and consumers, as

shown in analyses of the effects of regulatory policy on railroad foundings in early Massachusetts (Dobbin & Dowd, 1997) or interstate trucking firm failures in the 1980s (Silverman, Nickerson & Freeman, 1997). Or they may be the battles between rival organizational forms without apparent efficiency differences, as in the case of thrift-savings organizations that fought as much in the legislative arena as in the market (Haveman & Rao, 1997), or national coffee roasters in the U.S., that derived a competitive advantage over regional roasters through an international treaty (Bates, 1997). Evans (1995) detailed numerous ways that states act to create economic transformations, for example, by lending money or taking responsibility for high-risk activities such as research and development. Such efforts are overwhelmingly selective, aimed at promoting particular sectors over others.

Fifth and finally, public-centralized institutions may affect the legitimacy of particular organizational forms by influencing the definition of organizational propriety. This influence may be concrete, as when a law requires a certain organizational practice or office, or intangible, as when myths of efficiency develop to justify a practice that the state endorses but does not enforce (Dobbin & Sutton, 1998). Legitimacy, in turn, affects organizations' capacity to obtain the resources they need to survive (Meyer & Rowan, 1977). This feature of public-centralized institutions represents a link between this institutional form and the public-decentralized institutional form. Essentially, the state is in a particular position to influence culture – this is one of the characteristics that differentiates the state from other types of organizations.

Private-Centralized Institutions

The most ubiquitous role of private-centralized institutions is to internalize transactions in an organization. In his seminal paper, Coase (1937) addressed the question of why organizations exist. His central insight was that the governance of exchange within organizations as opposed to markets depended on the cost of transacting in each type of institution. In more recent work, Williamson (1985) and others have systematically investigated the effect of information, opportunism, and asset specificity on the governance of exchange, concluding that in some transaction environments, exchange is more efficient within an organization than the market. Further, the prevailing public-centralized institutions influence the attractiveness of various governance arrangements (Nee, 1992). Ficker (1999) illustrates the relationship between the environment of public-centralized institutions and other factors, and the market/organization trade-off in the evolution of the Mexican Central Railroad (MCR). The MCR was founded in 1880 into “a country characterized by economic backwardness and an incipient and precarious institutional framework.” The company initially

pursued a strategy of building main lines and depending on market transactions with railroads and other types of transportation organizations to supply freight. These market transactions did not materialize, however, due to the weak Mexican infrastructure, and the difficulties of organizational and technological coordination. In response to this failing of the market, the MCR switched to a strategy of internalization, extending its trunk lines and building branch lines to supply itself with freight.

Private-centralized institutions can also facilitate exchange between organizations. This can occur when organizations are part of a “super-organization” with its own rules and policies. An example is the diamond industry. Bernstein (1992) examines the rules that govern transactions in that industry, which rely on private-centralized institutions, and not the law. Members of a diamond bourse are governed by formal written rules that represent the codification of and are supported by industry norms. Bernstein finds that use of arbitration panels and mandatory pre-arbitration conciliation is a response to members’ need for speed, secrecy, and specialized knowledge of the industry. The industry enforces arbitration decisions with the threat of suspension of membership in the diamond bourse. In the case of non-compliance, a bourse faxes the individual’s picture to all other diamond bourses worldwide. Informed of non-compliance, members then refuse to trade with the individual in question because of the risk that they will be cheated. Through this reputation mechanism, the institution creates incentives for members to adhere to industry rules and norms in their transactions with other members.

Ingram and Simons (2000) describe a private-centralized institution that is even more comprehensive than the diamond bourse. They explain that in Palestine under the British Mandate (1922–1948), the state failed to provide public-centralized institutions to support economic exchange. A large group of cooperative organizations created a private-centralized substitute for these missing institutions, in the form of a comprehensive federation, called the Histadrut. The Histadrut has had as members, at various times, agricultural cooperatives (the kibbutzim and moshavim), workers’ cooperatives (in services, manufacturing and transportation), Israel’s largest conglomerate, a bank, credit cooperatives, housing cooperatives and consumer cooperatives. The Histadrut did a number of things to smooth transactions between these members. For example, the Histadrut directed its affiliated bank and its major marketing cooperative to give preferential service to other Histadrut members. It also arbitrated disputes for members, provided auditing services, gave seminars in accounting and management, arranged bulk purchases of raw materials, and maintained a pension fund. The institutional framework provided by the

Histadrut was a major benefit to its members – they had a failure rate one-fifth that of non-members during the period of the British Mandate.

Private-Decentralized Institutions

The archetype of the private-decentralized institution is the norm. Although norms are often unwritten and unspoken, the real contrast to centralized institutions is in enforcement. Norms rely on social relationships for their enforcement – the ultimate penalty for violating a norm is the cessation of a relationship, or in the extreme, ostracism from a group. Penalizing violations of norms typically falls to the affected parties rather than third parties, and it is the value of the relationship itself that provides the motivation to maintain the norms that surround it. As Homans ([1961] 1974, p. 76) puts it, “the great bulk of controls over social behavior are not external but built into the relationship themselves, in the sense that either party is worse off if he changes his behavior toward the other.”

The operation of norms among individuals within organizations is familiar. An illustrative case is Homans (1950) re-analysis of the bank-wiring room from the Hawthorne studies. That study documents the norms of the workers, governing how much to produce on a shift. Normative enforcement through relationships is clear – workers who did not work hard enough were insulted, and excluded from games, gambling, and the sharing of candy. Although such group processes may at first seem tangential to the organization’s strategy and performance, the truth is that norms interact with the rules of the organization, and that interaction has a fundamental influence on organizational success (Nee & Ingram, 1998; Zenger, Lazzarini & Poppo, this volume). It was General Electric’s group piece-rate incentive system that set the background for the development of the norms in the bank-wiring room, which generally acted to encourage productivity. In other instances, sub-organizational norms undermine the control system of the organization, and inhibit its pursuit of its goals (Shibutani, 1978).

Norms, or their analogs, also operate in the inter-organizational context. There is important historical research that illustrates the function of private-decentralized institutions in long-distance trade. With this type of trade, merchants can often profit from using other merchants as agents to sell goods, collect debts, and so forth. This agency relationship, however, raises the possibility that the agent will act opportunistically, keeping some or all of the monies owed. The Maghribi traders in the 11th century Western Mediterranean (Greif, 1994), and American merchants on the 19th century California coast (Clay, 1997) overcame this problem by forming coalitions, which allowed exchange to flourish. In both cases, merchants in the coalition conditioned future

use of other merchants as agents on those merchants' having acted in accordance with group norms in the past. For instance, when a Maghribi merchant was accused of cheating in 1041–1042, he found that “people became agitated and hostile and whoever owed [me money] conspired to keep it from me (Greif, 1994, p. 925).” By tying future economic gains to past behavior as an agent, merchants were able to ensure that the future gains to membership in the coalition were greater than the gains to cheating and being punished.

Additionally, there is a rapidly expanding body of empirical research on the relational governance of inter-organizational exchange. Uzzi (1996) describes the embeddedness of exchange in the Manhattan garment industry. As in the examples above, participants in that industry followed norms that encouraged them to deal fairly and flexibly with each other. Zaheer, McEvily and Perrone (1998) apply related arguments to explain the purchasing practices of large organizations. They find that even among large corporations, trust between buyers and sellers is an important input to reducing transaction costs. There is also a developing literature in strategic management that focuses on the role of relationships for the interorganizational transfer of knowledge (e.g. Darr, Argote & Eppler, 1995).

Public-Decentralized Institutions

We have left public-decentralized institutions for last because they are different from the previous three institutional forms. The difference is in terms of intentionality. Laws, organizational rules, and norms are provided consciously and even strategically by states, organizations and individuals. These institutions don't always have the effects that their designers and enforcers intend, but none the less, they emerge from some intent. In contrast, public-decentralized institutions might be called “pre-conscious.” As we have described, they are provided by amorphous civil society, and not by a specific actor. Public-decentralized institutions amount to culture – they are ideas about what practices and social designs are acceptable and desirable.

So why include public-decentralized institutions in the same theory as other institutional forms? The best reason is that despite their origins, public-decentralized institutions are comparable to other institutional forms in their operation (Scott, 1995). Cultural values structure the choices of actors, partly determining the alternatives that are considered and the attractiveness of each alternative. For example, a manager's evaluation of an alternative for an organization's strategy might be influenced by the alternative's propriety and legitimacy in much the same way that it might be affected by its legality (DiMaggio & Powell, 1983; Carroll & Hannan, 2000). Indeed, work on the cognitive processes by which public-decentralized institutions operate indicate

that they influence the value of alternatives – for example, the stock of a company may be discounted because the company's activities do not fit legitimate categories (Zuckerman, 1999).

Additionally, although public-decentralized institutions are not controlled by any specific actor, actors may still be strategic in the face of them (DiMaggio, 1988; Roberts & Greenwood, 1997). All of this is not to oversimplify this institutional form. It is particular among the institutional classes in the pre-conscious manner in which it may emerge, and operate to affect action. Still, public-decentralized institutions fit the description of new-institutional action that we began with. The influence of propriety and legitimacy on the effectiveness of organizational designs and strategies is undeniable. Likewise, as we describe below, public-decentralized institutions have a critical affect on the development of other institutional forms, and thereby present important strategic opportunities for organizations.

WHAT CAN STRATEGY DO FOR THE NEW INSTITUTIONALISM?

So far, we've explained what the new institutionalism is, and why scholars in strategic management are paying more attention to it. But what is to be gained by the whole enterprise? Of course, the subsequent chapters will provide the answer to that question. But before we turn you loose on them, we'll give you our own interpretation. From the inception of this volume, we were convinced that there was a real promise of "gains from trade" by bringing together new institutional and strategy research. We see the benefits flowing both ways – strategy research can help solve some of the core problems of the new institutionalism, and vice versa.

The potential contribution of strategy to the new institutionalism comes from its sophisticated conceptualization of the action of organizations. Organizations are obviously key to new-institutional theory, not only as a source of private-centralized institutions, but even more importantly as the vehicles for the pursuit of the most important human interests, both economic and social (Hannan & Freeman, 1977). Yet, the treatment of organizations in the various institutional theories has been soundly criticized. DiMaggio (1988) charged theories of public-decentralized institutions with suffering from a "metaphysical pathos," denying the capacity of organizations and individuals for self-interested, and strategic, action. Similarly, Granovetter (1985) claimed the same theories presented an "oversocialized" view of organizations and other actors, underestimating their autonomy from cultural influence. Institutional economists on the

other hand, have erred in the opposite direction in their treatment of organizations. North (1993), for example, identifies organizations as the chief vehicles of institutional change. Yet, his characterization of organizations as malleable, rational and decisive entities is in contrast to what we know about organizational change and strategy making.

Without a doubt one of the prime contributions of the chapters in this book is to develop more informed theories of the role of organizations in institutional change. Indeed, many of the chapters represent great leaps forward for this critical but understudied topic. Jaffee and Freeman, for example, offer a thrilling account of institutional change in real time. They courageously make predictions about the evolution of German taxation of stock-options as the institutions are unfolding. Their analysis highlights the role of interest-seeking organizations in institutional change. German tax law forms the background for competition among established organizational forms and their challengers. The authors challenge and refine existing ideas (e.g. North, 1993) by identifying organizational inertia as a key determinant of the strategies that organizations pursue to maintain or change specific institutions.

The idea that legal institutions can be the object of organizational strategies is also prominent in other chapters. Holburn & Vanden Bergh present a formal model of lobbying. Their model explicitly reflects a key challenge to any organizational attempt to influence the law – that there are multiple options as to where lobbying efforts should be targeted. Should an organization lobby a state's executive, or its legislators? Or perhaps the organization should bypass the law-makers and go directly to the agencies that enforce laws and regulations? The model presented in this paper yields prescriptions that organizations can apply to their lobbying strategies.

De Figueiredo and De Figueiredo address a different stage of the same problem, using a radically different methodology. They follow on the under-developed literature on strategic decision making (e.g. Schwenk, 1984; Zajac & Bazerman, 1991). Implicit in their approach is a sophisticated idea that has been slighted in the new institutionalism, that the influence of institutions depends on the perception of those institutions by the relevant actors. Their experiments yield insights into the very practical problem of how to help strategists to correctly analyze the institutional environment, and recognize the opportunities and challenges that it presents. Their results are cause for optimism, evidencing the utility of business-school courses on the strategy of dealing with institutions.

Two other papers in the volume examine organizations' role in institutional change, but focus on culture, rather than the law, as the context for organizations' strategies. In other words, they venture bravely into the void between

established institutional theories. Dowell, Swaminathan and Wade examine a fascinating instance of institutional entrepreneurship, the effort to establish standards for HDTV in the U.S. This chapter provides a uniquely lucid explanation of the sociology and psychology of “framing.” Framing is a conscious attempt to use cultural values to support a given action, or in this case, a direction for institutional change. Framing therefore represents a vast set of strategic opportunities for institutional entrepreneurs. The clear and compelling treatment of the topic in this chapter will encourage its application to other strategies and studies.

Rao’s chapter is similar in that it considers organizations’ influence on public-decentralized institutions, but it considers taken-for-grantedness, rather than a technological standard. Taken-for-grantedness presents a dilemma for strategy – it has been convincingly shown to influence organizational performance, but what can be done with that knowledge? How can organizations affect what others take for granted? Rao’s answer is that they can do so through demonstration. Specifically, he examines the demonstrations of reliability and quality that seeded the taken-for-grantedness of the automobile. In doing so, he gives fair treatment to the many other influences on the perception of the automobile. The result is a necessarily complex, but original and very promising set of ideas about the interdependence of organizations and public-decentralized institutions.

The interdependence between institutional forms demonstrated in Rao’s paper (and in most of the papers in this volume) is itself a substantial contribution to the new institutionalism. Research has so far been mainly within the quadrants of Fig. 1, with interdependencies between the quadrants going unexamined to the detriment of the theory. While the multi-form emphasis of so many of this volume’s papers redresses that theoretical neglect, it also has implications for the field of strategy. Unpacking the simultaneous influence of different types of institutions is necessary for effective strategizing. In many ways the theoretical and applied contributions of these papers are intertwined, as the next section demonstrates.

WHAT CAN THE NEW INSTITUTIONALISM DO FOR STRATEGY?

The previous section described the benefits that new institutional scholars can gain from taking strategy research seriously. But this is by no means a one-way street. Our rationale for including a volume on new institutionalism in the *Advances in Strategic Management* series is predicated on the idea that

the new institutionalism can help surmount some of the core challenges in strategy research.

The potential contribution of new institutional research to strategy comes from its highlighting of the interactive role that institutions play in both constraining and enabling organizational action. Institutions are frequently seen as background conditions or “shift parameters” that contour the expected payoffs associated with particular strategic actions (Williamson, 1991). But more than that, institutions directly determine what arrows a firm has in its quiver as it struggles to formulate and implement strategy, and to create competitive advantage. Given the importance of institutions for determining the success or failure of specific strategies or actors, consideration of ways to influence the creation and maintenance of favorable institutions is fundamental to any organization’s strategy. Hence, an understanding of institutional change, and the ways that firms can influence such change, becomes central to the study and practice of strategy.

Consider the taxonomy of institutions in Fig. 1. The vast majority of strategy research focuses on private-centralized institutions such as firms and the formal actions that they undertake. Strategy research has generated a post-adolescent, if not quite mature, body of literature that offers strong predictions and prescriptions for firms’ boundaries and competitive activity. Yet direct application of traditional strategy prescriptions to managing other types of institutions offers far less utility. How can a firm deal with private-decentralized institutions such as norms? Use of traditional strategic levers without consideration of variance in underlying norms of organization members can be ineffective, or can even backfire, as in Simon’s (1957) “unintended consequences.” As for altering norms themselves, this is frequently perceived as an organizational behavior or human resource management issue, and consequently outside the purview of strategy. Similarly, how can a firm deal with public-centralized institutions? With the exception of what is sometimes called the “non-market strategy” literature (Baron, 2001), these are largely seen as exogenous institutions that influence strategic and organizational choices in much the same way as production technology. Finally, the link between strategy research and public-decentralized institutions has generally been limited to the use of cultural distance measures (Hofstede, 1980) to predict the rate and mode of entry of multinational firms into specific host markets (Kogut & Singh, 1988; Hennart & Park, 1993).

The chapters in this book contribute to the development of deep insights into the influence of all four types of institutions on firm strategy, and vice versa. Even more exciting, many of the chapters set forth into new terrain regarding the *interactions* across types of institutions as well as their relationship to strategy. For example, Zenger, Lazzarini and Poppo propose a novel and

compelling extension to the theory of the firm by considering interactions between private-centralized and private-decentralized (or “formal” and “informal”) institutions. Starting from a few basic assumptions about differences in the characteristics of each of these types of institutions, they develop a series of bold propositions that potentially resolve several puzzles in the strategy and organization literature. The authors argue that although an organization cannot quickly alter private-decentralized institutions through the usual methods of changing formal organization structures, such formal changes can spark gradual changes in such norms. Coupling this argument with the common assumption that formal organizational structures are discrete (and hence can not be incrementally tweaked to achieve an optimal form), they provide a rationale for the seemingly-constant oscillation of structures that many organizations demonstrate (Nickerson & Zenger, 2002).

But the best way for us to convey what the new institutionalism can do for strategy research is to describe in detail the layout of this volume.

SPECIFIC RESEARCH QUESTIONS, AND THE LAYOUT OF THIS VOLUME

The above discussion noted general insights that can be fruitfully drawn from the new institutionalism into strategy research. We have organized the volume so as to highlight insights related to specific research questions in strategy. In their manifesto for the strategy field, Rumelt, Schendel and Teece (1994) suggest four fundamental questions in strategy research: (1) How do firms behave? (2) Why are firms different? (3) What limits the scope of the firm? (4) What determines success or failure in international competition? New institutional research, and in particular the research in this volume, speaks to each of these questions.

How do firms behave? Or, do firms really behave like rational actors, and, if not, what models of their behavior should be used by researchers and policy makers?

In “Policy and process: A game-theoretic framework for the design of non-market strategy,” Guy Holburn and Richard Vanden Bergh adopt a far-sighted rational action lens to explore interactions among agents of the state and their effect on how firms should try to influence legal/institutional framework. Expanding on positive political theory models of lobbying in the U.S., they demonstrate that different agents – regulatory agency officials, legislators, other elected officials – become the pivotal actors depending on the

distribution of preferences across these actors. Thus, a firm that wishes to influence a specific regulation should not necessarily lobby the regulator directly, but in many cases will need to direct its lobbying efforts towards other political actors.

In "Managerial decision making in non-market environments: A survey experiment," John de Figueiredo and Rui de Figueiredo test the assumption that managers are able to pursue far-sighted rational action. Noting that a large body of experimental literature raises questions about the rationality of managers, they conduct a series of experiments designed to test the ability of managers to make "optimal" decisions about activities designed to influence legal institutions, such as investing in lobbying activity. Their results indicate that although managers are competent at making optimal decisions when confronted with simple, single-stage problems, managerial decisions deviate significantly from optimal choices as problems become more complex.

In "Pretty pictures and ugly scenes: Political and technological maneuvers in high definition television," Glen Dowell, Anand Swaminathan, and Jim Wade study institutional change regarding the allocation of broadcasting spectrum in the U.S. They provide a case study of the various attempts by television broadcasters to fight FCC regulations in the mid-1980s that would take away unused spectrum from broadcasters and allocate it to other uses such as cellular communication. Initial attempts to get Congress to overturn this regulation foundered, due both to the difficulty of overcoming a "collective action" problem among the diverse broadcasters and to the lack of a resonant "frame" to motivate Congress. Yet subsequent attempts succeeded, once the broadcasters found a way to frame their need for spectrum in terms of U.S. manufacturing competitiveness vis-à-vis Japan, a particularly resonant frame in the late 1980s. Dowell et al. explain these outcomes through the lens of social movement theory, and particularly the role of framing problems in ways that motivate desired action. Rather than far-sighted rational actors, the managers and policymakers in this lens are characterized by subjective perception, and the institutional outcome is determined during the battle to socially construct the frame of the institutional change.

In "The evolution of university patenting and licensing procedures: An empirical study of institutional change," Bhaven Sampat and Richard Nelson take a still different view of actors' behavior and motivation. Drawing on a routine-based view of organizational action, they argue that actors develop "social technologies" to manage their various activities. As these social technologies diffuse and harden into standardized patterns of behavior, they become institutions. Hence, institutions arise through the boundedly rational attempts of actors to solve problems, notably problems associated with production and exchange. Sampat and Nelson study the diffusion of different social

technologies used by universities to manage their patenting and licensing activities, culminating in the technology transfer office commonly found at research universities today. Interestingly, they note that the diffusion of this institution was facilitated by the passage of the Bayh-Dole Act of 1980, for which universities actively lobbied – and the motivation for which seems to have been based on erroneous and inaccurate evidence of a university-industry technology transfer “market failure.”

Thus the four chapters in this section all focus on firms’ efforts to influence public and/or private institutions, and each takes a slightly different perspective on the fundamental strategic question: How do firms behave? It is interesting to note that this difference mirrors the differences in strategy literature writ large. But the consecutive presentation of these perspectives in this volume is informative in a way that heterogeneity in the broader literature is not. The papers here point to the fact that styles of choice depend on their context. The type of institutions that most constrain an actor greatly affect the appearance of the actor’s decisions. Actors attempting to influence the complex, but well-defined institutional structure represented by the U.S. government may seem intentional and calculative, but occasionally confused. When culture is the object or key constraint, decisions follow different styles because the rules of decision making are different. For example, the symbolic value of behavior may become more important.

Why are firms different? Or, what sustains the heterogeneity in resources and performance among close competitors despite competition and imitative attempts?

In “Competition, contingency, and the external structure of markets,” Ron Burt, Miguel Guilarte, Holly Raider, and Yuki Yasuda explore the implicit institutional foundation of market structure among industry competitors. They propose and demonstrate a network-based measurement of “effective competition” among rivals. They find that an apparent puzzle in prior literature can be explained by incorporating effective competition. Specifically, prior research has found that strong corporate culture is only erratically associated with firm performance. Burt et al. demonstrate that the relationship between corporate culture and firm performance is contingent on the level of effective competition faced by the firm – in highly competitive markets, strong culture enhances performance, while in low-competition markets culture has no impact on performance.

In “Institutional change in ‘real-time’: The development of employee stock options in German venture capital,” Jonathan Jaffee and John Freeman analyze the attempt by several young German law firms to gain legal clearance to implement “American-style” employee stock ownership plans for their clients who were start-up and venture capital firms, and the opposition to this from

several large, well-established law firms who did less work with start-ups. Conceptually, their study shows how firms can influence the institutional framework – in this case, concerning the legality of certain stock compensation policies – to further entrench their relative advantages over competitors. Institutions (and the ability to enact institutional change) thus become central features explaining sustainable performance differences among firms.

In “Institutional barriers to electronic commerce: An historical perspective,” Karen Clay and Robert Strauss study several historical precedents to Internet commerce. They note that Richard Sears, and later various credit card companies, faced challenges of opportunism associated with remote commerce that sound very familiar today, and they analyze the emergence of several institutions that ameliorated such opportunism in the past. In addition to pointing us toward institution-based solutions to the challenges facing current Internet businesses, Clay and Strauss underscore the competitive advantage that can accrue to a firm, or a group of firms, that successfully undertakes institutional innovation. By “solving” the remote commerce problem, Sears was able to grow quickly to dominate the mail order business in the late nineteenth century; thus, Sears’s institutional innovation provided the firm with a first-mover advantage that endured for nearly a century. Those businesses that establish private institutions to solve current challenges to Internet commerce may similarly enjoy enduring performance benefits. And, as in the preceding chapter, Clay and Strauss argue that performance differences between Internet- and brick-and-mortar businesses will turn largely on the outcome of battles over broad institutional issues such as taxation of Internet commerce.

These papers illustrate that sound new-institutional arguments are in the spirit of Henderson and Mitchell’s (1997) call for research that explores interactions between market effects and internal capabilities. Burt et al.’s study demonstrates how the competitive environment influences the value of a organization-specific institution (culture). Similarly, Jaffee and Freeman emphasize the significance of compliance between organizational form and the institutional environment. This valuable compliance creates strategic opportunities for organizations to manipulate the institutions that surround them. Finally, Clay and Strauss remind us that, given the appropriate market structure, a firm’s institutional innovations can potentially provide a source of sustained competitive advantage.

What limits the scope of the firm? Or, what is the function of or value added by the headquarters unit in a diversified firm?

In “Informal and formal organization in new institutional economics,” Todd Zenger, Sergio Lazzarini, and Laura Poppo note that prior scholarship on the theory of the firm has largely focused on either formal institutions such as

contracts, or on informal institutions such as norms, and rarely on the interactions between the two. Beginning with a few basic assumptions about the characteristics of each type of institution, they explicitly analyze interactions between formal and informal institutions. Zenger et al. derive a series of startling propositions that potentially resolve a number of puzzles that have challenged the theory of the firm over the last thirty years. Chief among these are: (1) when will formal and informal institutions act as substitutes, and when as complements; (2) what explains some organizations' apparent predilection for cycling (and recycling) through organization structures frequently; and (3) what precisely limits the size of the firm? The chapter suggests a number of fruitful directions for empirical testing as well.

In "Tests tell: Constitutive legitimacy and consumer acceptance of the automobile, 1895–1912," Hayagreeva Rao explores the growth in consumer acceptance of the automobile in the years following its initial commercialization. In particular, he examines the role of several activities – both those managed by the firm, such as advertising, and those propelled by actors outside the firm, such as auto demonstration races sponsored by social movement-like organizations of car enthusiasts – on auto sales. This study contributes a novel look at the way that social movement theory may explain the effectiveness of particular public-decentralized institutions, such as auto clubs. It also demonstrates how advertising and social movement things are differentially effective at different times, and also work as substitutes. As such, the chapter provokes consideration of the conditions under which a firm should strategically consider mobilizing forces outside its formal boundaries to enhance its competitive strategy.

These papers suggest a reframing of the definition, often used in strategy, of the firm as a nexus of contracts. It may be more useful to consider the firm as a nexus of institutions. This broader characterization takes explicit account of the multiple institutional forms that affect behavior or and within organizations. A more strategically tractable understanding emerges by recognizing the multiple institutional forms that constitute organizations – law, culture, norms, and rules. As Zenger et al. show, an organization is not merely the agency-theory driven rules of employment, but also the norms that are associated, but not completely coupled to them. As Rao shows, organizations have cultural identities that are separate from their product profiles, but fundamental to their effectiveness.

What determines success and failure in international competition? Or, what are the origins of success and what are their particular manifestations in international settings of global competition?

In "Learning about the institutional environment," Witold Henisz and Andrew Delios note that prior literature on FDI typically treats firms as homogeneous

while examining the relationship between FDI and national variation, or treats national institutions as homogeneous while examining the firm variation-FDI relationship. They explore the joint roles of heterogeneous firm experience and heterogeneous institutional environments in explaining the direction and mode of foreign direct investment. In their framework, a firm's experience provides firm-specific knowledge that moderates the influence of variation in institutional environment, thus leading multinational corporations with different patterns of experience to pursue different entry strategies and expect different performance outcomes in international competition. By recognizing variation in both firm experience and institutional environments, they are able to propose a wide range of empirically refutable implications that significantly extend current strategy research on international business.

In "Institutions and the vicious circle of distrust in the Russian market for household deposits, 1992–1999," Andrew Spicer and William Pyle explore the apparent failure of public and private institutions in Russia to support the development of a private market for household savings deposits. They analyze the events associated with these development attempts, and argue that the initial "institutional backdrop" at the birth of this sector contributed to a self-reinforcing cycle in which private commercial banks were unable, either individually or collectively, to win the trust of potential depositors. Of particular interest to strategy researchers, their research demonstrates how several traditional strategic prescriptions are implicitly predicated on deep assumptions about the institutional backdrop. For example, although advertising is often seen as an investment to demonstrate high quality, Spicer and Pyle suggest that in Russia, where consumers had a low level of market savvy and where regulatory institutions were not set up to enforce certain behaviors among banks, advertising apparently had either no relationship, or even an inverse relationship, with bank quality. Their analysis reinforces our understanding of the difficulty of simply "porting" from one nation to another successful businesses, or even successful institutions, without the appropriate supporting institutional backdrop.

These papers point the way towards a theoretically-informed refinement of the international business environment. They illustrate that every country represents a complex web of characteristics. Traditional ideas of "foreign and local" or one dimensional characterizations of a country (e.g. collectivist or individualist; common-law or Napoleonic Code) are insufficient. But beyond that point, which should be uncontroversial, new-institutional theory can pave the way for a rigorous analysis of the multi-faceted institutional environment that each country represents. Indexes of institutional stability or the environment for investing (e.g. Henisz, 2000) present a real opportunity for both researchers and strategists to incorporate institutional sophistication into their country-characterizations.

THE LAST WORD

We hope that this volume will inspire scholars of both the new institutionalism and of strategy to explore the exciting research opportunities lying at the juncture of fields. To the extent that it does, we are convinced that the methodological approaches demonstrated in this volume provide brilliant guideposts for such work. Let's do this again in ten years and see what we have wrought!

NOTE

1. Some seemed even to believe that the topic required new forms of scholarship. We are aware of one unfortunate full professor at a top-twenty U.S. business school who changed his title from "Professor of Strategy" to "Professor of New Economy."

REFERENCES

- Barnett, W. P., & Carroll, G. R. (1993). How institutional constraints affected the organization of early American telephony. *Journal of Law, Economics, and Organization*, 9, 98–126.
- Baron, D. (2001). Theories of strategic non-market participation: Majority-rule and executive institutions. *Journal of Economics and Management Strategy*, 10, 47–89.
- Barzel, Y. (1989). *The Economic Analysis of Property Rights*. New York: Cambridge University Press.
- Bates, R. H. (1997). *Open Economy Politics: The Political Economy of the World Coffee Trade*. Princeton, NJ: Princeton University Press.
- Baum, J. A. C., & Oliver, C. (1992). Institutional embeddedness and the dynamics of organizational populations. *American Sociological Review*, 57, 540–559.
- Bernstein, L. (1992). Opting out of the legal system: extralegal contractual relations in the diamond industry. *Journal of Legal Studies*, 21, 2583–2612.
- Calabrese, T., Baum, J. A. C., & Silverman, B. S. (2000). Canadian biotechnology start-ups, 1991–1997: The role of incumbents' patents and strategic alliances in controlling competition. *Social Science Research*, 29, 503–534.
- Carroll, G. R., & Hannan, M. T. (2000). *The Demography of Corporations and Industries*. Princeton, NJ: Princeton University Press.
- Carroll, G. R., & Teo, A. C. Y. (1998). *How Regulation and Globalization Affected Organizational Legitimation and Competition among Commercial Banks in Singapore, 1840–1994*. Presented at the 58th Annual Meeting of the Academy of Management, San Diego.
- Clay, K. (1997). Trade without law: private-order institutions in Mexican California. *Journal of Law, Economics and Organization*, 13, 202–231.
- Coase, R. H. (1937). The nature of the firm. *Economica*, 4, 86–405.
- Darr, E. D., Argote, L., & Eppler, D. (1995). The acquisition, transfer and depreciation of knowledge in service organizations: productivity in franchises. *Management Science*, 41, 1750–1762.
- DiMaggio, P. J. (1988). Interest and agency in institutional theory. In: L. G. Zucker (Ed.), *Institutional Patterns and Organizations: Culture and Environment* (pp. 3–21). Cambridge, MA: Ballinger.

- DiMaggio, P. J., & Powell, W. W. (1983). The iron cage revisited: Institutional isomorphism and collective rationality in organizational fields. *American Sociological Review*, 48, 147–160.
- Dobbin, F., & Dowd, T. J. (1997). How policy shapes competition: early railroad foundings in Massachusetts. *Administrative Science Quarterly*, 42, 501–529.
- Dobbin, F., & Sutton, J. R. (1998). The strength of a weak state: The rights revolution and the rise of human resource management divisions. *American Journal of Sociology*, 104, 441–476.
- Ellickson, R. C. (1991). *Order Without Law*. Cambridge, MA: Harvard University Press.
- Evans, P. (1995). *Embedded Autonomy: States & Industrial Transformation*. Princeton, Princeton University Press.
- Ficker, S. K. (1999). Institutional Constraints and the Strategy of the Firm: An American Railroad Corporation in Nineteenth-Century Mexico. Presented at the Conference on Institutions and Markets in Comparative-Historical Perspective, Palo Alto, CA.
- Fligstein, N. (1997). Fields, Power, and Social Skill: A Critical Analysis of the New Institutionalisms. Presented at the Conference on Power and Organization, Hamburg, FRG.
- Granovetter, M. S. (1985). Economic action and social structure: the problem of embeddedness. *American Journal of Sociology*, 91, 481–510.
- Greif, A. (1994). Cultural beliefs and the organization of society. *Journal of Political Economy*, 102, 912–950.
- Hannan, M. T., & Freeman, J. (1977). The population ecology of organizations. *American Journal of Sociology*, 82, 929–964.
- Haveman, H. A., & Rao, H. (1997). Structuring a theory of moral sentiments: Institutional and organizational coevolution in the early thrift industry. *American Journal of Sociology*, 102, 1606–1651.
- Henderson, R. M., & Mitchell, W. (1997). The interactions of organizational and competitive influences on strategy and performance. *Strategic Management Journal*, 18(Summer), 5–14.
- Healy, K. (2002). Sacred Markets and Secular Ritual in the Organ Transplant Industry. Paper presented at the Workshop on Economic Sociology, Princeton University, February.
- Henisz, W. J. (2000). The Institutional Environment for Economic Growth. *Economics and Politics*, 12(1), 1–31.
- Hennart, J. F., & Park, Y. R. (1994). Greenfield vs. acquisition: The strategy of Japanese investors in the United States. *Management Science*, 39, 1054–1070.
- Hofstede, G. (1980). *Culture's Consequences: International Differences in Work-related Values*. Beverly Hills: Sage Publications.
- Homans, G. C. (1950). *The Human Group*. New York: Harcourt Brace Jovanovich.
- Homans, G. C. (1974) [1961]. *Social Behavior: Its Elementary Form*. New York: Harcourt Brace Jovanovich.
- Ingram, P., & Clay, K. (2000). The choice-within-constraints new institutionalism and implications for sociology. *Annual Review of Sociology*, 26, 525–546.
- Ingram, P., & Inman, C. (1996). Institutions, intergroup rivalry, and the evolution of hotel populations around Niagara Falls. *Administrative Science Quarterly*, 41, 629–658.
- Ingram, P., & Simons, T. (2000). State formation, ideological competition, and the ecology of Israeli workers' cooperatives, 1920–1992. *Administrative Science Quarterly*, 45, 25–53.
- Knight, J. (1992). *Institutions and Social Conflict*. New York: Cambridge University Press.
- Kogut, B., & Singh, H. (1988). The effect of national culture on the choice of entry mode. *Journal of International Business Studies*, 19, 411–432.

- Meyer, J. W., & Rowan, B. (1977). Institutionalized organizations: formal structure as myth and ceremony. *American Journal of Sociology*, 83, 340–363.
- Murphy, K. M., Shleifer, A., & Vishny, R. W. (1992). The transition to a market economy: Pitfalls of partial reform. *Quarterly Journal of Economics*, 107, 889–906.
- Murphy, K. M., Shleifer, A., & Vishny, R. W. (1993). Why is rent-seeking so costly to growth? *American Economic Review*, 83, 409–414.
- Nee, V. (1992). The organizational dynamics of market transition: hybrid forms, property rights, and mixed economy in China. *Administrative Science Quarterly*, 37, 1–27.
- Nee, V., & Ingram, P. (1998). Embeddedness and beyond: institutions, exchange, and social structure. In: M. C. Brinton & V. Nee (Eds), *The New Institutionalism in Sociology* (pp. 19–45). New York: Russell Sage Foundation.
- Nickerson, J. A., & Zenger, T. R. (2002). Being efficiently fickle: A dynamic theory of organizational choice. *Organization Science*, forthcoming.
- North, D. C. (1981). *Structure and Change in Economic History*. New York: WW Norton.
- North, D. C. (1990). *Institutions, Institutional Change and Economic Performance*. New York: Cambridge University Press.
- North, D. C. (1993). Institutions and credible commitment. *Journal of Institutional and Theoretical Economics*, 149, 11–23.
- North, D. C., & Weingast, B. R. (1989). Constitutions and credible commitments: the evolution of the institutions of public choice in 17th century England. *Journal of Economic History*, 49, 803–832.
- Roberts, P. W., & Greenwood, R. (1997). Integrating transaction cost and institutional theories: Toward a constrained-efficiency framework for understanding organizational design adoption. *Academy of Management Review*, 22, 346–373.
- Rumelt, R. P., Schendel, D., & Teece, D. J. (1994). Introduction: Fundamental issues in strategy. In: R. P. Rumelt, D. Schendel & D. J. Teece (Eds), *Fundamental Issues in Strategy* (Chap. 1). Boston: Harvard Business School Press.
- Schelling, T. C. (1960). *The Strategy of Conflict*. Cambridge, MA: Harvard University Press.
- Schwenk, C. R. (1984). Cognitive simplification processes in strategic decision making. *Strategic Management Journal*, 5, 111–128.
- Scott, W. R. (1995). *Institutions and Organizations*. Beverly Hills, CA: Sage.
- Shibutani, T. (1978). *The Derelicts of Company K: A Study of Demoralization*. Berkeley: University of California Press.
- Silverman, B. S., Nickerson, J. A., & Freeman, J. (1997). Profitability, transactional alignment, and organizational mortality in the U.S. trucking industry. *Strategic Management Journal*, 18 (Summer), 31–52.
- Simon, H. A. (1957). *Models of Man*. New York: John Wiley & Sons.
- Skocpol, T. (1985). Bringing the state back in: Strategies of analysis in current research. In: P. B. Evans, D. Rueschemeyer & T. Skocpol (Eds), *Bringing the State Back In* (pp. 3–37). Cambridge: Cambridge University Press.
- Stark, D. (1996). Recombinant property in east European capitalism. *American Journal of Sociology*, 101, 993–1027.
- Stark, D., & Bruszt, L. (1998). *Postsocialist Pathways*. New York: Cambridge University Press.
- Strange, S. (1996). *The Retreat of the State*. Cambridge: Cambridge University Press.
- Teece, D. J. (1981). The market for know-how and the efficient international transfer of technology. *Annals of the American Academy of Political and Social Science*, 458, 81–96.
- Uzzi, B. (1996). The sources and consequences of embeddedness for the economic performance of organizations: The network effect. *American Sociological Review*, 61, 674–698.

- Weingast, B. R. (1993). Constitutions as governance structures: the political foundations of secure markets. *Journal of Institutional and Theoretical Economics*, 149, 286–311.
- Wholey, D. R., Christianson, J. B., & Sanchez, S. M. (1992). Organization size and failure among health maintenance organizations. *American Sociological Review*, 57, 829–842.
- Williamson, O. E. (1975). *Markets and Hierarchies: Analysis and Antitrust Implications*. New York: Free Press.
- Williamson, O. E. (1985). *The Economic Institutions of Capitalism*. New York: Free Press.
- Williamson, O. E. (1991). Comparative economic organization: The analysis of discrete structural alternatives. *Administrative Science Quarterly*, 36, 269–296.
- Williamson, O. E. (1994). Transaction cost economics and organization theory. In: N. J. Smelser & R. Swedberg (Eds), *The Handbook of Economic Sociology* (pp. 77–107). Princeton, NJ: Princeton University Press.
- Zajac, E., & Bazerman, M. (1991). Blind spots in industry and competitor analysis: Implications of interfirm (mis)perceptions for strategic decisions. *Academy of Management Review*, 16, 37–56.
- Zaheer, A., McEvily, B., & Perrone, V. A. (1998). Does Trust Matter? Exploring the Effects of Interorganizational and Interpersonal trust on Performance. *Organization Science*, 9, 141–159.
- Zuckerman, E. W. (1999). The categorical imperative: Securities analysts and the legitimacy discount. *American Journal of Sociology*, 104, 1398–1438.

PART 1:
HOW DO FIRMS BEHAVE?

This Page Intentionally Left Blank

POLICY AND PROCESS: A GAME-THEORETIC FRAMEWORK FOR THE DESIGN OF NON-MARKET STRATEGY

Guy L. F. Holburn and Richard G. Vanden Bergh

ABSTRACT

*We draw on the Positive Political Theory literature to develop insights into how firms decide whether to lobby legislatures or agencies in order to gain favorable policy outcomes. We present a simple structural model of the interaction among a firm, a legislature, an executive, a court and an agency to illustrate how, even if the **agency** has responsibility for implementing public policy, the firm will, under the right conditions, lobby the **legislature** instead to bring about a change in policy. Accordingly, we contribute to the existing non-market strategy literature by incorporating institutional players other than the legislature into the analysis, and by addressing the question of how firms allocate lobbying resources across the different branches of government.*

INTRODUCTION

The ubiquitous presence of the government in private sector business transactions, either directly or indirectly, means that the competitive environments in which firms operate can depend significantly on public policy. Consequently, firms have an incentive to design non-market strategies that shape government decisions on issues that affect firm performance. While market strategies consist of actions aimed at shaping interactions with competitors, customers and suppliers in the market place (e.g. pricing and investment decisions), non-market strategies consist of actions specifically designed to influence the institutional players who determine public policy – state and federal legislatures, executives, regulatory agencies and courts – and include activities such as contributing to electoral campaign funds, lobbying and litigation. In the broadest sense, non-market strategies are primarily concerned with using the public policy process to shape the allocation of property rights or the “rules of the game” that govern the interactions among firms, competitors and consumers in their market environment.

Although there is a burgeoning literature on the objectives of non-market strategies (Baysinger, 1984; Weidenbaum, 1980), the categories of firms that engage in extensive non-market activities (Pittman, 1976; Zardkoohi, 1985) and the specific activities and tactics used in implementing non-market strategies (Bonardi, 1999; Baron, 1996, 1995; Buchholz, 1990), relatively little research has been conducted on how the structure of government institutions affects the strategic choices that firms make when allocating resources in order to maximize their impact on public policy.¹ Although government decision-making procedures are frequently complex, involving multiple players – legislatures, executives, agencies and courts, each with their own preferences – and a variety of decision-making rules, many existing non-market studies focus on one stage or player in the broader policy process. Propositions derived from such research often lead to highly conditional strategic implications that are divorced from the broader institutional context. In this paper, we contribute to the current literature by drawing on the Positive Political Theory (PPT) literature to present an approach to the formulation of non-market strategy that builds on a structural model of the public policy process. By so doing, we focus attention on the fact that firm strategies aimed at influencing government decisions depend significantly on the “rules of the game” that define the institutional context and process by which public policy is determined.

The chapter consists of three sections. We begin first by briefly discussing the role of non-market strategy in determining firm performance and its relative importance compared to market-based strategies. With this as background, we then turn to a review and assessment of the non-market strategy literature,

focusing on the implicit or explicit assumptions typically made about the underlying public policy process. We find that the bulk of the literature does not address the full complexity of the policy process, often leaving aside interdependencies among executive, legislative, judicial and agency branches of government. Consequently, little attention has been paid so far to how firms should allocate resources across different branches of the government in order to achieve the greatest effect on ultimate public policy. We propose that by using game-theoretic techniques to model the structured series of interactions between the various government branches, it is possible to derive strategic prescriptions about which part of the government firms should devote non-market resources to.

In the third section, we illustrate the potential benefit of adopting a more formal approach to modeling non-market strategy by outlining a simple game where a firm interacts with an agency, executive, legislature and court. While the game is not designed to represent a specific situation and is quite simplistic in its scope, we employ it to demonstrate that firm decisions to lobby one branch of the government or another (e.g. an agency) are typically highly dependent on the relative preferences of the other players (e.g. courts, executive, legislature) as well as on the sequence of play and the decision rules at each stage. By incorporating the full policy “game”, we find that the firm need not directly lobby the player with specific responsibility for implementing policy decisions (e.g. an agency) in order to induce that player to shift policy in the firm’s favor. Indeed, in certain circumstances, lobbying the primary player will be fruitless in that it has no impact on policy. In these situations, which occur for example when an agency is effectively constrained by the future threat of legislative override, the firm will instead lobby the legislature – since the agency’s decisions track the legislature’s preferences, any change in the legislature’s ideal point will be mirrored by equivalent shifts in agency rulings. More generally, our goal here is to develop a framework that allows us to analyze the formulation of non-market strategies and to tease out testable hypotheses that will guide future empirical work.

THE ROLE OF NON-MARKET STRATEGY

In many industries, performance depends on the ability of firms to shape public policy as well as on the ability to influence competitor and consumer decisions directly in the market place. Governments can have a direct or indirect impact on firm profitability, for example with the passage of legislation limiting domestic or foreign entry by competitors into a market place, with agency rulings on final rates in regulated industries or on the protection of intellectual property, or

with judicial decisions in litigated disputes. The presence of the government, either as an active or passive influence on the nature of competitive markets, thus creates an additional opportunity for firms to improve their performance other than through the design of market strategies. Here, we define a firm's non-market strategy as a concerted set of actions aimed at influencing government decisions on public policies. Such actions consist of those taken in non-market arenas (e.g. lobbying legislators or agencies, contributing funds to electoral campaigns) as well as those taken in market arenas (e.g. plant location decisions, local component sourcing levels, or product pricing).² The balance of non-market and market *activities* in a non-market strategy depends on precisely whom in the policy-making process the firm wishes to influence; elected legislators, for instance, will have an interest in firm actions that directly improve their own welfare (e.g. campaign contributions) as well as in those that affect their voter-constituents (e.g. local employment levels, prices of widely consumed goods or services). Appointed regulators and agency officials, who are charged with implementing legislative directives, on the other hand, are more likely to respond to non-market strategies that are primarily focused on lobbying activities and the provision of information on policy options and policy consequences.

The relative importance of non-market compared to market strategies varies, depending on the actual or potential contribution of each to firm profitability. In some sectors of the economy, governments have a deep and direct impact on average profitability and on the distribution of profits among firms, for example in natural monopoly industries (water, electricity, gas) where entry, pricing and investment decisions are controlled frequently by regulatory agencies. In this and other cases where performance is determined more by government policy than by market forces, firms will naturally place a greater emphasis on designing non-market rather than market-based strategies. Lack of direct government influence in a market place, however, does not necessarily imply that non-market strategies play a reduced role. Indeed, in highly competitive industries such as consumer electronics, textiles and agriculture, the ability to obtain favorable public policies (e.g. import tariffs or quotas, direct producer subsidies) can be an important source of industry or firm-level rents. Irrespective of the degree of market competition, therefore, individual firm or industry-wide non-market strategies have the potential to improve overall performance.

The amount of resources that firms devote to non-market strategies depends not only on the expected policy gains as discussed above, but also on the costs of implementing the strategy, and hence on the expected financial returns from non-market investments. A variety of factors determine the supply cost of policy, including the "distance" between the firm's favored ideal and the status quo

policy, the degree of interest group competition or organized lobbying against the firm's proposal and the ideological preferences of the salient policy-makers. These costs vary across issues and also over time as successive generations of policy-makers with differing preferences come into office. In general, firms will invest in non-market assets up to the point where the marginal benefit, in terms of increased revenues, is equal to the marginal cost.

CURRENT APPROACHES TO NON-MARKET STRATEGY

Despite widespread managerial and academic acceptance that the non-market component is an important and complex factor in the design of business strategy, the non-market strategy literature has largely focused on one dimension, specifically the relationship between interest groups and elected legislators, and on the ways in which interest groups strive to influence legislative outcomes. Although public policies are frequently determined by non-legislative institutions, i.e. by regulatory agencies and courts through administrative rulings and judicial decisions – interactions between interest groups and these institutional players have received less attention in the existing non-market strategy literature. In addition, the non-market strategy implications of how different branches of government (i.e. executive, legislative, judicial and administrative branches) interact in the broader policy-making process have not been fully developed. Before returning to and elaborating on this gap in the literature, we briefly review the current research, paying close attention to the underlying models of the public policy process.

A small but growing body of research, both theoretical and empirical, has emerged over the last two decades exploring the strategies that interest groups employ to shape legislative decisions. The intellectual foundations for this research were developed in the economics field by Stigler (1971) and Peltzman (1976) during the 1970s with the "Economic Theory of Regulation" which questioned the traditional assumption of government behavior, namely that elected officials, or policy makers in general, act in ways that automatically maximize social welfare. Instead, elected legislators, who, as rational actors are assumed to pursue re-election goals, exchange regulatory policy favors, e.g. votes on legislative acts, in return for pecuniary or non-pecuniary resource transfers from interest groups. Theoretical attention has largely focused on understanding the conditions under which interest groups are likely to demand policy favors, through whatever means, and a number of empirical studies have attempted to identify, with mixed success, a relationship between interest group characteristics (e.g. size and concentration) and ultimate regulatory policies (Kalt

& Zupan, 1984; Pashigian, 1984; Nelson, 1982; Nelson & Roberts, 1989). Although the empirical evidence is not fully conclusive, the theoretical premise of this line of work, by applying the rational actor model to the public sector, has significantly shaped the direction of political economy research and has prompted a more critical approach to understanding the determinants of government policies.

The influence of the rational approach to business-government relations is apparent in the managerial scholarship which, in recent years, has turned its attention to the mechanisms and organizational strategies by which interest groups actually influence legislative outcomes. Here, electoral campaign contribution ("vote-buying") strategies dominate the literature: how much should firms or collective interest groups contribute to legislators (Austen-Smith, 1995; Grier et al., 1994; Yoffie, 1987), if at all, and if so, to which legislators and at what stage during the electoral cycle (Krehbiel, 1999; Romer & Snyder, 1994; Snyder, 1990, 1991; Stratmann, 1992, 1998)? The more useful lessons for managers in the design of such strategies derive from those studies that draw on the structural characteristics of the legislative process. Snyder (1990), for example, argues that since the passage of legislation requires only a simple voting majority, firms should concentrate their resources on legislators who are slightly opposed to the proposed policy change rather than those strongly in support or strongly against. Krehbiel (1999) generalizes Snyder's argument by demonstrating that firms should target not only financial but also informational resources at "pivotal" legislators. Austen-Smith develops this approach further by considering the lobbying implications of the role that gate-keeping committees play in the broader legislative procedure within the United States. Since House and Senate committees have agenda-setting powers within their policy domains, firms have the option to transfer lobbying resources from floor members to committee members and to shape legislative results at the agenda stage in addition to, or instead of, the final voting stage.

Although the political action literature offers some interesting insights into the design of effective strategies for shaping legislative outcomes, the literature as a whole remains somewhat narrow in its scope of study. First, there is an empirical emphasis on analyzing political strategies through firms' choices over *inputs*, especially campaign contributions, rather than through the impact on *outputs*, such as on final public policy decisions or even firm or industry financial performance. While data are usually more readily available on inputs than outputs, a natural drawback of this approach is that it becomes impossible to assess empirically the effectiveness of particular strategies. The question of whether, or under which conditions, firms should devote limited resources to non-market strategies therefore remains moot. A second bias exists, at least

empirically, in the concentration on campaign contributions or financial inducements, as interest groups' primary vote-buying mechanism. In practice, lobbying, or the provision of information about policy alternatives and outcomes, is an alternative means by which interest groups can shape legislative preferences, especially when legislators have imperfect information about policy proposals. Relatively few empirical studies, however, have been conducted on lobbying strategy, again reflecting the relative cost of collecting appropriate data.³ The insights on lobbying strategies stem mainly from a small theoretical literature that has primarily concentrated on the informational and competitive interest group conditions under which lobbying is most effective (Austen Smith, 1993; Austen Smith & Wright, 1994, 1996). As such, there has been little integration between the campaign contribution and lobbying strands of the literature, leaving basic questions about relative efficacies and optimal firm choices over each activity wide open for future research.

A more important feature of the broader non-market strategy literature's focus lies in the "dyadic" assumption about the nature of the public policy process with which firms, or interest groups more generally, interact. The implicit assumption in many non-market studies is that public policy is determined primarily by the legislature in isolation from other policy institutions, thereby allowing the researcher to focus on non-market strategy formulation in the context of a dyadic relationship between firms and legislators. The underlying model of the policy process is therefore of legislatures promulgating policy decisions independently of other players, i.e. agencies and courts.

In practice, however, public policies are frequently established not by legislatures but by regulatory agencies and the judicial system without direct legislative action (see Fig. 1). Agencies typically have powers to promulgate rules and to arbitrate on specific disputes within their policy domain, and courts similarly have powers to reverse agencies or to strike down unconstitutional legislation. Consequently, this naturally raises some basic questions, which so far have not been fully addressed, about non-market strategy: for example, when and how should firms lobby agencies or use the courts in order to affect public policy outcomes, and what are the implications for strategies targeted at legislators?

In this paper, we seek to build on the current literature by explicitly bringing in agencies and courts as additional players in the policy process and by exploring some of the implications for the design of non-market strategy. By so doing, we intend to develop a more integrated approach to non-market strategy that illustrates why firms lobby a particular branch of government, conditional on how that branch relates to the rest of the government. Admitting additional institutional players into the analysis of non-market strategy requires an understanding of how legislatures interact with agencies and courts, and

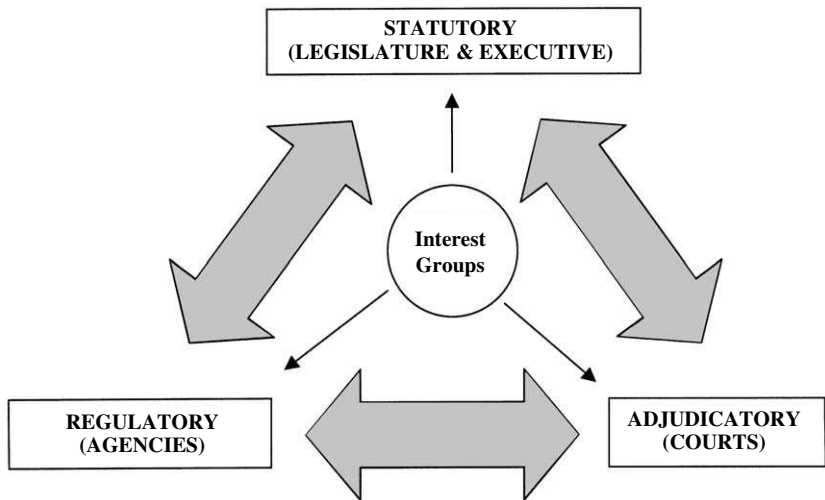


Fig. 1. Schema of Public Policy Arenas.

how these interactions result in public policy decisions. Understanding the behavioral model of players in the non-market arena is thus the first step in crafting non-market strategies designed to improve a firm's position.

In order to develop a model of non-market strategy formulation that integrates regulatory agencies, courts and legislatures, we draw on the Positive Political Theory (PPT) literature, a body of research that explicitly recognizes the interdependencies among institutional players in the policy process. As a branch of the political science field, PPT is centrally concerned with understanding the organizational arrangements of government and the implications for the design of public policy. One feature of the literature, which makes it especially attractive here, is that it uses game-theoretic models to analyze the interplay among executives, legislatures, agencies and courts, each of which is assumed to act rationally in a well-defined manner. Rather than provide an overview of this literature here, we instead describe several insights about the broader policy-making process that have emerged from PPT research and briefly illustrate initial implications for non-market strategy.⁴

(a) Agency – Legislature Interactions

Legislatures that delegate policy authority to expert administrative institutions (i.e. agencies) often build in a variety of safeguards to ensure that agency decisions do not stray too far or too fast from the legislature's ideal position

(Holburn & Vanden Bergh, 2001; McCubbins & Schwartz, 1984; McCubbins, Noll & Weingast, 1987, 1989; Vanden Bergh, 2000). Safeguards can take the form of procedural requirements that allow selected interest groups to participate in agency decision-making procedures, that alert legislatures to an agency's intention to make a ruling in the future or that leave key staff appointment decisions with the legislature. Legislatures are also able to influence agency decisions by monitoring activities through specialist committees that can ultimately initiate legislative bills to override the agency or who can conduct public inquiries. One consequence of these safeguards is that, if they substantively constrain the agency – for example, if the threat of ex post sanctions is strongly credible – then agency decisions will track the preferences of the legislature (Ferejohn & Shipan, 1990; Weingast & Moran, 1983). Thus, in this case, even though the agency is actively formulating public policy, it is doing so under the watchful eye of the legislature. In other instances, when the procedural and ex post constraints are less “binding”, the agency will have greater latitude to set policy more in accordance with its own preferences than with those of the legislature.

From a non-market strategy perspective, one implication of adopting a game-theoretic model of agency behavior is that, under certain conditions, firms have the option to influence agency decisions by lobbying the legislature rather than the agency directly. The conditions under which such a strategy is optimal will naturally depend on the agency's implicit freedom to make rulings that transgress from the legislature's ideal without incurring ex post punishment as well as on the relative cost to the firm of lobbying key legislators. More generally, the analysis suggests that firms can potentially substitute resources between lobbying agencies and legislatures and that the expected returns, in terms of policy gains, of lobbying one institution or the other will depend on the relative alignment of their preferences.

(b) Agency – Court Interactions

Regulatory agencies operate not only in the shadow of the legislature, but also under the threat of reversal by appellate and supreme courts, adding a further layer of strategic complexity into the public policy process. Reviewing courts may overturn agency decisions either on the basis that they did not possess the requisite statutory authority to set the particular policy or else on the grounds that they failed in their reasoning processes to sufficiently justify the new policy (Tiller, 1998; Tiller & Spiller, 1999; Spiller & Tiller, 1997). Agencies, looking ahead to the likelihood of future appeals, will therefore select policies that maximize their utility net of the expected costs of being overturned or of being

forced to reconsider their position. In the extreme case, this may imply that, in the presence of an opposed court, the agency chooses not to adjust the status quo when in more favorable circumstances it would in fact do so.

That firms can use litigation to challenge agencies and to gain favorable policy outcomes is a well-known aspect of non-market strategy. Local telecommunications firms in the U.S., for example, have frequently delayed the full onset of competition in their home territories by relying on lengthy and complex court cases to deter or hinder potential new entrants. MCI, for example, successfully sued the FCC and reversed its prior rulings which then allowed MCI to provide extended long-distance telecommunications services in direct competition with the incumbent, AT&T. In these and many other cases, interest groups have used the judicial process to reverse, modify or postpone the implementation of agency decisions in order to advance their competitive advantage.

Less obvious, however, are the implications from the analysis for the design of lobbying strategies. First, the threat of litigation by opposing groups suggests that the expected benefits to the firm of lobbying agencies are reduced as compared to the myopic approach to lobbying strategy that ignores interactions with courts and legislatures. Secondly, firms can potentially use lobbying and litigation as partial substitute strategies. Establishing a credible reputation as an aggressive and successful litigant, for instance, may serve to moderate initial agency decisions in the same manner as directly lobbying agency staff. Indeed, de Figueiredo and de Figueiredo (2001), in a model of two competing interest groups, find that under the right conditions, the existence of litigation completely eliminates lobbying activities.

(c) Court – Legislature Interactions

In the same way that agencies make rulings with an eye towards the threat of future override by legislatures or courts, recent scholarship suggests that such strategic behavior is also present in the judicial system (Gely & Spiller, 1990; Spiller & Gely, 1992; Vanden Bergh, 2000). Rather than acting purely on legalistic considerations (i.e. legal precedent) courts have been found to make decisions also on the basis of ideological and political factors. One study, for example, finds that while the estimated ideological preferences of the Supreme Court were an important factor in predicting the nature of labor union case decisions during the 1949–1988 period, court decisions were also sensitive to Congressional preferences (Spiller & Gely, 1992); with more ideologically conservative legislative committees, even liberal courts adjusted their pattern of decisions, presumably in order to avoid the possibility of

legislative overrule. These findings thus support the rational actor model of judicial behavior and the proposition that justices are politically sophisticated – that is, justices appear to consider the future political game that follows their rulings.

For firms aiming to influence regulatory policy through litigation, forum shopping decisions, as well as decisions to bring a case to court or not, will be informed by the ideological preferences of the legislature and of the court on the issue at hand. Even justices that have a historic record that is sympathetic towards the firm's case will be less willing to overturn agency rulings knowing that a strongly-opposed and active legislature could enact legislation reversing them. Similarly, firms may wish to delay or bring forward litigation strategies if expected changes in the political regime (for example, following member turnover in legislative committees) are likely to modify subsequent judicial behavior.

As a cumulative body of research, the PPT literature demonstrates, both theoretically and empirically, how public policies are determined by the structured interactions of self-interested government actors, namely executive, legislative, administrative and judicial bodies. Although a few studies have incorporated interest groups as additional institutional actors, these have typically done so in the context of a simplified version of the policy process (e.g. de Figueiredo et al., 1999; Spiller, 1990).⁵ Krehbiel (1999), for example, argues that firms should lobby “pivotal” actors, though the scope of analysis is confined to the legislature and pivotal legislators. Our goal here is to build on these studies by drawing on the structured-interaction approach of the PPT literature and by developing insights into the design of non-market strategy when we introduce firms as an additional actor into the broader policy process. In particular, we wish to understand how firms decide which branch(es) of government to lobby in order to gain favorable public policy outcomes.

A PPT FRAMEWORK FOR NON-MARKET STRATEGY FORMULATION

In this section we develop a theoretical framework in two parts. In the first part we model the interaction among the firm, the legislature, the executive and the agency in charge of implementing and enforcing regulation in the context of a complete information game, illustrating the strategic interaction among elected and appointed officials in determining policy outcomes. As discussed earlier, in large part, such strategic interactions are not assessed in the extant non-market strategy literature. More importantly, we explore the consequences of such interactions for non-market strategy and the allocation of lobbying

resources across different branches of government. We find that the firm's strategy and decisions about whom to lobby in order to affect policy outcomes depend on the type of political environment. In the second part of the framework we extend the initial game by introducing courts into the analysis. Doing so creates additional strategic challenges and opportunities for the firm, and modifies the conclusions reached in the first part.

In the initial game, there are four players, the firm, the legislature, the executive and the regulatory agency.⁶ The legislature is assumed to be bicameral with two relevant players, the House and Senate. Since the House and Senate committees with jurisdiction typically have control over policy offered on the floor of the legislature we simplify the analysis by assuming that the ideal point of the joint conference committee represents the bargaining outcome between the House and Senate.^{7, 8} In addition, due to the existence of the executive with veto power, the override majority (two-thirds of all members) in the House and the Senate is also important. The executive is assumed to incur political costs if a veto is overridden and will therefore only veto bills that are certain to be sustained.⁹ The regulatory agency is assumed to be a unitary actor whose preferences reflect those of the executive and legislature due to the appointment process.¹⁰

We assume that each political actor and the firm are rational and have well-defined policy preferences. The policy preferences of the political actors reflect the interests of relevant constituents and interest groups that determine their electoral success.¹¹ As such, the political actors maximize utility, which is assumed to be linear, symmetric and single peaked. We further assume perfect and complete information.

The sequence of the game is as follows. In the first stage, the firm decides where to allocate its lobbying resources in order to shift policy outcomes closer to its ideal point, F . Here the firm must choose between lobbying the agency, the executive, the conference committee or the veto supermajority point of the legislature. We assume that lobbying by the firm – either through the transfer of resources or through the provision of new information on policy alternatives or impacts – has the effect of marginally shifting the ideal point of the relevant political actor. In the second stage the agency (A will represent the ideal point of the agency) makes a rule along the relevant dimension. This decision establishes the status quo (x_0). The agency is assumed to make its decision on a continuum.¹² In the third stage, the relevant committees decide whether to offer an alternative policy (x) to overturn x_0 .¹³ As mentioned, this decision reflects the constraints imposed by the legislature, mainly that it must pass in both houses, and represents the post-bargaining outcome between the House and Senate committees. CC , for conference committee, represents the

committee's ideal point. In the fourth stage the executive (E) decides whether to sign or veto x . A veto reflects executive preference for x_0 vs. x . In the fifth and final stage, the maximum of the House and Senate two-thirds majority ($M_{2/3}$) determines whether the veto is overridden or sustained. An override reflects maximum super-majority preference for x vs. x_0 .

To derive the equilibrium, the game is solved by backward induction. Subgame perfection is assumed throughout. Each actor attempts to maximize utility subject to the constraints imposed by the institutions. That is, actors are sophisticated and recognize the effect of their actions on the subsequent play of the game. Since we assume that utility is linear and single-peaked, a policy that is closer to an actor's ideal point is preferred to one further away.

A policy equilibrium is assumed to be pareto efficient in the sense that there is no incentive to alter the policy because any alteration will result in at least one relevant actor being made worse off. Here, the *core* of the game, consisting of all pareto efficient policies, lies in the region between the conference committee, CC, and the veto super-majority point of the legislature, $M_{2/3}$. Thus, any policy ruling outside the core will be unstable in that it will be unacceptable to both CC and $M_{2/3}$ when compared to an alternative within the core – in which case CC will propose an alternative, x , within the core and $M_{2/3}$ will sustain it, if necessary overriding an executive veto. Precisely where in the core the equilibrium policy lies will depend on the relative configuration of preferences of all the players.

In each stage, the relative position of the actors determines the equilibrium. For expositional purposes, the model analyzes the implications of the conference committee, CC, being to the left of the legislative supermajority, $M_{2/3}$, in the policy space (a more liberal supermajority implies similar qualitative results). See Fig. 2. We also assume that the executive is relatively conservative compared to the supermajority, lying to the right in the policy space. Accordingly, the maximum of the House and Senate two-thirds majority is the relevant veto override point. Thus, at the fifth stage, $M_{2/3}$'s decision to override or sustain an executive veto depends on its relative position to the status quo, x_0 , and the conference committee's alternative, x . Since this is the final stage of the game the supermajority does not act strategically and simply chooses the action which immediately improves its utility; it will override the veto and restore x if x is closer to its preferred point than x_0 , otherwise it will sustain the veto and x_0 . The supermajority's preferred set consists of all x policies that are closer to its ideal point than the status quo.

In the fourth stage, the executive decides whether to veto x . Re-emphasizing the assumption that there is a substantial cost incurred if a veto is overridden (without offsetting benefit), the executive will only veto bills that the

supermajority will sustain (i.e. not override) and that consequently also leave the executive better off (i.e. x_0 is closer to the executive's ideal point than x). If either of these conditions does not hold, the executive will not veto the conference committee's alternative bill. For example, the executive will not veto bills that are in the supermajority's preferred set, irrespective of whether they make the executive worse or better off.

In the third stage, the conference committee will propose policy x subject to the constraints imposed by an executive veto and by a veto override by the supermajority. As such, CC will offer x that maximizes its utility and which avoids veto by E, by selecting the position in the supermajority's preferred set that is closest to CC. Therefore, given that $CC < M_{2/3} < E$, x will be just inside the left hand boundary of the preferred set. Here, the executive has no incentive to veto x since x is (weakly) closer to CC than x_0 , a veto would be over-ridden by the supermajority, and the conference committee cannot propose any other policies which simultaneously satisfy these conditions and which make it better off.¹⁴

In the second stage of the game, the agency makes a rule and establishes x_0 , recognizing the subsequent actions of CC, E and $M_{2/3}$. The rule depends upon the position of the agency (A) relative to the other players. In each case, A maximizes its utility subject to the future play of the other actors. Since A wishes not to be overturned in subsequent moves, it establishes the equilibrium of the game. It is helpful at this point to distinguish between several political regimes, defined by alternative preference configurations, in order to illustrate how the equilibrium policy outcome depends on relative preferences (see Fig. 2). In Regime 1, the agency is relatively liberal compared to the conference committee (which, as before, is liberal compared to the supermajority point).¹⁵ Thus, when $A < CC < M_{2/3} < E$, A makes a rule equal to CC's ideal point, since any $x_0 < CC$ will be replaced by $x = CC$ which is preferred by CC and $M_{2/3}$.¹⁶ In Regime 2, the agency is more moderate, lying between the conference committee and the supermajority on the policy dimension ($CC < A < M_{2/3} < E$). In contrast to Regime 1, the agency enjoys greater policy-making discretion in this situation since the conference committee and supermajority are "pulling" the agency in opposing directions. The agency is able to establish its ideal point as the equilibrium since any attempt by CC to propose a policy alternative closer to its ideal will be vetoed by the executive and sustained by the supermajority, thereby reasserting the agency-determined status quo. In Regime 3, where the agency is conservative relative to the legislative supermajority (i.e. $CC < M_{2/3} < A < E$), the agency is unable to obtain its ideal point as an equilibrium since, as in Regime 1, it lies outside the core. The best it can obtain, without triggering a sustainable alternative, is to set x_0 at the right hand boundary

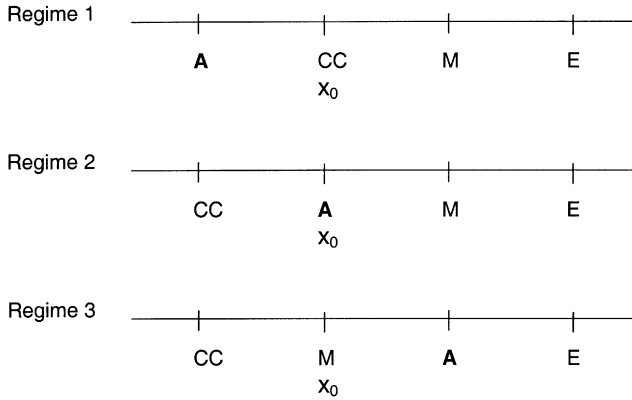


Fig. 2. Political Regimes and Policy Equilibria (x_0).

of the core, that is at $M_{2/3}$'s ideal point. To understand this, suppose that in fact the agency mistakenly set x_0 by an amount δ to the right of $M_{2/3}$. The conference committee could then propose an x equal to $M_{2/3}$ minus δ , which CC and $M_{2/3}$ would both prefer to x_0 , and which $M_{2/3}$ would sustain after an executive veto. $M_{2/3}$ minus δ would then become the equilibrium policy, making the agency worse off. We can see that the best the agency can achieve in this situation is to set x_0 equal to $M_{2/3}$ since CC cannot propose an alternative which CC and $M_{2/3}$ both prefer. This is therefore a stable equilibrium.

To summarize, in a complete information environment, the agency establishes the equilibrium through the initial rule x_0 subject to the constraints imposed by the existing institutional structure. Here there is no incentive for CC or E to alter x_0 as any change will make at least one party worse off. One can see that the equilibrium reflects the preferences of the electorate, the relative position of the actors within the institution as well as the institutional rules determining the play of the game.

We are now able to derive some propositions about the design of non-market strategies by analyzing the firm's decision in the first stage. We consider the firm's decision about which political actor to lobby in order to achieve a favorable change in policy equilibrium outcomes promulgated by the agency in the context of each of the three regimes outlined above. Beginning with Regime 2, where the agency is relatively moderate compared to the conference committee and supermajority, the firm will directly lobby the agency in an effort to move the agency's ideal point closer to the firm's. Since the agency has free reign to set policy at its own ideal point without the threat of legislative

override, any changes in its preferences, as induced by lobbying activities, will translate into corresponding changes in policy. However, when the agency is liberal relative to the committee and the veto override majority (Regime 1), directly lobbying the agency is a fruitless task: remember that when $A < CC < M_{2/3}$, the agency is effectively constrained to set the status quo at CC. Thus, even though the firm may succeed in moving the agency's ideal point, there will be no matching effect on the agency's equilibrium policy decision. Instead, the firm gains purchase by lobbying the legislative committee (CC) with jurisdiction over the policy issue since this will induce a shift in the agency's ruling in a bid to avoid CC proposing an alternative at its ideal. Similarly, in Regime 3 ($CC < M_{2/3} < A$), as in Regime 1, the firm moves policy not by lobbying the agency, but by lobbying the constraining institutional player, in this case the legislative supermajority congressman, $M_{2/3}$. Here, the agency establishes the equilibrium at the supermajority's ideal so any move in this ideal point will induce a subsequent shift in the agency's decisions as well.

Proposition 1: In the presence of complete information, a firm's decision about which branch of government to lobby depends on the type of political regime, as defined by the relative policy preferences of the political actors.

- A: If the agency is liberal relative to the committee and the veto override majority, then firms should allocate resources to changing the preferences of the *legislative committee* with jurisdiction over the policy issue (Regime 1).
- B: If the agency is moderate relative to the committee and the veto override majority, then firms should allocate resources to changing the preferences of the *agency* (Regime 2).
- C: If the agency is conservative relative to the committee and the veto override majority, then firms should allocate resources to changing the preferences of the relevant elected legislators likely to determine whether a veto can be sustained.

Further strategic implications may be derived once we recognize that agency preferences are not entirely independent of those of the legislature and executive, as has been implicitly assumed so far. U.S. agency heads or commissioners are typically appointed by the executive subject to the consent of the legislature, so the preferences of agency appointees are likely to reflect a combination of the appointing legislature's and executive's preferences. Firms thus have an opportunity to influence the nature of the agency's ideal position on any given policy issue by directly lobbying for the appointment of agency officials who are sympathetic to the firm's position. The above "regime" analysis, however, suggests that such lobbying activities will be more effective, in terms of ultimately affecting policy decisions, in some political regimes than others. For instance, in Regimes 1 and 3, where the agency is effectively

constrained by the legislative committee and supermajority legislator, respectively, inducing a shift in the agency's ideal point will have no impact on the agency's policy ruling. When the agency has some latitude, on the other hand, as in Regime 2, changes in agency preferences that arise from firm lobbying over appointment decisions, will translate into corresponding changes in policy. In this case, the firm has a strong incentive to lobby the executive and legislature to appoint their preferred agency officials. More generally, the wider the distance between CC and $M_{2/3}$, the more likely that lobbying over agency appointments will matter.

As discussed above, agency rules are subject to reversal by appellate courts. Reviewing courts may overturn agency decisions in favor of an existing status quo policy because an agency ruling did not possess statutory grounds or was not reasonable (Tiller, 1998; Tiller & Spiller, 1999; Spiller & Tiller, 1997).¹⁷ At the same time, legislatures may overturn court decisions through the statutory process. Thus, forward-looking justices incorporate the threat of a statutory response into their decision-making calculus (Vanden Bergh, 2000; Spiller & Gely, 1992; Gely & Spiller, 1990). Similarly, far-sighted agencies will promulgate rules taking into account the threat that the courts can overturn their decision in favor of the status quo (Spiller, 1992). As such, to create a more complete framework, we must explicitly incorporate the courts into our analytical framework. By explicitly recognizing the strategic interaction among agencies, courts and legislatures, firms can identify where lobbying activities are likely to have the greatest affect in terms of changing policy outcomes.

AGENCY-COURT-LEGISLATURE-EXECUTIVE-FIRM MODEL

With the introduction of courts into the general framework, there are now five players in the game, the firm, the legislature, the executive, the regulatory agency and the courts. Our assumptions about players' preferences from the initial game do not change, although the order of play now alters.

The sequence of the game is as follows. In the first stage, the firm decides which branch of government to lobby in order to shift policy outcomes closer to its ideal point, F. Here the firm must choose between lobbying the agency, the executive, the conference committee or the veto supermajority point of the legislature.¹⁸ In the second stage the agency (A represents the ideal point of the agency) makes a rule along the relevant dimension. This decision either reaffirms the status quo (x_0) or establishes a new rule, x , through the rulemaking process. In order to simplify the analysis we assume that the status quo resides at the conference committee's ideal point, CC. In the third stage, the court may

uphold the agency's rule x or overturn it in favor of the status quo. In the fourth stage the relevant committees decides whether to offer an alternative policy (x_a) to overturn x if the court upholds x .¹⁹ In the fifth stage the executive decides whether to sign or veto x_a . In the sixth and final stage, the maximum of the House and Senate two-thirds majority ($M_{2/3}$) determines whether the veto is overridden or sustained. An override reflects maximum super-majority preference for x_a vs. x .

As above, the equilibrium ruling by the agency is determined by solving the game via backward induction. In each stage, the relative position of the actors determines the equilibrium. For expositional purposes, the model analyzes the implications of the conference committee, CC, being to the left of the legislative supermajority, $M_{2/3}$, in the policy space (a more liberal supermajority implies similar qualitative results). We also assume that the executive is relatively conservative compared to the supermajority, lying to the right in the policy space. We thereby explore the affects of varying agency (A) and court (J) preferences on the allocation of lobbying resources by the firm. Figure 3 illustrates the setup.

Stages four, five and six are equivalent to stages three, four and five in the no-court game. That is, the supermajority will override the veto and restore an alternative statute x_a if x_a is closer to its preferred point than x , otherwise it will sustain the veto and restore the agency ruling x . The supermajority's preferred set consists of all policies that are closer to its ideal point than the agency's ruling x . In stage five, the executive decides whether to veto x . The executive will only veto bills that leave the executive better off (i.e. the agency's ruling x is closer to the executive's ideal point than the alternative policy proposed by CC, x_a). If either of these conditions does not hold, the executive will not veto the conference committee's alternative bill. In the fourth stage, the conference committee will propose an alternative policy x_a subject to the constraints imposed by an executive veto and by a veto override by the supermajority. As such, CC will offer x_a that maximizes its utility and which avoids veto by E, by selecting the position in the supermajority's preferred set that is closest to CC. Therefore, given that $CC < M_{2/3} < E$, x_a will be just inside the left hand boundary of the preferred set. Here, the executive has no incentive to veto x_a . Since x_a is (weakly) closer to CC than x , a veto would be over-ridden by the supermajority, and the conference committee cannot propose any other policies which simultaneously satisfy these conditions and which make it better off.

In the third stage of the game, the court decides either to support the agency's rule or to overturn it in favor of the status quo policy x_0 . The decision by the court depends upon its ideal policy position relative to the other players. In all environments, the court maximizes its utility subject to the future play of

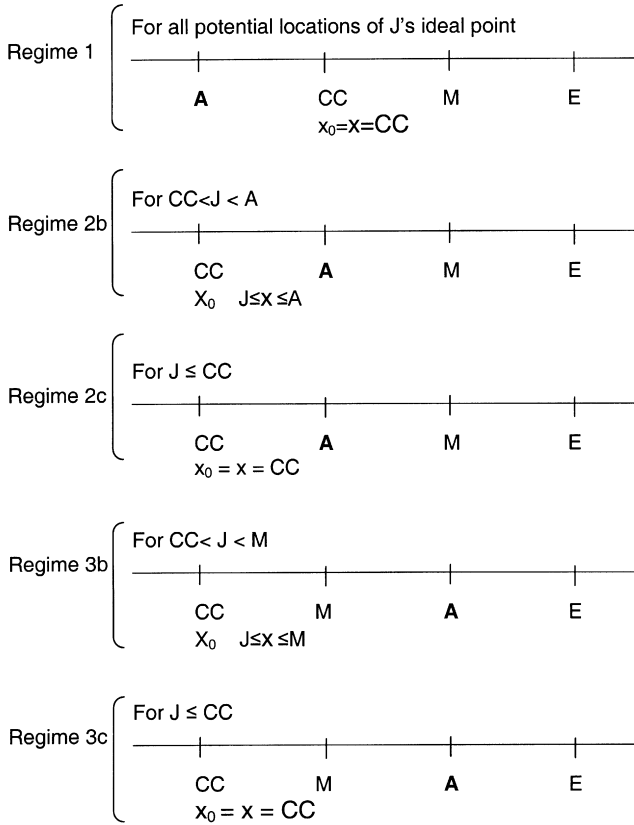


Fig. 3. Policy Equilibria in Political and Judicial Regimes.
 (Status quo x_0 is at CC and x is the equilibrium ruling by A).

the other actors. Since x_0 is assumed to be equal to the ideal point of the conference committee, the court knows that if it overturns the agency's rule x , then x_0 will be the final outcome of the game due to future actions in stages four through six. The court also knows that if the agency's rule x is an element of the core, then x will also be a stable outcome. Assuming that the ruling by the agency is an element of the core, then the court will uphold x if x is closer to the court's ideal point than x_0 , otherwise the court will overturn x . Theoretically, the agency could make a ruling x that is not an element of the core. This ruling, if upheld by the court, will lead to a statutory response by the legislature since the conference committee and the legislative supermajority

will be able to pass an alternative x_a that makes them each better off as compared to the agency's ruling. In this environment, the court will uphold the agency's rule x if the court expects the alternative statute (x_a) to be closer to its ideal point than the status quo, x_0 . Otherwise the court will overturn x in favor of the status quo.

In the second stage of the game, the agency makes a rule and establishes x , recognizing the subsequent actions of the court, the conference committee, the executive and the legislative supermajority. The rule depends upon the position of the agency relative to the other players. Since A can solve for subsequent moves, it establishes the equilibrium of the game. Given the addition of the courts, it is helpful at this point to distinguish between several political regimes, defined by alternative preference configurations, in order to illustrate how the equilibrium policy outcome depends on relative preferences (see Fig. 3). In Regime 1, the agency is relatively liberal compared to the conference committee (which, as before, is liberal compared to the supermajority point). No matter where the court's ideal point resides in the policy space, when $A < CC < M_{2/3} < E$, A makes a rule equal to CC 's ideal point, since any $x < CC$ will be overturned by the court in favor of $x_0 = CC$, which is preferred by CC and $M_{2/3}$.

In Regimes 2, the agency is more moderate, lying between the conference committee and the supermajority on the policy dimension ($CC < A < M_{2/3} < E$). With the introduction of the court, the agency is no longer able to establish its ideal point as the equilibrium policy in all cases. The agency must now consider the court's ideal point when promulgating rules. This is best illustrated in Regimes 2b and 2c.²⁰ Since the conference committee has jurisdiction to establish new policy in this area (x_0 or x_a), and the court can overturn agency rules in favor of x_0 , the agency's policy choice will move away from its ideal point closer to the conference committee's, depending upon the preferences of the court. If the court is moderate relative to the conference committee and agency (Regime 2b), then the initial policy choice of the agency will reside at the point in the policy space that makes the court indifferent to x and the status quo. In this way, the agency's rule is supported by the court and does not get overturned by the conference committee through a statutory process. As the court moves closer to the conference committee's ideal point, the agency's policy decisions also move closer to CC . Interestingly, for a relatively liberal court ($J < CC$), the agency makes a ruling equal to CC (Regime 2c). While on the surface the conference committee appears to be the critical political actor in this regime, it is the presence of the court and the threat of the court overturning the agency's policy choice that constrains the agency.

In Regimes 3, we observe a qualitatively similar dynamic. When the court is relatively conservative, the policy outcome is the same as in the no-court game above (Regime 3 in Fig. 2).²¹ The agency is able to choose a policy $x = M_{2/3}$ and avoid a judicial overrule or statutory response. As the court becomes relatively liberal, however (Regimes 3b and 3c), the agency must respond to the threat of the court overturning their ruling in favor of the status quo. Again, the presence of the court moves agency policy closer to CC as the court becomes more liberal. In the extreme case where the court is liberal relative to all other political actors, a conservative agency (Regime 3c) will make a relatively liberal ruling ($x = CC$) since it recognizes that the court can overturn their decision in favor of the status quo set by the conference committee. To understand this, suppose that in Regime 3c the agency mistakenly set x by an amount Δ to the right of CC. The court could then overturn x in favor of $x_0 = CC$. The legislature will not attempt to rewrite the law since the committee has no ability to write a statute that improves its utility. We can see that the best the agency can achieve in this situation is to set x equal to CC.

To summarize, in a complete information environment, the agency establishes the equilibrium through the initial rule x subject to the constraints imposed by the existing institutional structure. Here there is no incentive for the court, the conference committee or the executive to alter x as any change will make at least one party worse off. One can see that the equilibrium reflects the preferences of the relative position of the actors within the institution as well as the institutional rules determining the play of the game.

We are now able to derive additional propositions about the design of non-market strategies by analyzing the firm's decision in the first stage. We consider the firm's decision about which political actor to lobby in order to achieve a favorable change in equilibrium policy outcomes promulgated by the agency in the context of each of the regimes outlined above. The threat of the court overturning an agency rule does not affect the firm's behavior in Regime 1. For any ideological preferences of the court, the firm will lobby the conference committee (CC) in order to bring about more favorable policy outcomes. The relatively liberal agency is effectively constrained to confirm the status quo policy. Even if the firm can allocate resources to change (marginally) the preferences of any other political actor besides the conference committee, there will be no matching effect on the equilibrium policy outcome. Thus, the firm gains purchase by lobbying the legislative committee with jurisdiction over the policy issue since this will induce a shift in the agency's ruling in a bid to avoid the committee proposing an alternative at its ideal.

In Regimes 2 and 3 (see Fig. 3), the presence of the court may change the firm's lobbying decision as compared to our earlier discussion when we assumed

no judicial oversight. In Regime 2, when the agency is relatively moderate compared to the conference committee and supermajority, the firm will directly lobby the agency only if the court is more conservative than the agency. For relatively liberal courts however, the firm will only bring about change in policy by lobbying the conference committee. This is a much different result from the previous game. In Regimes 2b and 2c, the agency no longer has complete discretion to set policy because the threat of the court overturning the agency's decisions becomes credible. Thus, the agency will be careful to avoid an overruling by the court. The agency will pay close attention to the relative position of both J and CC when making policy choices. To effect favorable policy in this regime, the firm must therefore attempt to change marginally the preferences of the conference committee. In Regime 3, a similar dynamic is at play. Whereas in the no-court environment, the agency was constrained only by the supermajority veto override legislator, with the presence of the court, the pivotal political actor shifts depending on the policy preferences of the court. As the court becomes more and more liberal (movement of court's preferences closer to CC relative to $M_{2/3}$ in Regime 3, Fig. 3), the agency will need to pay closer attention to the relative position of J and CC as opposed to $M_{2/3}$. Again, to effect favorable policy in this regime, the firm must attempt to change marginally the preferences of the conference committee.²²

Proposition 2: In the presence of complete information, a firm's decision about which branch of government to lobby depends on the type of political regime, as defined by the relative policy preferences of the political actors.

- A: If the agency is liberal relative to the committee and the veto override majority, then firms should allocate resources to changing the preferences of the *legislative committee* with jurisdiction over the policy issue (Regime 1) irrespective of the policy preferences of the court.
- B: If the agency is moderate relative to the committee and the veto override majority, then firms should allocate resources to changing the preferences of:
 1. the *agency* if the court is conservative relative to the agency
 2. the *legislative committee* (Regimes 2b and 2c) if the court is more liberal than the agency.
- C: If the agency is conservative relative to the committee and the veto override majority, then firms should allocate resources to changing the preferences of:
 1. the relevant *supermajority legislators* likely to determine whether a veto can be sustained, if the court is conservative relative to the agency.
 2. the *legislative committee* (Regimes 3b and 3c) if the court is more liberal than the agency.²³

Further strategic implications may be derived once we recognize that court preferences are not entirely independent of those of the legislature and executive.

Judicial selection methods vary across political environments and reflect either broad electoral preferences or a bargaining game between the executive and legislature. Vanden Bergh (2000) shows that judicial selection rules affect judicial behavior. In some states in the U.S., court justices are re-elected through a competitive electoral process, whereas in other states they are selected through a non-competitive retention process. In the latter case, if a justice is voted out of office, then the legislature and executive appoint a replacement. Firms thus have an opportunity to influence the nature of the court's preferences by directly lobbying for the appointment of certain justices who are sympathetic to the firm's position or by supporting the campaign of a particular justice. Our "regime" analysis, however, suggests that such activities will be more effective, in terms of ultimately affecting policy decisions, in some political regimes than others. For instance, in Regime 1 (Fig. 3), where the agency is effectively constrained by the legislative committee or supermajority legislator, respectively, inducing a shift in the court's preferences will have no impact on the agency's policy ruling. When the court is a pivotal actor, on the other hand, as in Regimes 2b and 3b, changes in court policy preferences that arise from firm's allocating resources to affect appointment decisions or electoral outcomes, will translate into corresponding changes in policy. In these cases, the firm has a stronger incentive to invest in lobbying activities that improve the electoral chances of their preferred justice.

Although we have presented a simple, stylized model of the policy making and lobbying process, which ignores naturally important considerations such as interest group competition in lobbying and the existence of uncertainty about ideal points or about the costs of making policy changes, it does nonetheless provide a theoretical basis upon which future complexities can be built. Additionally, this simple framework provides new insights into the design of effective lobbying strategies. Perhaps most importantly, by explicitly recognizing the strategic nature of the interactions among political actors, firms can identify where in government their lobbying activities are likely to have the greatest impact in terms of changing policy outcomes. Thus, if the firm wishes to shift agency-determined regulatory policy in its favor, directly lobbying the agency responsible for the issue at hand need not be the most effective strategy. Deciding whether to lobby the agency or instead the political principals that oversee the agency will generally depend on their relative preferences and the nature of the decision-making game. When political actors are a binding constraint on the agency, the firm will shift its lobbying efforts towards the politicians and away from the agency, even though the agency continues to promulgate and enforce policy. This result substantially modifies the conclusions of much of the existing "Business and Public Policy Management" literature

which, absent recognition of the strategic nature of the policy “game”, typically prescribes lobbying strategies uniformly directed at *agencies* in regulatory situations (Buchholz, 1990; Baron, 1996).

EXTENSIONS

Incomplete Information

A natural limitation of the model is the assumption about complete and perfect information – each player is assumed to know the play of the game as well as the other players’ preferences. In practice, while legal constitutions provide some clarity about the jurisdictions of different players and the sequence of policy-making games, it is likely that uncertainty exists about players’ policy preferences. Agencies, for example, operating under fixed budgets, may not have sufficient resources to communicate the necessary information about policy alternatives and consequences to legislators in order to determine their ideal positions on a given issue. When agencies are uncertain about the precise location of constraining actors’ ideal points, they may mistakenly issue a policy ruling that does not fall within the core, thereby triggering a legislative response that establishes a new equilibrium. Note that in the complete information scenario, our model predicts that we should not observe *legislative* activity at all: as a rational far-sighted actor, an agency responds to changes in legislative preferences such that its ruling establishes a new equilibrium, implying that acceptable legislative alternatives will not arise. The gains realized by the firm are thus limited by the response of the agency. However, if an agency does promulgate a policy outside the core, a legislative response will occur as the conference committee takes advantage of the opportunity to move policy closer to its ideal. From the firm’s perspective, such out-of-equilibrium mistakes by the agency can ultimately be beneficial. In the right political environments, the legislative response will bring about new policy equilibria even closer to the firm’s ideal point than a pareto efficient policy decision made by the agency. This is easily seen when the firm has an ideal policy that is more liberal than the legislative committee’s, the veto override majority’s and the agency’s ideal points (as in Regime 3 in Fig. 2). In this environment, a pareto efficient policy by the agency is equal to the ideal point of the veto override legislator’s ideal policy. Yet if the agency makes a mistake by ruling outside the core (to the right of $M_{2/3}$ in Regime 3), then the legislative committee with jurisdiction will offer a statute to overturn the agency decision. This statute will be closer to the firm’s ideal policy. On the other hand, by a similar reasoning, if the firm is relatively conservative compared to the other players (the

firm's ideal is to the right of the agency in Regime 3 in Fig. 2), an agency mistake would leave the firm worse off.

A strategic implication of this analysis is that firms have an opportunity to influence policy outcomes not only by lobbying for shifts in political actors' preferences, but also by lobbying to alter the informational asymmetries that exist between players. Take, for example, the latter scenario described above: lobbying activities that ultimately result in a reduction in agency uncertainty about the supermajority's ideal point would reduce the probability of a legislative response that would leave the firm worse off than a more-informed agency ruling that is not overturned. If the agency understands the relative configuration of preferences but not their precise locations, then the firm would send an informational message to the agency about the location of the supermajority's ideal point. Since the agency knows that the firm will benefit from a correct rather than mistaken agency ruling, the agency will interpret the firm's message as credible. Such messages would be regarded as disingenuous, however, if the firm stood to gain from agency mistakes, suggesting that the firm would withhold from informational lobbying when the firm and agency are on opposite sides of the policy core.

The opportunity for a firm to engage in such informational lobbying depends on the firm having better information than the agency about other players' preferences. This could arise when agencies have limited budgets that do not allow them to sufficiently work with political actors when formulating policy. Firms, by contrast, do not operate under statutory fixed budgets and are thus able to invest in lobbying activities as long as they provide a net benefit. Thus, when agencies are at risk of making a mistaken ruling that would ultimately leave the firm worse off following legislative override, firms have an incentive to invest in learning more about politician's preferences and to subsequently lobby the agency with such information.

Competing Interest Groups

In addition to a simplified informational environment, the model also assumes that only a single private interest group (a firm or industry-wide association) is seeking to influence public actors in the policy process. Interest groups, however, frequently compete in the non-market arena with other organized interests who have opposing policy preferences (Austen-Smith & Wright; 1994, 1996). Real estate developers, for example, often must counteract powerful environmental groups in their plans to construct on greenfield sites; power generation companies commonly find that local residents, concerned about perceived pollution and health externalities, vigorously oppose the siting of new

power plants. Jaffee and Freeman (in this volume) demonstrate how divisions can exist between firms within the same industry; while some German law firms advocated the adoption of legislation enabling clients to implement employee share option plans (ESOPs), other law firms, threatened by subsequent loss of competitive advantage, opposed such proposals. In these and other similar situations, organized interests attempt to shift policy-makers' positions in opposite directions.

Although the model presented here does not incorporate multiple interest groups, it is flexible enough to accommodate them as an extension in future research. The analysis of which public player is pivotal will not change since this is determined by the institutional rules that govern their interactions. Opposing interests will thus concentrate on lobbying the same actor for preferential policy treatment. Predicting the outcome of the lobbying game with opposing interests will depend on factors such as the sequence of interest group moves and their relative resources. For each player, however, the marginal return to lobbying expenditures is likely to be reduced in the face of competition. Indeed, as Jaffee and Freeman discuss, the presence of multiple venture capital firms and law firms with opposing views on the merits of German ESOP legislation partially explained the lack of legislative or judicial decisions to change the existing status quo policy.

Political Dynamics

A further limitation of our model is that it represents a one-period game among the players. To this extent, expectations of future games, and policy outcomes, are not incorporated in the analysis of players' behavior. If firms expect future political preferences to change, however, they may take actions to prevent current agency-determined policies from shifting further away from their ideal position. Suppose, for example, that the firm anticipates that a new executive, with stronger anti-firm policy preferences, will gain office at the next election. Since agency heads or commissioners are generally appointed by the executive this would imply a rightward shift in the agency's preferences. If the agency's ideal position lies in the core and is able to set policy at its ideal, then the change in executive preferences will lead to a similar shift in the agency's equilibrium policy ruling.

To protect itself against such adverse policy shifts, the firm may lobby political actors to increase the agency's cost of changing policy, thereby insulating the status quo against future reversals or modifications. One way to do this is to lobby for structural changes in the agency's decision-making process

that restrict agency discretion on a particular aspect of policy, such as the ability to determine specific rates or competition policy in regulated industries. Indeed, there is a growing body of empirical evidence that legislatures do act in a far-sighted manner when deciding upon the appropriate level of discretion for regulatory agencies. Holburn and Vanden Bergh (2002) find that in the U.S. during the 1970s and 1980s, Democrat-controlled state legislatures were more likely to enact legislation requiring regulatory agencies to admit consumer advocates to administrative hearings when they were relatively uncertain about being re-elected. By specifying this requirement in statute, incumbent Democrat coalitions increased the prospect of the policy remaining in effect even if future political generations were expected to have opposing preferences.

Alternatively, firms may lobby for changes in the selection procedures of agency heads which effectively lock-in agency preferences for a period of time. For example, increasing the number of agency commissioners may allow an executive to make new appointments that alters the identity of the median commissioner and hence the nature of agency policies that are decided by a majority vote. Similar results may be achieved by adjusting the tenure of agency commissioners or by changing selection rules (e.g. which selecting parties have jurisdiction).

From a strategic perspective, therefore, dynamic political considerations influence the incentives of firms to lobby for changes in the relationship between political actors and agencies. While in a static setting, firms may opt to lobby agencies in order to shift agency preferences and policy rulings, they may instead lobby the legislature and executive to initiate legislation that insulates the status quo from future political generations.

This chapter has focused solely on *where* in the government institutional structure firms should allocate their lobbying efforts. Other strategic decisions relate to the *framing* of policy issues and informational arguments, and the *type* of resources to expend on lobbying. Such decisions will be informed additionally by the government location of firm lobbying activities. Dowell, Swaminathan & Wade (in this volume) argue that the appropriate framing of the rationale for High Definition Television (HDTV) technological standards was an important factor in the emergence of a powerful collective lobbying group that successfully influenced the nature of the adopted standards. The appropriate frame is likely to depend on whether the firm or industry association is lobbying legislators or agencies. Legislators are expected to respond especially to lobbying by their own electoral constituents since the latter can establish a direct personal connection to perceived future electoral fortune. Agency staff, on the other hand, will be less sensitive to partisan political considerations and more responsive to “public interest” arguments that bear on

the agency's statutory mandate. Framing issues in the context of broader societal welfare will thus be more appropriate when attempting to persuade agencies of the merits of a particular policy position.

In prescribing such lobbying strategies, this chapter has worked on the assumption that managers act in a highly calculative and hyper rational manner, correctly determining the relative advantages of alternative courses of action. As the experimental research by de Figueiredo and de Figueiredo (in this volume) suggests, however, this may be a strong assumption. As non-market strategy problems become more complex, including situations with increasing numbers of stages or players, de Figueiredo and de Figueiredo find that managers experience greater difficulty in reaching the optimal solution. The need for managers to identify pivotal agencies, legislators, legislative committees or executives, particularly under conditions of imperfect information, may thus be too complex a demand when managers are boundedly rational. If managers rarely participate in the non-market arena, the condition of bounded rationality could be extreme. Hence, while this chapter presents a clear prescriptive model for non-market strategy formulation, the costs of implementation for individual firms may be prohibitively high. One implication is that future research seeking empirical evidence for the propositions contained here should potentially focus on firms that specialize in designing and enacting lobbying strategies for clients who wish to outsource such operations. On balance, lobbying specialists are more likely to reach the optimal lobbying strategy than less experienced firms.

CONCLUSION

Cursory glances at the daily headlines of most major newspapers demonstrate the myriad of ways in which governments directly or indirectly affect firms' performance. Even in industries where government has little or no direct role, government control over the levers of macroeconomic policy (e.g. interest rates, money supply) and microeconomic policy (e.g. price structures) ensures that short-run and long-run returns to private sector investment are closely tied to government economic policy decisions. Consequently, non-market strategies that influence public policy outcomes can potentially improve firms' overall performance. In contrast to the voluminous competitive market strategy literature, however, the ways in which firms formulate and implement non-market strategy has received relatively little academic attention. In part, we suggest this research deficit is due to the lack of a theoretical framework that links non-market outcomes – i.e. public policies – to the complex interplay between political-institutional actors, such as legislatures, agencies and courts, and private sector interest groups, and that also yields empirically refutable

hypotheses. Although the recent political action literature is a considerable advance on the traditional “life cycle” approach to public policy formulation in that it explicitly models political actors as rational agents, this stream of research is almost exclusively focused on the firm-legislature relationship, leaving aside the influence of other government branches in the broader policy process.

In this chapter, we take a small step towards developing a conceptual approach to non-market strategy analysis and formulation by drawing on the Positive Political Theory literature. PPT models the policy process as a well-defined game among institutional actors who have clear objectives and action spaces. By integrating the firm as an additional strategic player into an analysis of the policy process, we derive some preliminary implications for the design of non-market lobbying strategies. Chief among these is the firm’s decision about which branch of government to lobby to achieve the greatest impact on policy outcomes. We find that, due to the strategic interactions of the political-institutional players, firms will not necessarily lobby regulatory agencies, for example, even when agencies are actively implementing and shaping policy, under conditions when the agencies are especially sensitive to the preferences of the legislature, executive or judiciary. More generally, we argue that firms need to understand the broader public policy game – as defined by the sequence of play, decision-making rules and players’ preferences – in order to identify where their lobbying activities will have the greatest leverage.

NOTES

1. For an exception, which draws on the heterogeneity across countries in national government institutions and the associated implications for lobbying strategies, see Hillman and Keim (1995).

2. We depart here from conventional definitions of non-market strategy by including market and non-market based activities in the mechanisms. Baron (1995, p. 47), for example, defines non-market strategy as “a concerted pattern of actions taken [by the firm] *in the non-market environment* to create value by improving its overall performance”, emphasis added. We focus attention on the *purpose* of non-market strategy – to influence public policy – rather than on the *means* by which it is implemented.

3. For an exception, however, see De Figueiredo and Silverman (2002).

4. For conference volumes on the topic see, Positive Political Theory and the Law, *USC Law Review* (1995); Conference. Regulating Regulation, *Journal of Law and Contemporary Problems* (Vol. 57, Winter/Spring, 1994); Conference on the Economics and Politics of Administrative Law and Procedure, *Journal of Law, Economics and Organization* (Vol. 8, 1992).

5. Spiller (1990) and de Figueiredo et al. (1999) argue that interest groups provide informational, as well as financial, benefits to legislatures, reducing the informational disadvantage that legislature’s face vis-à-vis agencies. Although these papers also

examine the relationship among interest groups, legislatures and agencies, the focus is on understanding how information asymmetries affect the incentives of legislatures to allow interest groups to lobby agencies. The paper here relates to these analyses by considering, under the perfect information scenario, the conditions under which interest groups choose to lobby agencies or legislatures.

6. The policy space is assumed to be a single dimension. Poole and Rosenthal (1990) found that one dimension captured most of the spatial information. Where an issue maps onto the dimension may change over time (and is critical to the analysis), however, they found that "economic" votes (e.g. rate regulation) tend to line up on one dimension. Snyder (1990) offers a critique to Poole and Rosenthal.

7. Since the committees are constrained by the veto prospects, this assumption does not affect the qualitative results of this analysis. Qualitatively similar results would obtain with more complex committee floor interaction at the expense of notational complexity.

8. Shepsle and Weingast (1987, 1989) argue that committee control of the policy results from their gate-keeping and ex-post veto powers. For a critique of this argument see Krehbiel (1991).

9. Relaxing this assumption would not change the equilibrium of the game. Examples of cost include, loss of reputation or loss of electorate confidence in the President's ability to lead. Certainty is possible because in this game information is assumed to be complete.

10. There is controversy in the literature over the source of an agency's preferences. For example, Bawn (1995) posits that preferences are endogenous to the designed procedures. On the other hand, Epstein and O'Halloran (1996) aver that the agency's preferences are aligned with the executive. In this paper, we follow Vanden Bergh (2000), arguing that an agency's preferences reflect a bargaining game between the executive and legislature depending on the appointment rules. In most circumstances, agency heads are appointed by the executive on the advice and consent of the legislature. In other environments agency heads are elected. In an elected state, agency preferences may be quite different from a bargaining outcome between the executive and legislature. We also consider this type of environment in our analysis.

11. Mayhew (1974) shows that the electoral connection suggests that legislators take into account the interests of voters, that is, the electoral consequences of their actions. Legislators and the Executive are therefore constrained by this connection. Regulatory agents are discussed below.

12. This is not an unreasonable assumption, as an agent often has discretion over rules on any one dimension.

13. Both x and x_0 are elements of X , the set of possible policy alternatives. X is a subset of $\mathbf{R}^1$ (the entire policy space).

14. We assume that the executive expends political capital or incurs adverse costs if its veto is overturned by the legislature.

15. This preference ordering can occur if the tenure cycle of the agency's political appointees is not aligned with that of the legislature or executive. Re-elections can generate new political coalitions with preferences differing from those of the incumbent agency – whose preferences are likely to reflect those of prior political generations – or at least until the legislature has an opportunity to replace key agency officials through the appointment process. State public utility commissioners, for example, are generally appointed for five or more years and typically remain in office beyond one political election cycle. This preference ordering can also occur when agency actors are elected to office.

16. We define $P^{M2/3}(x_0)$ as the set of policies $M_{2/3}$ prefers to x_0 . The only time that x will not be an element of $P^{M2/3}(x_0)$ is if $M_{2/3} < A < CC < E$ and A chooses $x_0 = A$. In this case $x = CC$ will be outside of $P^{M2/3}(x_0)$, and will be signed by E . However, this x_0 would not arise if one assumes subgame perfection throughout, since A will adjust x_0 to avoid CC offering an alternative policy that is preferred by E and $M_{2/3}$. That is, A will rule $x_0 = CC$ in this environment.

17. We assume throughout this paper that the court makes a decision to overturn an agency ruling in favor of the status quo. As such, the court faces a choice between two alternative policies. In some political environments, courts do have the authority to overturn an agency ruling in favor of any alternative policy. Under this type of decision rule, the court makes a decision on a continuum. The implications of this alternative decision process does not change the qualitative results in this paper, mainly that firm resource allocation decisions depend upon the nature of politics within a given institutional environment. One can, however, derive distinct hypotheses regarding resource allocation by comparing and contrasting court decision rules. We leave this for future research.

18. As before, we assume that lobbying by the firm has the effect of marginally shifting the ideal point of the relevant actor.

19. The conference committee will only act if an agency rule is different from the status quo and the rule is upheld by the Court. As above, all policy choices, x_a , x and x_0 are elements of X , the set of possible policy alternatives. X is a subset of $\mathbf{R}^1$ (the entire policy space).

20. In Regime 2a, the court's ideal point J is greater than A and the optimal ruling by the agency is equal to A . We do not show this regime in the figure.

21. In this regime (3a), the court's ideal point is to the right of M 's. The pivotal politician is thus M and the Agency's optimal ruling is located at M 's ideal point.

22. This is consistent with de Figueiredo and de Figueiredo (2001) which argues that litigation only makes sense in certain states of the world.

23. For regime 2b and 3b the exact point of the agency ruling (x) will be in a range of values depending upon the location of the status quo and the court. It will be the point at which the court is indifferent between the status quo and the agency ruling. Thus, given the configuration of preferences and the location of the status quo in 2b and 3b, the agency ruling $x = 2J - x_0$.

ACKNOWLEDGMENTS

The authors thank Brian Silverman, Witold Henisz and participants at the 2001 Columbia Strategy Workshop for helpful comments and suggestions.

REFERENCES

- Austen-Smith, D. (1993). Information and Influence: Lobbying for Agendas and Votes. *American Journal of Political Science*, 37, 799–834.
- Austen-Smith, D. (1995). Campaign Contributions and Access. *American Political Science Review*, 89, 566–581.

- Austen-Smith, D., & Wright, J. R. (1994). Counteractive Lobbying. *American Journal of Political Science*, 38, 25–44.
- Austen-Smith, D., & Wright, J. R. (1996). Theory and Evidence for Counteractive Lobbying. *American Journal of Political Science*, 40, 543–564.
- Baron, D. P. (1995). Integrated Strategy: Market and Non-market Components. *California Management Review*, 37, 47–65.
- Baron, D. P. (1996). *Business and Its Environment*. Prentice Hall, Inc. New Jersey
- Baron, D. P. (1997). Integrated Strategy and International Trade Disputes: The Kodak-Fujifilm Case. *Journal of Economics and Management Strategy*, 6, 291–346.
- Baron, D. P. (1999). Integrated Market and Non-market Strategies in Client and Interest Group Politics. *Business and Politics*, 1, 7–34.
- Bawn, K. (1995). Political Control versus Expertise: Congressional Choices about Administrative Procedures. *American Political Science Review*, 89, 62–73.
- Baysinger, B. (1984). Domain Maintenance as an Objective of Business Political Activity: An Expanded Typology. *Academy of Management Review*, 9, 248–258.
- Bonardi, J. P. (1999). Market and Non-Market Strategies during Deregulation: The Case of British Telecommunications. *Business and Politics*, 1(2), 203–231.
- Buchholz, R. A. (1990). *Essentials of Public Policy for Management*. Prentice Hall, Inc. New Jersey.
- de Figueiredo, J. M., & de Figueiredo, R. J. P. (2001). The Allocation of Resources by Interest Groups: Lobbying, Litigation and Administrative Regulation. Unpublished manuscript.
- de Figueiredo, J. M., & Silverman, B. S. (2002). Academic Earmarks and the Returns to Lobbying. Unpublished Manuscript.
- de Figueiredo, R. J. P., Spiller, P. T., & Urbiztondo, S. (1999). An Informational Perspective on Administrative Procedures. *Journal of Law, Economics, and Organization*, 15, 283–305.
- Epstein, D., & O'Halloran, S. (1994). Administrative Procedures, Information, and Agency Discretion. *American Journal of Political Science*, 38, 697–722.
- Epstein, D., & O'Halloran, S. (1996). Divided Government and the Design of Administrative Procedures: A Formal Model and Empirical Test. *The Journal of Politics*, 58, 373–397.
- Ferejohn, J., & Shipan, C. (1990). Congressional Influence on Bureaucracy. *Journal of Law, Economics and Organization*, 6, 1–20.
- Gely, R., & Spiller, P. T. (1990). A Rational Choice Theory of Supreme Court Statutory Decisions with Applications to the State Farm and Grove City Cases. *Journal of Law, Economics and Organization*, 6(2), 263–300.
- Grier, K. B., Munger, M. C., & Roberts, B. E. (1994). The Determinants of Industry Political Activity, 1978–1986. *American Political Science Review*, 88, 911–926.
- Hillman, A., & Keim, G. (1995). International Variation in the Business-Government Interface: Institutional and Organizational Considerations. *Academy of Management Review*, 20, 193–214.
- Holburn, G. L. F., & Vanden Bergh, R. G. (2002). Political Dynamics and Institutional Reform: The Diffusion of Utility Consumer Advocacy Offices in the United States. Unpublished Manuscript.
- Kalt & Zupan (1984). Capture and Ideology in the Economic Theory of Politics. *American Economic Review*, 74, 279–300.
- Kiewiet, D. R., & McCubbins, M. D. (1988). Presidential Influence on Congressional Appropriations Decisions. *American Journal of Political Science*, 32, 713–736.
- Krehbiel, K. (1991). *Information and Legislative Organization*. The University of Michigan Press, Ann Arbor.

- Krehbiel, K. (1999). Pivotal Politics: A Refinement of Non-market Analysis for Voting Institutions. *Business and Politics*, 1, 63–81.
- Mayhew, D. (1974). Congress: The Electoral Connection. Yale University Press, New Haven.
- McCubbins, M. D., Noll, R. G., & Weingast, B. R. (1987). Administrative Procedures as Instruments of Political Control. *Journal of Law, Economics and Organization*, 3, 243–277.
- McCubbins, M. D., Noll, R. G., & Weingast, B. R. (1989). Structure and Process, Politics and Policy: Administrative Arrangements and the Political Control of Agencies. *Virginia Law Review*, 75(2), March, 431–508.
- McCubbins, M. D., & Schwartz (1984). Congressional Oversight Overlooked: Police Patrols vs. Fire Alarms. *American Journal of Political Science*, 28, 165–179.
- Nelson, R. A. (1982). An Empirical Test of the Ramsey Theory and Stigler-Peltzman Theory of Public Utility Pricing. *Economic Inquiry*, 20, 277–290.
- Nelson, R. A., & Roberts, M. J. (1989). Ramsey Numbers and the Role of Competing Interest Groups in Electric Utility Regulation. *Quarterly Review of Economics and Business*, 29, 21–42.
- Pashigian, P. B. (1984). The Effect of Environmental Regulation on Optimal Plant Size and Factor Shares. *Journal of Law and Economics*, 27, 1–28.
- Pittman, R. (1976). The Effects of Industry Concern in Three 1971 U.S. Senate Campaigns. *Public Choice*, 23, 71–80.
- Poole, K. T., & Rosenthal, H. (1991). Patterns of Congressional Voting. *American Journal of Political Science*, 35(1), 228–278.
- Romer, T., & Snyder, J. (1994). An Empirical Investigation of the Dynamics of PAC Contributions. *American Journal of Political Science*, 38, 745–769.
- Schwartz, E., Spiller, P., & Urbiztondo, S. (1994). A Positive Theory of Legislative Intent. *Law and Contemporary Problems*, 57.
- Shepsle, K. A., & Weingast, B. R. (1987). The Institutional Foundations of Committee Power. *American Political Science Review*, 81(1), 85–104.
- Shepsle, K. A., & Weingast, B. R. (1989). Penultimate Power: Conference and the Legislative Process (Unpublished Manuscript).
- Snyder, J. (1990). Campaign Contributions as Investments: The U.S. House of Representatives, 1980–1986. *Journal of Political Economy*, 98, 1195–1227.
- Snyder, J. (1990). Committee Power, Structure Induced Equilibria and Roll Call Votes (Unpublished Manuscript).
- Snyder, J. (1991). On Buying Legislatures. *Economics and Politics*, 3, 93–109.
- Spiller, P. T. (1990). Politicians, Interest Groups, and Regulators: A Multiple-Principals Agency Theory of Regulation, or Let Them Be Bribed. *Journal of Law & Economics*, 33(1), 65–101.
- Spiller, P. T. (1992). Agency Discretion Under Judicial Review. *Mathematical and Computer Modelling*, 16(8/9), 185–200.
- Spiller, P. T. (1992). Governmental Institutions and Regulatory Policy: a Rational Choice Analysis of Telecommunications Deregulation. In: Quirk & Rich (Eds), *Improving Public Policy Decision Making*.
- Spiller, P. T., & Gely, R. (1992). Congressional Control or Judicial Independence: The Determinants of U.S. Supreme Court Labor-Relations Decisions, 1949–1988. *Rand Journal of Economics*, 23(4), 463–492.
- Spiller, P. T., & Tiller, E. H. (1997). Decision Costs and the Strategic Design of Administrative Processes and Judicial Review. *The Journal of Legal Studies*, XXVI, 347–370.
- Stratmann, T. (1992). Are Contributors Rational? Untangling Strategies of Political Action Committees. *Journal of Political Economy*, 100, 647–664.

- Stratmann, T. (1998). The market for congressional votes: Is timing of contributions everything? *Journal of Law & Economics*, 41, 85–113.
- Tiller, E. H. (1998). Controlling Policy by Controlling Process: Judicial influence on Regulatory Decision Making. *Journal of Law, Economics and Organization*, 14(1), 114–135.
- Tiller, E. H., & Spiller, P. T. (1999). Strategic Instruments: Legal Structure and Political Games in Administrative Law. *Journal of Law, Economics and Organization*, 15(2), 349–377.
- Vanden Bergh, R. G. (2000). The Evolution of Institutions: Politics and Process in the American States (Ph.D. Dissertation). University of California, Berkeley.
- Weidenbaum, M. (1980). Public Policy: No Longer a Spectator Sport for Business. *Journal of Business Strategy*, 1, 40–53.
- Weingast, B. R. (1981). Regulation, Reregulation and Deregulation: The Political Foundations of Agency Clientele Relationships. *Law and Contemporary Problems*, 44(1), 147–177.
- Weingast, B. R., & Marshall, W. J. (1988). The Industrial Organization of Congress; or, Why Legislatures, Like Firms, are not Organized as Markets. *Journal of Political Economy*, 96(1), 132–163.
- Weingast, B. R., & Moran, M. J. (1983). Bureaucratic Discretion or Congressional Control? Regulatory Policymaking by the Federal Trade Commission. *Journal of Political Economy*, 91(5), 765–800.
- Wilson, J. Q. (1980). The Politics of Regulation. In: Wilson (Ed.), *The Politics of Regulation*. Basic Books, New York.
- Yoffie, D. B. (1987). Corporate Strategies for Political Action: A Rational Model. In: A. A. Marcus, A. M. Kaufman & D. Beam (Eds), *Business Strategy and Public Policy: Perspectives from Industry and Academia* (pp. 43–60). New York: Quorum Books.
- Zardkoohi, A. (1985). On the Political Participation of the Firm in the Electoral Process. *Southern Economic Journal*, 51, 804–817.

MANAGERIAL DECISION MAKING IN NON-MARKET ENVIRONMENTS: A SURVEY EXPERIMENT

John M. de Figueiredo and Rui J. P. de Figueiredo, Jr.

ABSTRACT

In this paper we consider a number of experiments to determine whether aspiring managers can solve non-market strategy problems. Utilizing a survey of nearly 300 MBA students, we show that with simple, single-stage problems, managers are very competent in reaching the optimal choice given their non-market environment. As problems become more complex, however, they have much greater difficulty in arriving at the optimal result. In this regard, analysts must use some caution when applying theories and evaluating empirical results concerning non-market behavior.

INTRODUCTION

Over the past decade, scholars and business executives alike have recognized that, in many businesses, effective non-market strategy is critical to sustaining profits and growth (Baron, 1996, 1997, 1999; Krehbiel, 1999; de Figueiredo & Spiller, 2000; de Figueiredo & Tiller, 2001; Henisz & Zelner, 2001).¹ But while scholars have made substantial progress in generating both positive and normative analyses of the components of effective non-market strategy, it is still an open question as to what degree practitioners are capable of

incorporating institutional and strategic analysis into their thinking. Indeed, as has been commented on before, managers are only marginally trained to cope with such problems (Weingast, 1987). As an important precursor to this research, therefore, it is necessary to understand the extent to which managers can solve non-market strategy problems, and the extent to which such solutions can be incorporated into mainstream strategic planning.

In practice, natural experiments to answer this type of question are difficult to identify in the real-world. As has long been recognized, the difficulty of controlling for a variety of complex and often times idiosyncratic strategic situations makes it hard to localize the extent to which the non-market component has either been effectively pursued or resulted in favorable competitive advantage. This problem suggests that an alternative approach, experimentation, might help us identify how managers can cope with non-market strategy problems. Since experimentation allows for independent controls over the strategic setting and institutions, it has the advantage of decomposing the “compositeness” of field data (Smith, 1989).

Experimentation as a way of identifying managerial behavior has a significant tradition in both economics and psychology (see Camerer, 1999). In general, there have been two uses of experiments: (1) to evaluate the predictions of particular theories; and (2) to identify appropriate behavioral assumptions for further development of theory. In economics, one example of the rich use of experimentation has been in testing theories of markets and the institutional rules that undergird them. Experiments have allowed economists to evaluate both the speed and efficiency of market institutions (Smith, 1989). At the same time, they have allowed analysts to refine behavioral assumptions about agents’ choices (Camerer & Lovo, 1998; Chacon & Camerer, 1996; Cox, Smith & Walker, 1988; Smith, 1989).² As Roth (1991, p. 107) comments, experiments can help “game theory bridge the gap between the study of ideally rational behavior and the study of actual behavior.” In psychology, experimentalists have also undertaken a large research program to identify the degree to which individuals can act as the *homo economicus* of economic theory. Thaler (1980) for example, relates the “endowment effect” to an aversion for losses (see also Kahneman, Knetsch & Thaler, 1990). Similarly, Kahneman and Tversky (1979) provide experimental results which identify the ability, and inability, of individuals to conduct Bayesian inference in the way rational agents are posited to behave (see also Bazerman, 1985). In both of these fields, experimentation has allowed analysts to modify the axiomatic assumptions of behavior to develop richer theories of social interaction (see, e.g. O’Donoghue & Rabin, 1999; Hermalin & Isen, 2000).

Although there has been an enormous amount of work completed on repeated “market” games and games with uncertainty (see Kagel & Roth, 1995;

Camerer, 2001, for a nice summary), there has been little applied to non-market strategy. Given the significance of non-market strategy in today's markets, it is useful to undertake a similar exercise to determine the extent to which executives in firms can solve strategic non-market problems. In order to evaluate whether managers are so able, we present the results of a survey experiment that sheds light on this question. The survey was given to almost three hundred masters degree students at two leading business schools in the United States.³ Each respondent was given a series of non-market situations pertaining to activity in administrative regulatory institutions (government agencies and the courts). In addition, various versions of the questions – with more or less information – were given to explore the effects such information has on non-market decision-making. The survey was designed to answer two questions. First, in general, can managers solve simple non-market problems? Second, to what extent does greater uncertainty or less clarity affect these decisions? Although our survey cannot differentiate whether the performance in the questions is attributable to issues of bounded rationality and an inability to solve expected value calculations, or to the precise nature of non-market settings, it does show that expected value calculations in non-market environments are not trouble-free for managers.

To summarize our results, we find that when the problem is fairly simple – requiring only a single non-market forum to assess – the students (managers) perform well: almost ninety percent propose the optimal strategy. When the problem is more complex, however, performance drops considerably: when students must allocate resources across a range of non-market outcomes, only two-thirds propose the payoff maximizing strategy. Finally, when non-market strategy requires thinking not just about the suppliers of policy but also competitors, the managers (students) performed even worse, only half achieving the optimal strategy.

A second set of results has to do with the degree to which managers can cope with less precision in non-market problems. To that end, we modify the survey for certain respondents as a way of understanding how managers deal with such uncertainty. First, we examine the extent to which managers can learn over time about non-market problems. In this case, the managers were given a series of questions of the same form to determine if the answers converged to the “right” or optimal one, over the series. In fact, our results here are heartening for the economic analysis of non-market strategy; they show that as students are given problems of a similar form, over time they reduce the losses from suboptimal answers. We also provided some students with less information – for example, with only “fuzzy” probabilities that map actions into outcomes. Here we found that, in general, managers who are faced with

only imprecise information about outcomes perform similarly to those with more precise information. However, one result stands out: when faced with low probability events, the managers perform more poorly with less information. Finally, in the same vein, we also assessed the degree to which managers would be willing to seek outside advice on non-market problems. It is easy to theorize that the relationship between advice-seeking and competence could be either positive – as more competent managers are also better at recognizing the need for outside advice – or negative – as less competent managers garner more value from outside advice. Here, two results stand out. First, managers who are worse at solving these problems are more likely to hire advisors to solve the problems for them. Second, managers faced with greater uncertainty are more likely to seek outside advice than those in which the environment is more precisely understood.

The paper therefore provides some behavioral regularities which are important to incorporate in developing the literature on non-market strategy. In particular, as managers face more and more complex non-market environments, they are more likely to act in a boundedly rational way. Thus, hyper-rational, non-market, complex theoretical models of non-market strategy may be beyond the capabilities of managers to play optimally in real life. On the other hand, when faced with repeated situations, the traditional *homo economicus* might become a reasonable approximation of managerial behavior.

The paper proceeds as follows. In the following section we describe the survey instrument used and the sample of students. In Section II, we present the analysis of increasingly complex strategy situations and how respondents performed. In addition, we also examine the nature of biases, if any, in responses. In Section III, we turn to extensions of the basic instrument to determine how uncertainty, fuzziness and learning affect the basic results. In Section IV we offer concluding remarks.

I. THE SAMPLE AND INSTRUMENT

In order to explore whether managers can solve simple games involving non-market strategy, we conducted a survey of 289 MBA students at two leading business schools. We purposefully chose MBA students (as opposed to, say, political science students) because, first, in general, most of these students have been managers prior to entering business school; and second, these are students who will confront non-market problems as they progress through their careers. Thus, to a first approximation, they are more likely to reflect the managerial ranks than are other types of students.

In Business School No. 1, we randomly sampled one-half of the first year class (179 students). All students were enrolled in a “core” sequence that included economics, statistics, strategic management, organizational behavior, communications, and accounting. In the economics course, all students were introduced to simple, one-stage Nash equilibrium games and simple repeated games concepts.

In Business School No. 2, we sampled 60% of the second year class (110 students).⁴ All students had completed their core sequence and had completed over 75% of their electives. These students had also taken many of the same core courses as those students had in Business School No. 1, and were introduced to the same game theoretic concepts. In addition, the students in Business School No. 2 who took the survey were also concurrently enrolled in a required second year course on business-government relations.

In Table 1 we present the descriptive statistics for the sample. The full sample is presented in the first column, followed by the Business School No. 1 and Business School No. 2 subsamples. We believe this group is fairly representative of middle management at most companies. The average age of the sample is 27–30 years, with 4–6 years work experience. Two thirds are male, and roughly two thirds are native English speakers. Nearly a quarter have another graduate degree (besides the MBA they are earning), and nearly 85% have GMAT scores higher than 700. The risk tendencies of members of the group are presented in a number of other questions. While it is hard to measure risk tendency, we ask a number of questions to attempt to capture the differing dimensions of risk. With respect to risk for body injury, 90% of students wear seat belts, and less than 10% smoke regularly. With respect to financial risk tendencies, over 20% of survey participants have over 90% of their savings in stocks, which, for this age group, would be considered an “aggressive” financial strategy. Only 10% of respondents have children, which might lower their tendency to take many types of physical and financial risks. We do not expect students to have special knowledge of the scenario used in the survey, as only 5% could name the Federal Communications Commission (FCC) Chairman, and only 7% had sued or have been sued.

Although the two business schools are relatively homogeneous by most measures, there are some small differences. The average age and the number of years work experience at Business School No. 2 is higher than at Business School No. 1. However, Business School No. 1 has a higher proportion of individuals with graduate degrees, and has higher mean GMAT scores.

The survey instrument was given to each student in class, and they had 25 minutes to complete the exercise. Students were told the survey was optional; names and other identifying information were not requested. To motivate the

Table 1. Participant Characteristics.

Variable	Full Sample	Business School No. 1	Business School No. 2
CALCULATOR USED	0.08	0.08	0.08
MALE	0.67	0.71	0.61
AGE < 26	0.13	0.18	0.05
AGE 27–30	0.71	0.70	0.72
AGE 31–35	0.14	0.09	0.21
AGE 35 +	0.01	0.02	0.01
GRADUATE SCHOOL	0.24	0.27	0.20
0–3 YRS WORK EXP	0.22	0.26	0.16
4–6 YRS WORK EXP	0.58	0.58	0.60
7–9 YRS WORK EXP	0.13	0.10	0.17
10+ YRS WORK EXP	0.04	0.05	0.03
ENGLISH FIRST LANGUAGE	0.63	0.65	0.59
WEAR SEAT BELTS	0.89	0.87	0.94
SMOKER	0.09	0.09	0.08
PARENT	0.09	0.07	0.11
90% INVESTMENT STOCKS	0.22	0.23	0.20
FORMER GOVT EMPLOYEE	0.16	0.16	0.15
SUED BEFORE	0.07	0.07	0.07
GMAT <500	0.00	0.00	0.01
GMAT 500–549	0.00	0.01	0.00
GMAT 550–599	0.01	0.00	0.03
GMAT 600–649	0.08	0.08	0.08
GMAT 650–699	0.34	0.31	0.39
GMAT 700–800	0.49	0.53	0.43
CAN NAMECHAIRMAN FCC	0.05	0.06	0.05

Note: Means presented. All variables are dummy variables. $N = 289$ for full sample, $n_1 = 179$ for Business School No. 1, $n_2 = 110$ for Business School No. 2.

students to do well, we positioned the survey as a way for diagnosing their understanding of non-market issues, and as a competition between Business School No. 1 and Business School No. 2, to see which students performed better.⁵ No questions were permitted after the instrument was handed out.

A copy of most of the questions in the survey instrument is found in the Appendix. The survey instrument was piloted on ten students and modified before being issued to the 289 MBA students. We have included five questions on the survey instrument displayed, although the actual survey instrument had only three questions for each student (hence the varying sample sizes). The actual three questions and the ordering of the questions differed across students, so that we could conduct statistical tests of question independence. Every student received a question about Lobbying Alone (Question 1). Every student received

a question about Lobbying and Litigation – some received a question with fixed probabilities, while others received a question with fuzzy probabilities (Question 2). Finally, every student received a question about competitive lobbying. Some received a question without learning, while others received a question with learning (Question 3). We include the descriptive statistics for the questions answered in Table 2.

Table 2. Survey Responses.

Variable	Obs	Mean	Std. Dev	Min	Max
CALCULATOR USED	289	0.08	0.28	0	1
MALE	289	0.67	0.47	0	1
AGE <26	289	0.13	0.34	0	1
AGE 27–30	289	0.71	0.46	0	1
AGE 31–35	289	0.14	0.35	0	1
AGE 35+	289	0.01	0.12	0	1
GRADUATE SCHOOL	289	0.24	0.43	0	1
0–3 YRS WORK EXP	289	0.22	0.42	0	1
4–6 YRS WORK EXP	289	0.58	0.49	0	1
7–9 YRS WORK EXP	289	0.13	0.33	0	1
10+ YRS WORK EXP	289	0.04	0.20	0	1
ENGLISH FIRST LANGUAGE	289	0.63	0.48	0	1
WEAR SEAT BELTS	289	0.89	0.31	0	1
SMOKER	289	0.09	0.28	0	1
PARENT	289	0.09	0.28	0	1
90% INVESTMENT STOCKS	289	0.22	0.41	0	1
FORMER GOVT EMPLOYEE	289	0.16	0.37	0	1
SUED BEFORE	289	0.07	0.26	0	1
GMAT <500	289	0.00	0.06	0	1
GMAT 500–549	289	0.00	0.06	0	1
GMAT 550–599	289	0.01	0.10	0	1
GMAT 600–649	289	0.08	0.27	0	1
GMAT 650–699	289	0.34	0.47	0	1
GMAT 700–800	289	0.49	0.50	0	1
CAN NAME CHAIRMAN FCC	289	0.05	0.22	0	1
Q1 LOBBYING	274	50.31	12.36	0	149
Q2A 50% LITIGATION	159	46.60	22.22	0	100
Q2B 20% LITIGATION	154	27.52	35.14	0	200
Q2C 50% CONSULTANT PAY	145	20.23	32.29	0	160
Q2D 20% CONSULTANT PAY	140	11.19	28.58	0	160
Q2A MODERATE LITIGATION	110	46.72	23.31	0	160
Q2B LOW LITIGATION	108	21.50	33.15	0	200
Q2C MODERATE CONSULTANT PAY	103	23.04	39.57	0	151
Q2C LOW CONSULTANT PAY	98	17.80	36.81	0	200
Q3A 50M COMPETITOR	275	98.45	12.85	50	150

Table 2. Continued.

Variable	Obs	Mean	Std. Dev	Min	Max
Q3B 100M COMPETITOR	275	133.02	47.10	0	200
Q3C EQUILIBRIUM COMPETITOR	170	127.79	58.65	0	200
Q3D EQUILIBRIUM OWN	165	25.28	52.17	0	201
Q3C 125M COMPETITOR LEARNING	98	151.08	57.74	0	199
Q3D 175M COMPETITOR LEARNING	98	31.63	72.53	0	225
Q3E 195M COMPETITOR LEARNING	97	25.25	68.64	0	245
Q3F LEARNING EQ COMPETITOR	97	143.68	51.12	0	399
Q3G LEARNING EQ OWN	93	28.70	59.56	0	200

II. SURVEY QUESTIONS AND RESULTS

Lobbying Alone

In this section of the paper, we review the main questions that help us to understand the ability of students to solve non-market games. We are trying to determine if students can solve economic games that are couched as non-market strategy games, and whether there are systematic mistakes that are made by managers.

In order to provide context, students were asked questions about an adapted version of a model presented in de Figueiredo and de Figueiredo (2002). The first question that is asked of every student is a question about lobbying alone. Students are asked to choose a non-market investment level that is optimal given the structure of the payoff function. This question is designed to reflect the managerial decision making about how much to lobby, thinking of lobbying as an investment.

By construction, the payoff function maintains constant marginal returns, with a peak at a \$50 million investment in lobbying. We illustrate this in Fig. 1, as a loss function. Firms are at threat to lose \$200 million if they do nothing. By investing some amount, they mitigate that loss. Every \$1 million investment enhances the probability of being successful in lobbying by 2% to a maximum of 100%. Thus, the minimum of the linear loss function resides at a corner of a \$50 million investment. It is also important to recognize that the loss function is not symmetric, because the (opportunity) cost of underinvesting by \$1 million (a missed opportunity to earn \$3 million – 2% of \$200 million, less the \$1 million in lobbying expense) is greater than the cost of overinvesting (above \$50 million which costs simply \$1 million for every million invested). So, by construction, there are smaller penalties for overshooting, than for undershooting investment.

Figure 1 illustrates that nearly 90% of respondents chose to invest in lobbying at the highest expected payoff level. Further, despite the asymmetric marginal benefit function, nearly twice as many respondents underinvested in lobbying than overinvested in lobbying. The underinvestors did 20% worse than the overinvestors on average. Nevertheless, the results for underinvestors (relative to overinvestors) were better than expected because those who underinvested tended to underinvest only slightly; those who over invested, overinvested substantially. On the whole, this first question provides us with a good

Q1: Expected value calculation to lobbying under constant marginal benefit.

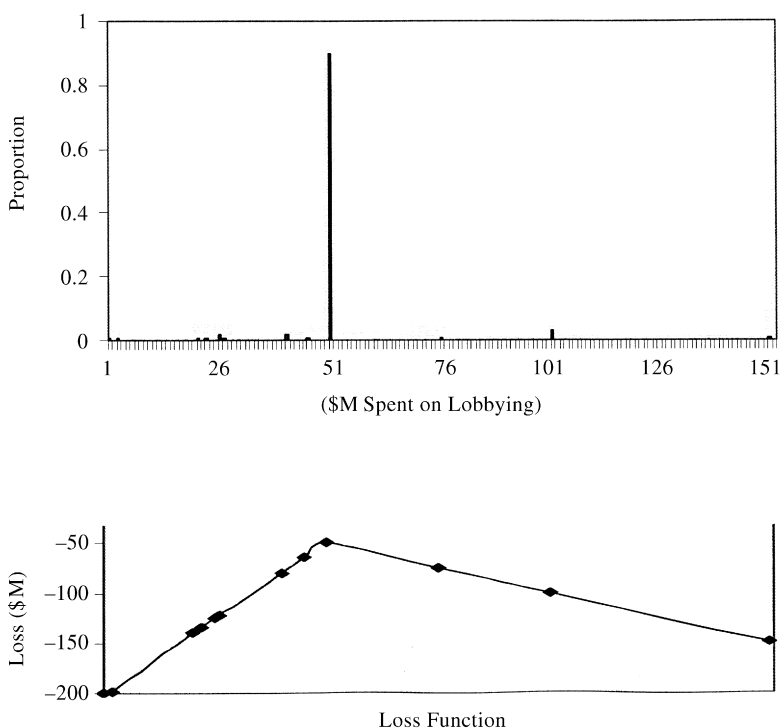


Fig. 1. Lobbying Alone.

Note: The bottom graph represents the payoff function (loss function) to lobbying. The top graph represents the frequency distribution of chosen investment levels in lobbying, given an experimental question with the payoff function in the bottom graph. Nearly 90% of respondents chose the level of investment that would maximise his/her payoff (minimize the loss).

benchmark to what is a relatively straightforward non-market question: how much should I invest in lobbying if I consider the lobbying game in isolation? Nine out of ten MBA students are able to lobby optimally.

Lobbying and Litigation with Certainty

The second question explores the relationship between lobbying and litigation. In reality, firms can lobby regulators for a favorable rule in an administrative agency. However, if the rule is unfavorable, parties are usually free to litigate in court over the outcomes. This question is designed to see if managers can solve this simple two-stage game when all information is deterministic.

The first part of question two offers the same game as before, but in this case appends a litigation outcome. The first part of the question assumes that there is a 50% chance the court will rule in your favor; all else carries over from the previous question. The question poses: given this second stage litigation, how much should the firm invest in lobbying in the first stage? The nature of this question is actually quite simple to implement in practice, but sometimes difficult for managers to internalize. A 50% probability of being overturned means merely that the expected payoff to lobbying is now cut in half. That is, lobbying investment is loss with certainty, but the expected payoff to investment is half of what it was without the court. This results in a symmetric loss curve as the marginal payoff to small amounts of lobbying is flatter. Thus, the global minimum of the loss function is still at a \$50 million investment, but expected return is $-\$150$ million rather than $-\$50$ million as it was in the previous question. Overshooting and undershooting are equally costly because of the symmetry of the loss curve. We illustrate the results in Fig. 2.

In this part of the question, 64% of respondents chose the payoff-maximizing level of investment in lobbying, conditional on a 50/50 chance of winning in litigation *ex post*. Nearly 10% of respondents chose to overinvest (investing up to \$100M) while over one-quarter of respondents chose to underinvest.

In the second part of the question, we try to explore the behavior of managers further along this dimension. Given the same setup, we ask a subset of students to evaluate how much they would invest in lobbying given a 20% probability of a favorable outcome in court.⁶ This question is designed to show that if there is a very high probability of losing in court, it may make sense not to invest in lobbying at all, even if you can obtain a favorable outcome. The way this question is structured, the optimal investment is \$0. This is because the loss function when lobbying is always lower than the loss function when there is not lobbying – that is there is a corner solution to the problem (see Fig. 3).

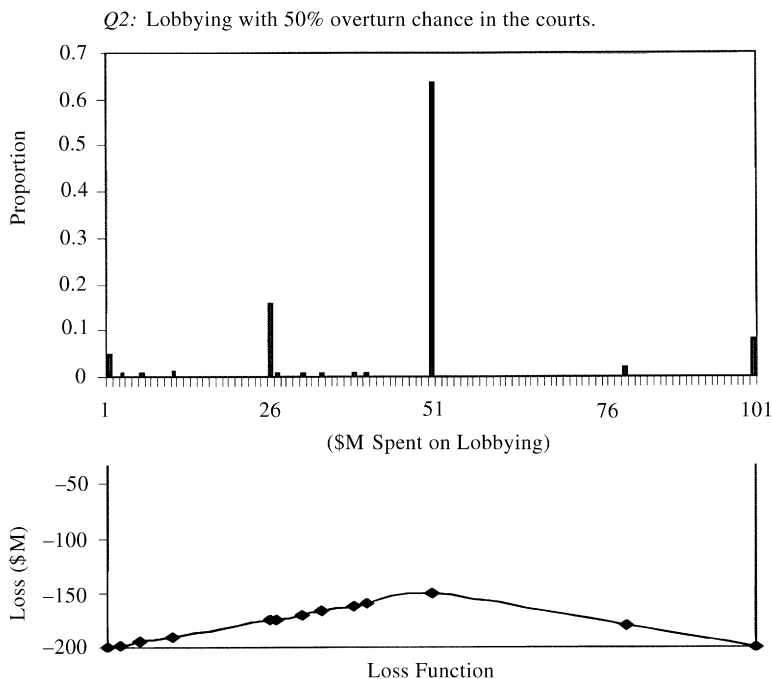


Fig. 2. Lobbying and Litigation.

Note: The bottom graph represents the payoff function (loss function) to lobbying when there is a 50% probability that the court will overrule the agency. The top graph represents the frequency distribution of chosen investment levels in lobbying, given an experimental question with the payoff function in the bottom graph. Almost 64% of respondents chose the payoff maximizing level of lobbying. Ten percent chose over-investment, while 26% chose to under-invest.

Almost 35% of respondents chose the payoff maximizing level of lobbying, choosing to invest nothing, while nearly two-thirds chose to overinvest. In this situation, however, small overinvestments do not hurt you very much, because the loss curve does not become steep until after \$50 million. Almost 27% chose to invest between \$40M and \$50M in lobbying, which was the policy payoff maximum level in the 50% probability question. Ten percent opted to spend more than \$50 million.

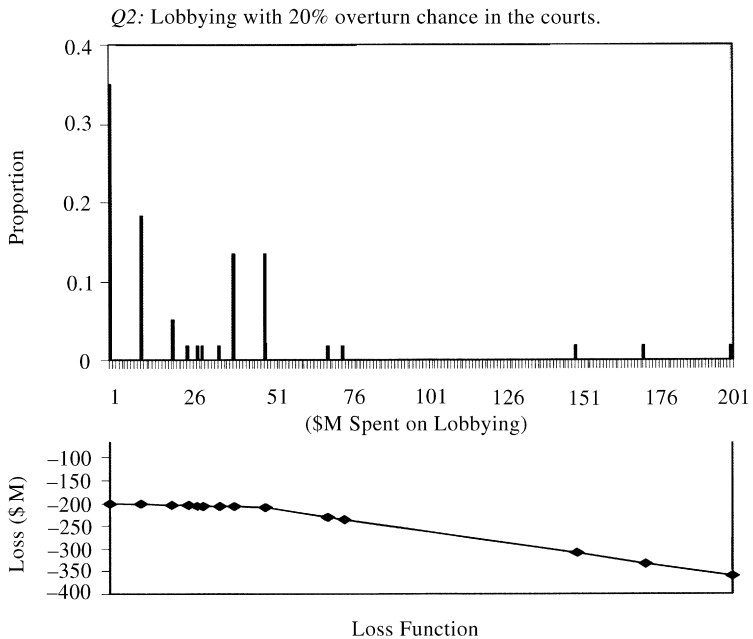


Fig. 3. Lobbying and Litigation (20%).

Note: The bottom graph represents the payoff function (loss function) to lobbying where there is a 20% probability that the court will overrule the agency. The top graph represents the frequency distribution of chosen investment levels in lobbying, given an experimental question with the payoff function in the bottom graph. The maximum payoff occurs at \$0M investment. Almost 35% of respondents chose the payoff maximizing level of lobbying, while nearly two-thirds chose to over-invest. Almost 27% chose to invest between \$40M and \$50M in lobbying.

Competitive Lobbying

The third question we pose to MBA students focuses on competitive lobbying. The recent literature on competitive lobbying (e.g. Austen-Smith & Wright, 1996; Groseclose & Snyder, 1996) suggests that lobbying does not occur in a vacuum, but in competition with opposing interest groups. Thus, while the first

two questions focused on non-competitive lobbying, the competition between interest groups is explicitly addressed in this question.

Here, we posited that there are two groups competitively lobbying for rents. They move sequentially, so that the MBA student's firm moves first, and is then followed by the competitive firm. (This insures a unique equilibrium.) While the marginal value to lobbying is the same as it was before, the value is determined by the net difference in lobbying expenditures between the student and the competitor. Whereas before, in Question 1, \$50 million was the global maximum (\$50M times 2% = 100% probability of not losing \$200 million), in this case, the returns to lobbying would be (Your Expenditure – Competitor Expected Expenditure) times 2%. This problem calls to mind the model of Groseclose and Snyder (1996) of buying supermajorities, where we have sequential moves and the optimal strategy of the first-mover often limits the second-mover's ability to "cherry pick" cheap individuals from thin majorities.

To solve this problem, which can be complex, students have to recognize that if they invest too little, the second mover will come in and obtain all the rents. So while the optimal investment may first seem to be \$50 million, it is clear that the competitor will come in investing \$100 million and win the \$200 million prize with certainty. Students must recognize that there is a blocking strategy so that their investment will prevent entry. If we assume that the ties are given to the first mover, then the blocking strategy is to invest \$150 million or more to earn the \$200 million with certainty. There is no incentive for the second mover to invest, because the return is at best zero. We illustrate this as a loss function in Fig. 4.

Figure 4 provides the frequency distribution for respondents. A full 50% of the respondents chose to invest at the payoff maximizing level, \$150 million.⁷ This group would effectively block second movers from lobbying for policy, and receive the highest payoff. Alternatively, 50% of respondents did not respond with the optimal value. Nearly 20% overinvested in lobbying, while 30% underinvested in lobbying. Those who underinvested took a particularly large hit, because not only did they not prevent lobbying by a second-mover, but they also lost their investment. (Contrast this to those who overinvested, and lost their investment, but blocked entry by a second mover.) Those respondents who underinvested found themselves substantially worse off than those who overinvested. The extra loss attributed to underinvestment was nearly three times, on average, the loss from overinvestment. Taken together, it would suggest that managers might have some difficulty thinking about competitive lobbying and lobbying coupled with litigation, which does not afflict managerial decision making in non-competitive lobbying.

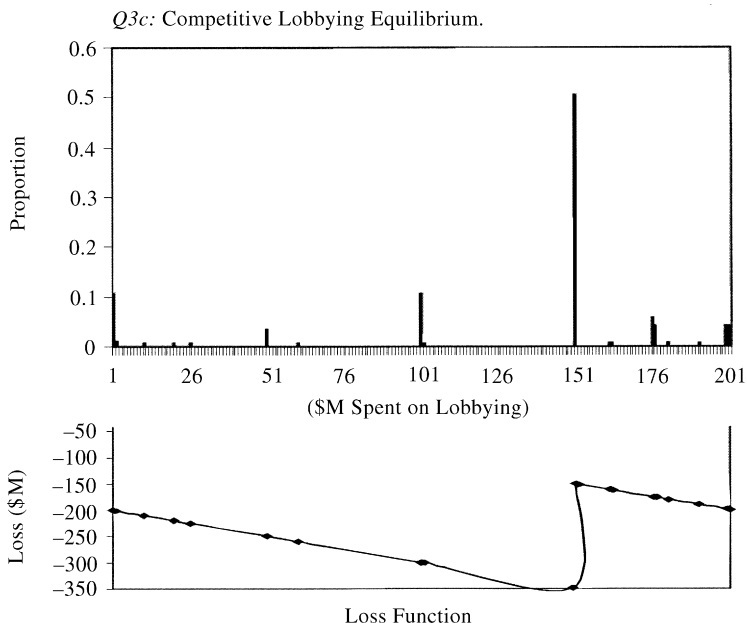


Fig. 4. Competitive Lobbying Equilibrium.

Note: The bottom graph represents the payoff function (loss function) to competitive lobbying. The top graph represents the frequency distribution of chosen investment levels in lobbying, given an experimental question with the payoff function in the bottom graph. The maximum payoff occurs when one deters a second mover from lobbying. This occurs when the lobbying expenditure by the first mover is sufficiently high. Approximately 50% of respondents chose the maximum of the payoff with 30% under-investing and 20% over-investing.

III. EXTENSIONS

In this section, we introduce three extensions to the analysis.⁸ The first is an analysis of learning. Although managers may be asked to make decisions regarding non-market strategy, it is unclear whether managers are “thrown” into this situation or gradually “ramp-up” to learning about the situation over time.⁹ While learning over many years is difficult to replicate in a survey, we have introduced one question with learning. On the competitive lobbying question, we included, on some surveys, a question where students were “guided” to the correct answer with leading questions. Rather than just asking

them for the answer on the question, we asked a series of questions so that they could triangulate on the equilibrium. In doing this, we hoped that they would understand in what range the equilibrium existed, and they would figure out how to solve this problem. Thus, as an addendum to question 3, we asked how much the student would spend in lobbying if the competitor had spent \$50, \$100, \$125, \$175, and \$195 million respectively. We expected to have students learn that the answer was somewhere between \$125 and \$175 million, and the intuition would be clear.

Figure 5 illustrates the results under learning. Although 7% fewer of the respondents chose the correct answer in this situation than did in the situation without learning, a much higher percentage chose to over-invest rather than under-invest in the optimal lobbying effort. This is quite interesting. It is likely due to the nature of learning mechanism. The survey asked what one expected the second mover to invest if you, as the first mover invested \$100, \$125, \$175, and \$195 million. \$175 million is the first answer where blocking occurs. Thus, 22% of respondents focused on the \$175 million expenditure and wrote this as their level of investment, rather than understanding the precise mechanism by which blocking occurred, and thus obtain blocking a cheaper price. Nevertheless, students did seem to learn, and greatly increased their investment over the “non-learning” situation, because they tended to over-invest for benefit, rather than under-invest. This in turn results in a much better situation for the firm, than under-investment.

The second extension we introduced was uncertainty. As with many problems managers face, the exact probabilities are unknown. Rather, managers, through advisors or other means, usually obtain fuzzy probabilities of success, such as “low,” “moderate,” “high.” We incorporated this thinking into the analysis of litigation. We replicated question 2 on litigation. However, instead of giving the participants concrete probabilities of winning in litigation, we asked them how much they would invest in lobbying (the lobbying production function is known with certainty) when the probability of winning in litigation was “moderate” and “low.”

The mean value of investment under moderate probabilities of winning was \$46.7M, compared to \$46.6M investment for a 50% chance of winning for a non-overlapping set of questionnaires. The optimal lobbying investment in this situation is \$50M. This would suggest that students impute a 50% chance of winning in litigation when they are told they have a “moderate” chance. When told they have a “low” chance of winning, the mean investment in lobbying is \$21.5M, compared to \$27.5M when told they have a 20% chance of winning in litigation. This result suggests two things. First, the respondents likely impute a probably lower than 20% when they are told they have a “low” probability

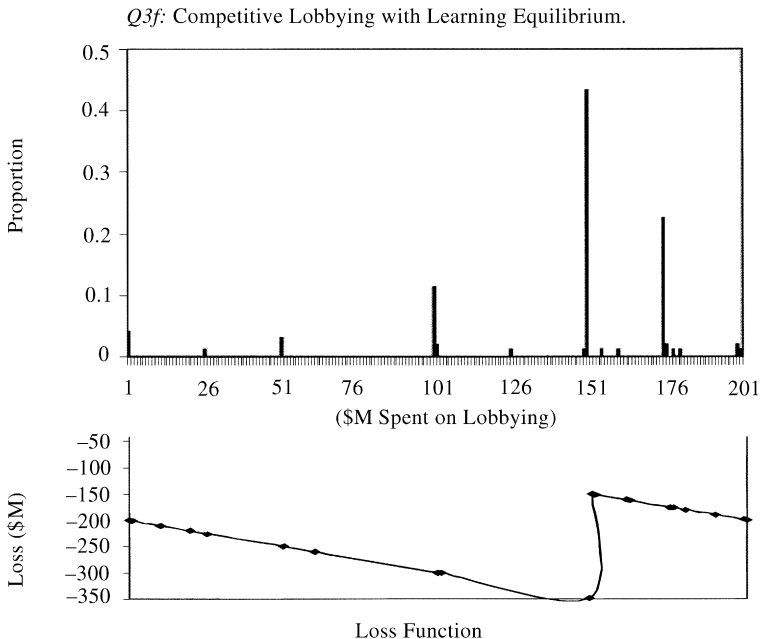


Fig. 5. Competitive Lobbying with Learning.

Note: The bottom graph represents the payoff function (loss function) to competitive lobbying. The top graph represents the frequency distribution of chosen investment levels in lobbying, given an experimental question with the payoff function in the bottom graph. The maximum payoff occurs when one deters a second mover from lobbying. This occurs when the lobbying expenditure by the first mover is sufficiently high. Approximately 43% of respondents chose the maximum of the payoff function with 25% under-investing and 32% over-investing. Note that 23% of respondents chose to invest \$175M, the tipping point in the survey question.

of winning. Second, it may be better to have fuzzy low probabilities, because then the managers choose closer to the optimal investment, \$0M.¹⁰ Notably, this result is consistent with the literature on low and negative probabilities and human behavior. That is, managers have difficulty in dealing with low probability events and making what are profit-maximizing decisions in this realm (Kahneman & Tversky, 1979; Viscusi, 1999).

A final extension we introduced into the survey was on the reliance on consultants. A common critique made of the non-market strategy field is that managers do not have to know non-market management tools, because they

have people who can advise them.¹¹ We tackle this critique in this survey instrument. These experienced advisors and consultants can offer managers the “right” answer. We append to question two additional questions related to how much a manager would be willing to pay an advisor under conditions of certainty and under conditions of uncertainty. This allows us to identify how confident managers are in their calculations. It also allows us to see if managers are willing to pay advisors more under uncertainty (where supposedly their value is higher) than under conditions of certainty. The relationship between advice-seeking and competence could be either positive – as more competent managers are also better at recognizing the need for outside advice – or negative – as less competent managers garner more value from outside advice. We ask these questions for both the high and low probability of winning in litigation, where the consultant offers advice on optimal lobbying investments.

The first interesting result is that managers are willing to pay consultants more the higher the likelihood of winning in court. With a 50% chance of winning in litigation, a consultant is paid on average \$20M by the firm, but with a 20% chance of winning in litigation, a consultant is paid \$11M by the firm. This is interesting because the consultant may be of more value when the probability is low, and the firm should not waste its time investing. To examine whether there were systematic differences between those respondents who chose the payoff-maximizing investment and those who did not, we constructed a dummy variable that is equal to one if the optimal investment is chosen and zero otherwise. With a 50% chance of winning, the consultant makes no more nor less money from those people who chose the correct answer than those who chose the incorrect answer (correlation -0.02). However, at the 20% chance of winning in litigation, the manager is likely to pay the consultant more if the manager is unable to calculate the correct answer herself. This correlation of -0.25 is statistically significant and suggests that managers who cannot solve these problems do hire consultants, especially when probabilities are in the tails.

We then compared these results to the results with fuzzy probability. The mean consulting fee for a firm that faces a moderate probability of winning in litigation is \$23.0M, and is \$17.8M for one that faces a low probability of winning in litigation. First, recognize that both of these numbers are higher than the fees they earn under certainty (paragraph above). So managers pay advisors more the greater the uncertainty. Second, managers still pay advisors more the better the outcome. That is, high probabilities of winning are associated with higher fees, despite the fact that the loss function may be quite steep at low probabilities.

IV. CONCLUSION

As we commented in the introduction, as the field of non-market strategy develops, it is crucial to understand the degree to which practitioners conform to the stark behavioral assumptions underlying theories. As a means of assessing the degree to which the rational assumptions underlying these theories are reasonable approximations, we provided almost three hundred prospective managers with increasingly complex strategic situations. Our purpose was to assess the degree to which the respondents could solve such problems. Indeed, to the extent that the question had precise answers which could be analytically derived, the first set of results speak more to the question of whether managers can solve expected utility problems – in this case couched as non-market problems.

In general our results show that with simple, single-stage problems, managers are very competent in reaching the optimal choice given their environment. As problems become more complex, however, they have much greater difficulty in arriving at the optimal result. In this regard, analysts must use some caution when evaluating and applying theoretical results. Thus, moving ahead, we must be careful about how we interpret non-market strategy empirical results. Studies with non-results do not necessarily suggest that non-market strategy is not important, but perhaps that managers have not figured out their optimal strategy.

That said, however, our results also include a strong optimism for economic analysis of non-market strategy. For they show that at least in the stark environment we provided, that over time, when confronted with similar problems, managers can adapt and learn about their environment and improve their performance. Further, our final set of results show that when faced with problems they cannot solve, managers will seek outside advice in order to reach a “better” solution.

Certainly, experiments in non-market strategy are useful in simplifying and controlling for a variety of complex situations that make it hard to identify the extent to which the non-market component has either been effectively pursued or resulted in favorable competitive advantage. As extensions to the work, it would be interesting to replicate this experimental study in other settings. Jaffee and Freeman (2002), in this volume, suggest sophisticated non-market strategies are being employed in Germany. Creating experiments to compare the sophistication of managers in other national settings, to those in the United States, would be an interesting comparative study to follow. In addition, replicating our experiments in a market setting would be a useful extension of this work, because our survey cannot determine whether any shortcomings by respondents is due to general problems of bounded rationality, or specifically

to the non-market setting of the questionnaire. Identifying analogous questions and setups in the market environment, and conducting a similar survey, would allow us to separate out the effects of our results that are generated from the “non-market environment” vs. the “analytical setup and bounded rationality.” Nevertheless, there are a number of reasons to believe that the application of traditional self-interest calculations might have to be modified when moving to non-market settings. Therefore, a final path to pursue is to separate out exactly how non-market strategy questions might differ systematically from market strategy questions.

NOTES

1. To cite some examples, Baron (1997) shows how in one case, the trade dispute between Kodak and Fujifilm, the rents earned by proper execution of non-market strategy are potentially substantial. de Figueiredo and Spiller (2000) similarly argue that the enormous rents on the table for telecommunications firms were distributed differentially in the United States and Europe among interexchange carriers and local exchange carriers depending on the regulatory environment in the United States and Europe. de Figueiredo and Kim (2001) estimate that even the recent administrative regulatory dispute over payphone calling card pricing for long distance calls could cost the players \$400 million over ten years. Finally, de Figueiredo and Silverman (2002) determine there are hundreds of millions of dollars at stake every year in the university appropriations process, which can be attributed to university lobbying.

2. A primary example of this type of approach has been the identification that agents tend to bid on “the risk-averse side of Vickrey’s linear bid function” in first-price sealed bid auctions (Smith, 1989, p. 158). This finding led to further experimentation and subsequent theory modification to include heterogeneity in risk-aversion in auctions (Kagel & Levin, 1986).

3. As we discuss below, MBA students were engaged as a proxy for managers.

4. The sample for Business School No. 2 may not be random. In Business School No. 1, 179 of 180 students took the survey, representing nearly three full sections of the MBA class. To the extent that students are somewhat randomly assigned to sections, then the Business School No. 1 sample is random. In Business School No. 2, 40% of students did not attend the class. To the extent that the missing students were non-random, then the sample can be considered non-random. Nevertheless, the results from the two schools are remarkably similar.

5. For MBA students, this actually serves as a good motivator.

6. de Figueiredo and de Figueiredo (2002) develop a game theoretic model showing how lobbying and litigation are linked.

7. Because the tie-breaking rule was not clear in the survey instrument, we have combined those that answered \$150M and \$151M into “\$150M,” and report our results as thus going forward.

8. We conducted an econometric analysis of the data to see if we could determine what are the characteristics of managers who perform well in these types of non-market strategy questions. The results were, on the whole, not statistically significant. We did

find in one regression on competitive lobbying without learning, that students with GMAT scores above 700 did save up to \$45M in excessive losses at a 95% level of statistical significance. There were also small performance differences in performance across business schools.

9. In interviews with executives, the former seems to be a better characterization. These interviews have suggested that managers move along in their careers managing their market strategies. When they reach the executive ranks, they are suddenly confronted with a host of non-market issues, with which they are ill-equipped to contend.

10. There is another way of looking at this. That is, assume that managers choose the optimal investment with fuzzy probabilities. What is the probability they impute? Unfortunately we cannot answer this question, because with constant marginal benefit, managers should only choose \$0M or \$50M. To the extent that there is a bimodal distribution, reflecting indifference between the two corners, and we see \$21.5M as the outcome, this would suggest that 40% of managers believed that “low” meant less than 25% probability of winning, while 60% of managers believed that “low” meant 25% or greater probability of winning. Note, however, that this strictly economic interpretation must be taken carefully, as the results of question two are at odds with this outcome.

11. It is also interesting to note how frustrated many respondents were with the idea of having to make decisions with fuzzy probabilities. Many students, after the questionnaire was completed, wanted to know how they were supposed to make decisions without knowing the actual numbers.

12. This is somewhat akin to managers don’t need to know accounting, because they can hire accountants. Managers only need to know how to “manage.”

ACKNOWLEDGMENTS

Prepared for the “New Institutionalism in Strategic Management” Conference at Columbia University, April 21–22, 2001. Thanks to John Carroll and Roberto Fernandez for comments on survey design, and to Brian Silverman for helpful comments.

REFERENCES

- Austen-Smith, D., & Wright, J. R. (1994). Counteractive Lobbying. *American Journal of Political Science*, 38, 25–44.
- Baron, D. P. (1996). *Business and its Environment*. Englewood Cliffs N.J.: Prentice-Hall.
- Baron, D. P. (1997). Integrated Strategy and International Trade Disputes. *Journal of Economics and Management Strategy*, 6(2), 291–346.
- Baron, D. P. (1999). Integrated Market and Non-market Strategies in Client and Interest Group Politics. *Business and Politics*, 1(1), 7–34.
- Camerer, C. F. (1999). Behavioral Economics: Reunifying Psychology and Economics. *Proceedings of the National Academy of Sciences*, 96(19), 10575–10577.
- Camerer, C. F., & Lovo, D. (1999). Overconfidence and Excess Entry: An Experimental Approach. *American Economic Review*, 89(1), 306.

- Camerer, C. F. (2001). *Behavioral Game Theory: Experiments in Strategic Interactions*. Princeton: Princeton University Press.
- Chacon, G. P., & Camerer, C. F. (1996). Loss-avoidance and Forward Induction in Experimental Coordination Games. *Quarterly Journal of Economics*, 111(1), 165–194.
- Cox, J., Smith, V., & Walker, J. (1988). Theory and Individual Behavior of First Price Auctions. *Journal of Risk and Uncertainty*, 1, 61–99.
- de Figueiredo, J. M., & de Figueiredo, R. J. P., Jr. (2002). Lobbying, Litigation, and Administrative Regulation: The Allocation of Resources by Interest Groups. *Business and Politics* (forthcoming).
- de Figueiredo, J. M., & Silverman, B. S. (2002). Academic Earmarks and the Returns to Lobbying. MIT Sloan Working Paper #4245-02.
- de Figueiredo, J. M., & Kim, J. J. (2001). Strategic Information Transmission in Lobbying: Getting Ready for Litigation. *MIT Working Paper*.
- de Figueiredo, J. M., & Tiller, E. H. (2001). The Structure and Conduct of Lobbying: An Empirical Analysis of Corporate Lobbying at the Federal Communications Commission. *Journal of Economics and Management Strategy*, 10(1), 91–122.
- de Figueiredo, R. J. P., Jr., & Spiller, P. T. (2000). Strategy, Structure and Regulation: Telecommunications in the New Economy. *Michigan State University – Detroit College of Law Review*, 1, 253–285.
- Groseclose, T., & Snyder, J. (1996). Buying Supermajorities. *American Political Science Review*, 90(2), 303–315.
- Henisz, W., & Zelner, B. (2001). The Institutional Environment for Telecommunications Investment. *Journal of Economics and Management Strategy*, 10(1), 123–148.
- Hermalin, B. E., & Isen, A. (2000). *The Effect of Affect on Economic and Strategic Decision Making* (Working Paper). University of California, Berkeley.
- Holburn, G., & Vanden Bergh, R. (2002). Policy and Process: A Game-Theoretic Framework for the Design of Non-market Strategy. *Advances in Strategic Management* (Vol 19, pp. 33–66). Elsevier Science.
- Jaffe, J., & Freeman, J. (2002). Institutional Change in Real-Time: The Development of Employee Stock Options for German Venture Capital. *Advances in Strategic Management* (Vol 19, pp. 219–246) Elsevier Science.
- Kagel, J., & Levin, D. (1986). The Winner's Curse and the Value of Time: Results from a Natural Experiment. *American Economic Review*, 76, 894–920.
- Kagel, J., & Roth, A. (Eds) (1995). *Handbook of Experimental Economics*. Princeton: Princeton University Press.
- Kahneman, D., Knetsch, J. L., & Thaler, R. H. (1990). Experimental Tests of the Endowment Effect and the Coase Theorem. *Journal of Political Economy*, 98(6), 1325–1348.
- Kahneman, D., & Tversky, A. (1979). Prospect Theory: An Analysis of Decision Under Risk. *Econometrica*, 47(2), 263–291.
- Krehbiel, K. (1999). Pivotal Politics: A Refinement of Non-market Analysis for Voting Institutions. *Business and Politics*, 1(1), 63–82.
- Neale, M., & Bazerman, M. H. (1985). The Effects of Framing and Negotiator Overconfidence on Bargaining Behaviors and Outcomes. *Academy of Management Journal*, 28(1), 34–49.
- O'Donoghue, T., & Rabin, M. (1999). Doing It Now or Later. *American Economic Review*, 89(1), 103–124.
- Roth, A. E. (1991). Game Theory as a Part of Empirical Economics. *Economic Journal*, 101, 107–114.
- Smith, V. (1989). Theory, Experiment and Economics. *Journal of Economic Perspectives*, 3, 151–169.

- Thaler, R. (1980). Toward a Positive Theory of Consumer Choice. *Journal of Economic Behavior and Organization*, 1 (March), 39–60.
- Viscusi, W. K. (2000). How do Judges think about risk? *American Law and Economics Review*, 1(1–2), 26–62.
- Weingast, B. R. (Ed.) (1987). *Political Economy and the Business School Curriculum*. St. Louis: Washington University in St. Louis Press.

APPENDIX

Non-Market Strategy Questionnaire

The attached questionnaire is completely OPTIONAL and ANONYMOUS. It will be used for the purposes of teaching and research. In addition, any published results or presentation of the results will be only concerning aggregate patterns of responses. All answers will be kept otherwise anonymous.

You have 25 minutes to answer the questions. You may use blank spaces for calculations. You may use a calculator or other electronic device.

Please note: the boxed part of each question is identical across all questions.

Please answer the two questions below before you leave.

PLEASE WRITE THE TIME ON THE CLOCK WHEN YOU TURN IN THIS SURVEY. TIME ON CLOCK AT COMPLETION: _____

DID YOU USE A CALCULATOR OR COMPUTER TO ANSWER ANY OF THE QUESTIONS? YES NO

Please answer the following questions. We are not asking for your name or otherwise identifying information. You will remain anonymous. This is for our records so that our analysis can be complete.

Sex: Male Female

Age (circle): less than 26 26-30 31-35 35+

In addition to your current MBA studies, have you completed any other graduate degree? Yes No

Number of Years Full Time Work Experience (circle): 0-3 4-6 7-9 10+

Native Language: English Other (Specify)_____

Do you wear seat belts more than 80% of the time? Yes No

Do you smoke more than one cigarette a week? Yes No

Do you have children? Yes No

Is more than 90% of your wealth in the stock market? Yes No

Have you worked for more than three months as a government employee? Yes No

Have you ever been sued or sued anyone? Yes No

GMAT math score (circle):

less than 500 500-549 550-599 600-649 650-699 700+ don't know

Please name the Chairman of the Federal Communications Commission, if you know who it is, otherwise write a zero?

NAME: _____

Note: Boxed part of question is identical across all questions.

Q1. You are the Chief Operating Officer of Telco, Inc., a very large telecommunications company. Many of your activities are regulated by the Federal Communications Commission (FCC), a government commission that oversees the activities of telecommunications companies.

Under current regulations set by the FCC, you are required to pay, for the next year only, a tax to the government of \$200 million. The \$200 million tax will be transferred to LDCo, a long distance company. After next year, the tax will end.

The Board and CEO have come to you and are asking you to eliminate that tax. Studies have shown that this reduction in tax will have NO impact on the volume of calls or the prices which are charged by you or the long distance companies. It only impacts the amount of the transfer. So if the tax is eliminated, you, as Telco, will earn an additional \$200 million next year.

In this question only, your sole means to influence the FCC is to lobby (or persuade) the FCC to lower the tax. No other actors can influence the outcome. For every \$1 million you spend, there is a 2% increase in the probability (to a maximum of 100%) that the FCC will vote in your favor and eliminate the tax. So, for example, if you spend \$10 million, there is a 20% chance the FCC will vote for no tax, and an 80% chance the FCC will keep the \$200 million tax.

In this question, there are no other avenues of appeal of an FCC decision. The FCC decision is final.

How much do you spend on lobbying?

PUT ANSWER HERE: _____

Note: Boxed part of question is identical across all questions.

Q2 and 3. You are the Chief Operating Officer of Telco, Inc., a very large telecommunications company. Many of your activities are regulated by the Federal Communications Commission (FCC), a government commission that oversees the activities of telecommunications companies.

Under current regulations set by the FCC, you are required to pay, for the next year only, a tax to the government of \$200 million. The \$200 million tax will be transferred to LDCo, a long distance company. After next year, the tax will end.

The Board and CEO have come to you and are asking you to eliminate that tax. Studies have shown that this reduction in tax will have NO impact on the volume of calls or the prices which are charged by you or the long distance companies. It only impacts the amount of the transfer. So if the tax is eliminated, you, as Telco, will earn an additional \$200 million next year.

In this question only, regulation in the United States is a two-stage process. First, you can lobby (or persuade) the FCC to lower the tax. No other actors can influence the outcome at this stage. For every \$1 million you spend, there is a 2% increase in the probability (to a maximum of 100%) that the FCC will vote in your favor and eliminate the tax. So, for example, if you spend \$10 million, there is a 20% chance the FCC will vote for no tax, and an 80% chance the FCC will keep the \$200 million tax.

Second, a party who does not like the ruling, can appeal to the court. You are relatively certain LDCo would be unhappy by a lower tax, and would choose to take the case to court if the FCC decided to lower the tax. In general, the court can uphold the decision of the FCC, or, if the decision is different from the original state of affairs, revert back to the original state of affairs. The original state of affairs is a taxation rate of \$200 million.

In our example, if the FCC ruled \$200 million then the Court could only rule \$200 million. However, if the FCC ruled no tax, the Court would have a choice of upholding the FCC decision at no tax or reverting back to \$200 million.

Suppose that the FCC is deciding between reducing the one-time tax to nothing or keeping the tax at \$200 million, and the court will then rule.

[The questions appear on the next page.]

Note: Boxed part of question is identical across all questions.

Q4. You are the Chief Operating Officer of Telco, Inc., a very large telecommunications company. Many of your activities are regulated by the Federal Communications Commission (FCC), a government commission that oversees the activities of telecommunications companies.

Under current regulations set by the FCC, you are required to pay, for the next year only, a tax to the government of \$200 million. The \$200 million tax will be transferred to LDCo, a long distance company. After next year, the tax will end.

The Board and CEO have come to you and are asking you to eliminate that tax. Studies have shown that this reduction in tax will have NO impact on the volume of calls or the prices which are charged by you or the long distance companies. It only impacts the amount of the transfer. So if the tax is eliminated, you, as Telco, will earn an additional \$200 million next year.

In this question, your sole means to influence the FCC is to lobby (or persuade) the FCC to lower the tax. For every \$1 million you spend, there is a 2% increase in the probability (to a maximum of 100%) that the FCC will vote in your favor and eliminate the tax. So, for example, if you spend \$10 million, there is a 20% chance the FCC will vote for no tax, and an 80% chance the FCC will keep the \$200 million tax.

Unfortunately, in this question only, you are not alone. LDCo is also lobbying. The FCC decision will be determined based on the *net* amount spent on lobbying. In particular, the probability the FCC will vote to eliminate the tax will increase by 2% for every \$1 million *more* that Telco spends than LDCo (to a maximum of 100%). If the net amount is zero or negative then the FCC will not change the tax. So for example, if LDCo spends \$100 million on lobbying and Telco spends \$125 million then the probability that the FCC will rule to eliminate the tax will be 50% (i.e. \$125 million spent by Telco - \$100 million spent by LDCo = \$25 million $\times$ 2% = 50% chance the FCC will rule to eliminate the tax).

In this question, there are no other avenues of appeal of an FCC decision. The FCC decision is final.

[The questions appear on the next page.]

A. Suppose the court has a 50% chance of upholding the FCC and a 50% chance of overturning the FCC ruling. How much would you as Telco spend on lobbying?

PUT YOUR ANSWER HERE: _____

B. Suppose the court has a 20% chance of upholding the FCC and an 80% chance of overturning the FCC ruling. How much would you as Telco spend on lobbying?

PUT YOUR ANSWER HERE: _____

C. How much would you as Telco pay a litigation consultant, who is very experienced at making these calculations, as a consulting fee, to advise you on your optimal expenditures, assuming the chances in court are 50/50 as in part A of this question?

PUT YOUR ANSWER HERE: _____

D. How much would you as Telco pay a litigation consultant, who is very experienced at making these calculations, as a consulting fee, to advise you on your optimal expenditures, assuming the chances in court are 20/80 as in part B of this question?

PUT YOUR ANSWER HERE: _____

A. If LDCo has spent \$50 million on lobbying, how much will you as Telco spend on lobbying?

PUT YOUR ANSWER HERE: _____

B. If LDCo has spent \$100 million on lobbying, how much will you as Telco spend on lobbying?

PUT YOUR ANSWER HERE: _____

C. If you do not know how much LDCo has spent on lobbying, how much would you expect them to spend, assuming they spend money first, and they know Telco receives a 2% higher probability of winning for every million dollars more Telco spends than LDCo spends?

PUT YOUR ANSWER HERE: LDCo: _____

How much would you as Telco spend on lobbying, given LDCo has already spent the amount you have calculated above?

PUT YOUR ANSWER HERE: Telco: _____

Note: Boxed part of question is identical across all questions.

Q5. You are the Chief Operating Officer of Telco, Inc., a very large telecommunications company. Many of your activities are regulated by the Federal Communications Commission (FCC), a government commission that oversees the activities of telecommunications companies.

Under current regulations set by the FCC, you are required to pay, for the next year only, a tax to the government of \$200 million. The \$200 million tax will be transferred to LDCo, a long distance company. After next year, the tax will end.

The Board and CEO have come to you and are asking you to eliminate that tax. Studies have shown that this reduction in tax will have NO impact on the volume of calls or the prices which are charged by you or the long distance companies. It only impacts the amount of the transfer. So if the tax is eliminated, you, as Telco, will earn an additional \$200 million next year.

In this question, your sole means to influence the FCC is to lobby (or persuade) the FCC to lower the tax. For every \$1 million you spend, there is a 2% increase in the probability (to a maximum of 100%) that the FCC will vote in your favor and eliminate the tax. So, for example, if you spend \$10 million, there is a 20% chance the FCC will vote for no tax, and an 80% chance the FCC will keep the \$200 million tax.

Unfortunately, in this question only, you are not alone. LDCo is also lobbying. The FCC decision will be determined based on the *net* amount spent on lobbying. In particular, the probability the FCC will vote to eliminate the tax will increase by 2% for every \$1 million *more* that Telco spends than LDCo (to a maximum of 100%). If the net amount is zero or negative then the FCC will not change the tax. So for example, if LDCo spends \$100 million on lobbying and Telco spends \$125 million then the probability that the FCC will rule to eliminate the tax will be 50% (i.e. \$125 million spent by Telco - \$100 million spent by LDCo = \$25 million $\times$ 2% = 50% chance the FCC will rule to eliminate the tax).

In this question, there are no other avenues of appeal of an FCC decision. The FCC decision is final.

[The questions appear on the next page.]

A. If LDCo has spent \$50 million on lobbying, how much will you as Telco spend on lobbying?

PUT YOUR ANSWER HERE: _____

B. If LDCo has spent \$100 million on lobbying, how much will you as Telco spend on lobbying?

PUT YOU ANSWER HERE: _____

C. If LDCo has spent \$125 million on lobbying, how much will you as Telco spend on lobbying?

PUT YOU ANSWER HERE: _____

D. If LDCo has spent \$175 million on lobbying, how much will you as Telco spend on lobbying?

PUT YOU ANSWER HERE: _____

E. If LDCo has spent \$195 million on lobbying, how much will you as Telco spend on lobbying?

PUT YOU ANSWER HERE: _____

F. If you do not know how much LDCo has spent on lobbying, how much would you expect them to spend, assuming they spend money first, and they know Telco receives a 2% higher probability of winning for every million dollars more Telco spends than LDCo spends?

PUT YOUR ANSWER HERE: LDCo: _____

How much would you as Telco spend on lobbying, given LDCo has already spent the amount you have calculated above?

PUT YOUR ANSWER HERE: Telco: _____

PRETTY PICTURES AND UGLY SCENES: POLITICAL AND TECHNOLOGICAL MANEUVERS IN HIGH DEFINITION TELEVISION

Glen Dowell, Anand Swaminathan and
James Wade

INTRODUCTION

The development of High Definition Television in the United States has been a nearly twenty year odyssey, one that has been characterized by political maneuvering, alliance formation and disbanding, and technological change that has frustrated actors' attempts to shape HDTV development in their own favor. In this paper, we treat technological standards as institutions, and we use the HDTV story to describe how social movements affect technological change, and how the effectiveness of social movements is in turn influenced by technological change. We find that actors involved in HDTV used collective action framing processes with varying degrees of success, and that the success of their framing attempts can be explained partly by standard explanations of what affects the efficacy of frames (Benford & Snow, 2000), and partly by the nature of technological change, which can render previously effective frames null and void.¹

The issues of how institutions change, and how private actors such as organizations and individuals bring about change in institutions, remain of key interest to new institutional theorists (Ingram & Clay, 2000). In this paper, we use social movement theory, and particularly the framing processes that the theory describes, in order to illustrate one way in which organizations collectively act to change institutions as represented by technological standards. Standards are institutions in the sense that they constrain firms' choices (Ingram & Clay, 2000), because the existence of technological standards limits the technological choices available to firms. We also show that social movement theory can shed light on the process of technological evolution, a topic that is of central concern to organizational theorists.

In recent years the emergence of many new knowledge-based industries has been credited to key technological innovations. For instance, Hafner and Lyon (1996) argue that the development of browser technologies set the stage for the commercialization of the worldwide web. Similarly, the development of recombinant DNA technology made the emergence of the biotechnology industry possible (Barley, Freeman & Hybels, 1992). While technological innovations are indeed important and even critical drivers of industry evolution, focusing on these innovations alone leads to an incomplete, undersocialized understanding of industry evolution.

A more complete account would include an analysis of the political and social aspects that come into play in the evolution of new industries. Collective action, for instance, is often required to capitalize on opportunity structures created by new technologies (Rao, Morrill & Zald, 2000; Swaminathan & Wade, 2001). In particular, collective action can lead to agreement on a technological standard, which, in turn can enhance the diffusion of knowledge across industry participants and legitimate the new technology (Aldrich, 1999).

We are not the first to suggest that technological change is socially embedded, of course. Pinch and Bijker (1987) recognize that technological evolution is not a linear, determinate process, and different actors may identify different shortcomings with any given technological artifact. Tushman and Rosenkopf (1992) describe various classes of technology and argue that open complex systems, which are comprised of a set of closed subsystems that are linked together, are especially likely to be subject to political processes, because such systems cut across organizational populations and may have consequences for national economies. In such cases, any technological alternative is unlikely to meet the preferences of all interested groups. Thus choosing a technological standard will inherently involve political processes.

While it has been recognized that technological trajectories are subject to political actions and that collective action may be required in order to affect

trajectories, we do not know enough about *how* organizations engage in collective action and act to affect technological evolution (Aldrich, 1999). In order to gather support for a technological innovation, organizations must often mobilize resources and garner the support of key constituents, including customers, suppliers, and regulatory agencies. In this paper we use social movement theory as a framework to view the actions that firms take to mobilize the resources necessary to enact technological change. Social movement research is concerned with understanding how actors generate collective action in order to effect social change. Swaminathan and Wade (2001) argue that the emergence of new organizational forms resembles a social movement process. Here, we extend their analysis and demonstrate that technological change is often influenced by social movement activity. In the HDTV case that we examine, major networks and other broadcasters that try to mobilize support for their retention of broadcasting spectrum under the guise of supporting the development of HDTV, undertake the social movement activities.

Technological evolution and collective action may affect each other, as collective action can influence technological standards, but evolution of technologies can impact the effectiveness of a social movement's attempt to create collective action. When new technology-based industries emerge they generate counter-movements led by firms from other industries that seek the same resources (Swaminathan & Wade, 2001). Technological change may influence this process by dissolving boundaries between what were once separate industries with distinct resource spaces. Rapid technological change, then, may make it more difficult for an industry to establish a standard because such changes will create opportunities for new entrants, who are unlikely to have the same interests as those who entered earlier. While new technological developments generate opportunities, technology, in turn evolves from the interaction of collaboration among alliance partners and competition across alliances.

Our objective in this chapter is to explore these issues by analyzing the technological evolution of the High Definition Television (HDTV) industry in the United States. The evolution of HDTV provides an excellent setting in which to discuss institutional effects, as the technology evolved through actions by individual firms and collective bodies, within an institutional framework, providing a setting in which technological progression was shaped by collective action. The HDTV case shows that organizations do not simply act within a given institutional framework, but actively attempt to shape and control institutions through collective action.

The HDTV case allows us to consider how framing attempts must evolve to account for changes in the environment in which the frames are cast. The nature of technological evolution, especially within open systems, is rife with potential

for unintended consequences. As an open system, HDTV is comprised of several subsystems, each of which evolved somewhat independently of the others, and the evolution at the subsystem level caused broadcasters and television manufacturers to change their frames and react to unforeseen threats. We examine the effect that the evolution of these subsystems has on the interactions between groups of firms, and on the degree to which different populations of firms interact with each other.

NEW TECHNOLOGIES AND COLLECTIVE ACTION

In technology-intensive industries fragmented interests across industry participants often complicate the collective action process. Such competition can hinder collective action as industry participants promote their own interests and fail to advance the interests of the overall industry in obtaining legitimacy (Aldrich & Fiol, 1994; Garud, 1994). Competing technological communities can arise as industry participants sort themselves out into competing alliances in order to determine standards in an era of ferment (Wade, 1995; Vanhaverbeke & Noorderhaven, 2001). Moreover, the competition between competing standards is not always based on clearly defined technical criteria. Because technological systems are complex, which system is technically superior often depends on the evaluative criteria chosen (Garud & Rappa, 1994). Competing groups strategically use political tactics to “frame” technology adoption issues to favor their own technology. In the early VCR industry, for example, Matsushita emphasized the longer playing time of VHS systems, while Sony stressed the superior picture quality of its BETA system.

It is especially important for firms to motivate collective action when the technology in question is an open system (Tushman & Rosenkopf, 1992). Such systems consist of multiple subsystems that interact with each other, so that the performance of the whole system is affected by the performance of each subsystem. These subsystems may evolve independently of the system itself, which can have unforeseen consequences for actors who attempt to shape the evolution of the system. Moreover, having heterogeneous actors involved in the creation of a new industry makes it difficult for a new organizational form to develop its own distinct identity and attain a taken-for-granted status (McKendrick & Carroll, 2001).

In open system technologies, the emergence of collective action is also particularly crucial because of intense competition from older established technologies. Older technologies are particularly difficult to replace when network externalities are present. Network externalities occur when the utility that a user derives from a good increases with the number of other adopters

(Katz & Shapiro, 1985). Thus, for instance, the worth of a telephone increases to the extent that the total number of telephones in use rises. Once an older technology has accumulated extensive organizational and consumer support, advocates of a new technological approach not only have to provide the base technology but also a host of supporting technologies at a cost that may be prohibitive (Wade, 1995).

Customers using the older technology will be reluctant to switch because their investments in the older system would be lost (Hannan & Freeman, 1989; Lieberman & Montgomery, 1988). This problem may be exacerbated by the fact that the price of products based on new technology will have to remain quite high until there are many adopters. Few actors, however, may be willing to adopt new technology unless the price is low so that switching costs can be minimized.

Because of such externalities, new technological approaches may not be pursued, even when the technological capability for introducing and developing them is present. For instance, although high-speed rail is clearly feasible and has been introduced in a number of countries, there has been virtually no impetus to introduce this technology in the United States. In many cases the reluctance to pursue new technological options occurs because of the high costs involved in introducing the technology and uncertainty about whether the technology will diffuse.

Similar circumstances impeded the introduction of high definition television (HDTV) in the United States. High definition television was first developed in Japan by NHK. NHK had been broadcasting in high definition in Japan since 1979, and in 1981 it demonstrated its technology to the Society of Motion Picture and Television Engineers (SMPTE). The technology that NHK brought to the U.S. was developed in conjunction with the Nippon Broadcasting Company and consisted of 1,125 lines of resolution at a 60 MHz refresh rate. The existing standard in the United States (known as the NTSC standard)² was 525 lines of resolution; NHK's system, operating at a resolution that was more than double the existing standard, produced pictures that were astoundingly clear.

Despite its technical success, NHK's demonstration generated little interest among most U.S. broadcasters and on the part of the National Association of Broadcasters (NAB), the association that represented broadcast stations. This lack of interest was not entirely surprising because of the huge investments that would be required by broadcasters to switch to the HDTV format. Broadcast stations would have to potentially buy new cameras, new transmitting equipment, and perhaps even build new types of TV transmitting towers, changes that would be extremely costly. Because of the high switching costs,

the majority of broadcasters showed little interest in the new technology. Consistent with the standards literature, these high switching costs would seem to indicate that the future of HDTV in the United States was relatively gloomy. How, then, did HDTV gain a foothold in the United States? One answer to this question lies in how key actors with a stake in this technology developed a compelling collective action frame that unified disparate organizational populations and motivated collective action.

SOCIAL MOVEMENTS AND FRAMING PROCESSES

Collective action frames are systems of shared beliefs that justify the existence of social movements and spur collective action (Klandermans, 1997). These frames are the products of strategic efforts by groups of people to fashion shared understandings of the world that legitimate and motivate collective action. Thus, frames involve agency, as it is the actors within the social movement who create frames and attempt to use particular frames to incite action aimed at achieving a particular goal.³

Three framing tasks must be accomplished if collective action is to occur. The three core framing tasks are diagnostic framing, prognostic framing and motivational framing (Snow & Benford, 1988). In diagnostic framing, activists focus on a problem that demands redress and identify the agents responsible. This usually includes identifying opponents (Benford & Hunt, 1992; Morris, 1992). Prognostic framing involves having a plan to resolve the problem. Motivational framing provides motives or reasons why the collective action should be undertaken. There is likely a correspondence between the diagnostic and prognostic frames, as the diagnosis of the problem and antagonists tends to limit the solutions and strategies available to a social movement organization (Benford, 1987). This interdependence between the diagnostic and prognostic framing is consistent with existing ideas of social constraints on technological change (Nelson & Winter, 1982; Pinch & Bijker, 1987).

Attending to the three core framing tasks allows an organization to generate consensus on the problem and solution, and action toward resolution of the problem. Merely using the three framing tasks does not guarantee a social movement organization's success, however. Instead, the degree to which the tasks are effective depends upon the resonance of the frames and possibly whether the frame connects to a "master frame." As Klandermans (1997) notes, one viable strategy for organizations advocating change is to establish linkages to other organizations and populations. In order to do this, the change agents must have a resonant collective action frame, which requires that the frames be empirically credible and consistent.⁴ Characteristics that are most often

associated with a frame's resonance are those that connect the frame to the cultural meanings and symbol systems of the movement's audience (Babb, 1996). A frame that is empirically credible will connect more closely with the movement's audience (Benford & Snow, 2000). Empirical credibility describes whether advocates of the frame can point to convincing events in the world that tend to support their claim. For example, student activists in the Chinese democracy movement could point to political reforms in the Soviet Union under the Gorbachev regime to argue credibly that similar change was possible in China (Zuo & Benford, 1995).

Another characteristic that determines a frame's resonance is its consistency (Benford & Snow, 2000), or the extent to which a movement's articulated beliefs match its actions. Social movement research, however, has not really examined cases where there are inconsistencies between articulated beliefs and actual beliefs of movement actors. Just as organizations may adopt structures to meet institutional demands while decoupling these structures from their technical core (Meyer & Rowan, 1977), organizations or groups may adopt frames that are inconsistent with their true beliefs in order to achieve their goals. Such strategic action is especially likely when multiple actors with divergent interests must be mobilized.

Mobilizing multiple actors requires the creation of a master frame, a collective action frame that is more inclusive and flexible. Most collective action frames are associated with the narrow interests of a particular organizational population. Master frames, however, help in establishing linkages to other organizational populations. According to Benford and Snow (2000), master frames are broad in scope and serve as sort of a master algorithm for multiple social movements. For instance environmental justice frames (Cable & Shriver, 1995), rights frames (Valocchi, 1996) and choice frames (Davies, 1999) can be conceptualized as master frames which have been utilized by a variety of social movements. We argue below that the national competitiveness frame is a master frame that is often used by organizations trying to influence Government policy.

FRAMING PROCESSES IN THE HDTV INDUSTRY

Framing processes often play a critical role in technological evolution and standard setting. A technological standard creates institutional constraints on actors, narrowing their perception of what is possible and limiting their search for new solutions (Nelson & Winter, 1982). Essentially, the movement toward a technological standard is a socially embedded process, in which the standard itself becomes an institutional force, as it takes on a taken for granted status.

During this standard setting process, however, the involved actors often have diverging interests and preferences. Political conflict and competing alliances often emerge as these actors try to influence the standard setting process and the evolution of the new technology. The framing actions used by these actors and the persuasiveness of these frames in gaining support from important stakeholders often determines the path of technological evolution and which actors will benefit. In the case of HDTV, these framing actions influenced the direction of HDTV evolution, and were in turn influenced by evolution of HDTV components.

In the remainder of this section, we describe and analyze the first two framing attempts that broadcasters made as they attempted to retain spectrum. In both the framing efforts, broadcasters attempted to stop land mobile organizations from obtaining spectrum that the broadcasters owned but were not using. In the first effort, broadcasters utilized direct confrontation with land mobile companies and argued that television signals would be compromised if spectrum were given to the land mobile companies. In the second effort, broadcasters re-framed their arguments to portray the spectrum decision as being related to issues of technological progress and national competitiveness. We describe below why the first frame failed, necessitating the second frame, and why the second effort was more successful. In the final sections of this paper, we outline how technological evolution and conflicting interests challenged the second frame, and how this affected the process of introducing HDTV to the U.S. market.

The Creation of a Collective Action Frame Among Broadcasters

It is commonly found in research on social movements that the identification of a threat from an opposing party encourages cooperation and the generation of a collective action frame (Benford & Snow, 2000). In 1985, broadcasters recognized the threat posed by new regulations proposed by the FCC that would take away some of the unused spectrum allocated to broadcasters and allocate it to users of two way radios such as police and departments and ambulance services as well as cellular services.⁵ Spectrum simply refers to the channels that broadcasters were allocated, each channel occupying 6 MHz of radio frequency. While the broadcasters were allocated almost one hundred channels, half of these went unused. Some were set aside to provide for future use, while others acted as buffers between existing channels to prevent interference. So, for example, a TV channel that broadcasts on channel 3 has an empty channel above and below it on the spectrum in order to minimize interference from

channels 2 and 4. In addition, spectrum in the UHF frequency was largely unused, as up to 50 channels were set aside for television use, while most cities had at most 8 or 10 broadcast channels (Brinkley, 1997).

At the time that the FCC considered reallocating unused spectrum, it is unclear what, if any, plans the broadcasters had for using the spectrum in the future. Brinkley (1997) argues that the broadcasters had no uses in mind, but were still adamant that the spectrum was rightfully theirs and were desperate to retain it. Mueller (1983) contends that the broadcasters should possess full property rights over the spectrum that had been granted to them, and should therefore have the ability to develop or sell the spectrum as they see fit. In any event, the broadcasters saw the reallocation as a real threat to their future options for the spectrum, and sought a method of preventing the FCC from turning the spectrum over to land mobile companies.

Failure of Initial Collective Action Frame

By 1986, it appeared that the land mobile lobby, the association representing two-way radio users and cellular services, had the congressional support that it needed to take these “excess” channels away from the broadcasters. This threat was the catalyst that spurred the broadcasters to develop a common collective action frame. Specifically, the NAB, which represented all television and radio broadcasters and the Association of Maximum Service Telecasters (which later became MSTV), which represented television affiliates of the major networks, began to search for ways to forestall the spectrum reallocation.

The initial framing effort centered on maintaining the status quo. The diagnostic and motivational framing focused on convincing the FCC that the spectrum was necessary for the broadcasters to continue to provide quality television signals. The NAB’s chief lobbyist, John Abel, claimed that allocating their spectrum to Land Mobile uses would cause interference with existing television broadcasts and that consumers had a right to clear television reception (Brinkley, 1997). His hope was that this argument would convince the FCC to act in the broadcasters’ best interests.

Land mobile lobbyists, led by Motorola (which made most of the two way radios), argued, however, that allocating the channels to two-way radio users, such as police and ambulance services that provide life saving services, was more important than clear television reception. Moreover, land mobile lobbyists stressed that the channels were not currently being used, and that the television broadcasters should not be allowed to maintain possession of the channels if they had no plans to use them. While the threat of having their spectrum taken away unified the broadcasters by identifying a common threat

(diagnostic framing) and clear reasons for undertaking collective action (motivational framing), the NAB had little success in blocking the new regulations. By 1985 the NAB had clearly lost the battle to the land mobile firms and the FCC had decided to reallocate UHF spectrum to land mobile uses in some large urban markets.

At the same time that the NAB was fighting to keep the broadcasters' unused spectrum away from the land mobile firms, Japan's NHK was attempting to get its HDTV system approved as a worldwide standard. Up to this point, the NHK system had drawn little interest from broadcasters in the United States, because they perceived that a substantial investment would be required for them to switch to high definition signals. The turning point in NAB's interest in HDTV came when their lobbyists realized that it required more than one channel of bandwidth (Brinkley, 1997, p. 10). Since it required so much bandwidth to transmit, HDTV created the possibility that broadcasters would need their entire spectrum. Seizing on this fact, the NAB began to promote the promise of HDTV to anyone who would listen. The NAB began by inviting NHK to demonstrate their system in the United States, which was ironic considering that the broadcasters would soon resort to framing the Japanese as the enemy in the HDTV issue (Dupagne & Seel, 1998).

A Revised Collective Action Frame

As they started to create a new frame to combat the imminent loss of their spectrum, the broadcasters moved away from targeting land mobile use as their opponent, and started stressing two points. First, they noted that the NHK system was designed for broadcast from satellites, so they warned that steps must be taken to ensure that HDTV would be made freely available to American consumers. As Edward Fritts, president of the NAB argued, "It is a fact that consumers will be able to enjoy this improved broadcast in the near future. The question is whether the FCC will let them." He added that the FCC plan to reallocate spectrum to land mobile use would "preclude America's broadcasters from developing HDTV as a free over-the-air service to the nation (Abramovitz, 1987)." Broadcasters were particularly worried about satellite and cable substitutes, as those technologies were not constrained by the 6 MHz bandwidth restrictions that faced broadcasters. Thus, championing HDTV could potentially help broadcasters retain spectrum and fight the cable and satellite threats (Dupagne & Seel, 1998, p. 173).

Second, and perhaps more importantly, the NAB cautioned that acceptance of the NHK system would cement Japanese dominance of the consumer electronics industry in the United States. As Holburn and Vanden Bergh (this

volume) point out, one of the key issues in lobbying is often which party the lobbyist targets. Direct lobbying of the agency that makes a decision may in fact be less critical than attempting to affect the position of the legislative body. In this instance, some of the success of the NAB's use of HDTV as a bargaining chip stemmed from the impact of this argument on members of Congress. Representative Ritter of Pennsylvania, for instance, exclaimed that the U.S. had been "watching from the sidelines as the Japanese have taken this technology and run with it so many laps around this course (Brinkley, 1997, p. 28)."

The broadcasters, in fact, had a variety of parties who needed to be attracted to their framing efforts. The most immediate target was the FCC, and the committees that it formed, as the FCC has the mandate of regulating the broadcasters and any organizations that use spectrum in the United States. The FCC, in turn, is directly responsible to Congress, and its five commissioners are political appointees; therefore it was logical that the broadcasters attempt to enlist members of Congress in the collective action framing. Finally, electronics manufacturers and content producers (movie and television studios) represented important constituents to be attracted to the broadcasters' cause, as the new television sets needed for HDTV would be built by the electronics firms, and the electronics manufacturers could have influence with Congress because of the jobs that might be created by HDTV production.

The second framing attempt of the broadcasters was much more successful than their first. Indeed, a 1987 ruling set aside spectrum for the broadcasters' use, with the expectation that the spectrum would be used to provide HDTV. The outcome of each framing effort by the broadcasters can be traced to the relative resonance of each frame. As we discussed earlier, two important determinants of a frame's resonance are its empirical credibility and whether it is linked to a master frame; we now turn to discussing these issues for the second framing attempt in order to understand why it succeeded where the first attempt failed. The comparison of the two framing efforts is summarized in Table 1.

Frame Resonance: Empirical Credibility

As we discussed earlier, a frame's resonance partly depends upon its empirical credibility, which is the degree to which the actors creating the frame can point to events that support their claims. In the first framing attempt, there was little evidence that the use of spectrum by land mobile firms would lead to interference. The second framing attempt had a greater degree of empirical credibility, as the broadcasters linked U.S. participation in HDTV to the potential for the U.S. to regain competitiveness in consumer electronics. Specifically, the majority of

Table 1. Comparison of Broadcasters' Initial and Subsequent Framing Efforts.

Dimension of the Frame	Framing Attempt No. 1	Framing Attempt No. 2
Diagnosis	Impending Loss of Spectrum to Land Mobile Firms	Impending Loss of Spectrum to Land Mobile Firms
Prognosis	Directly Lobby FCC for Retention of Spectrum	Promote HDTV, which required more than 1 channel
Motivation	Crowding of Signals in Usable Spectrum will Cause Deterioration in Television Signal Quality	Failure to Act on HDTV will Cause U.S. to Miss out on Key Technology for 21st Century and Cost Jobs
Master Frame	None	Loss of U.S. Economic Competitiveness
Empirical Credibility	Limited	Strong, as Foreign Firms were Dominant in Electronics and Other Markets

statements were positioned in a negative manner, in effect saying that failure to participate in HDTV meant that the U.S. would never again have a domestic consumer electronic base. Fears of Japanese dominance in HDTV were bolstered by Japanese firms' dominance of almost every facet of the American consumer electronics market.

This frame had strong empirical credibility because U.S. consumer electronics firms had experienced precipitous erosion in competitive position over the previous ten years. By 1985, U.S. firms were effectively out of the VCR market and Zenith was the only U.S. television manufacturer left in the market. Indeed, as Tables 2 and 3 show, by 1990 Zenith had only a 12% market share of the color TVs sold in the United States and an even smaller share of world production.

These arguments appeared convincing to most observers and were buttressed by numerous reports justifying the importance of HDTV. For instance in 1989, the Economic Policy Institute issued a report claiming that without HDTV two million jobs would be sacrificed and that it would result in a 225 billion dollar trade deficit over the next 20 years. The fervor over HDTV led to the impression that "to miss out on HDTV is to miss out on the 21st Century" (Beltz, 1991, p. x, quoting Representative Ritter; see also Hart & Tyson, 1989).

Table 2. The U.S. Color Television Market in 1990.

COMPANY	COUNTRY	BRAND	SHARE (%)
THOMPSON	France	RCA (16.60%)	22.25
		GE (5.65)	
NORTH AMERICAN PHILIPS	Netherlands	Magnavox (7.75%)	12.20
		Sylvania (3.20%)	
		Philco (0.65%)	
		Philips (0.60%)	
ZENITH	United States	Zenith	11.65
SANYO	Japan	Sears (4.90%)	6.70
		Sanyo (1.50%)	
		Fisher (0.30%)	
MATSUSHITA	Japan	Panasonic (3.20%)	6.55
		Quasar (1.85%)	
		JVC (1.50%)	
SONY	Japan	Sony	6.50
SHARP	Japan	Sharp	5.00
TOSHIBA	Japan	Toshiba	4.00
EMERSON	United States	Emerson	3.80
MITSUBISHI	Japan	Mitsubishi	3.50
HITACHI	Japan	Hitachi	2.50
MOBIL	United States	Montgomery Ward	2.40
GOLDSTAR	South Korea	Goldstar	2.00
SAMSUNG	South Korea	Samsung	1.80
OTHERS	—	—	9.15

Source: Dupagne, M., & Steel, P. B. *High Definition Television: A Global Perspective*, p. 138. Iowa State University Press, Ames, Iowa.

*Frame Resonance:
Linkage to a Master Frame*

A frame that has empirical credibility may create legitimacy for its proponent's stance, but it does not necessarily lead to successful mobilization of other actors. Linking issues to a master frame can bring organizations from different populations together. In the case of the broadcasters' initial framing attempts, we can see evidence that the failure of the first attempt can be partially attributed to its inability to engage other actors in the broadcasters' battle. There were no other industries or organizational populations that stood to gain by broadcasters' retention of spectrum at the expense of land-mobile companies. In fact, companies that made products necessary for land mobile use, such as

computer chip and cellular phone manufacturers, would have stood to gain from the land mobile growth that would accompany the spectrum reallocation.

In seizing upon HDTV and American participation in electronics markets the broadcasters were successful in tying the spectrum issue to a master frame. As we discussed above, master frames are broad in scope and serve as sort of a master algorithm for multiple social movements (Benford & Snow, 2000). In this instance, the master frame was U.S. economic competitiveness in world markets, which was of widespread concern in the mid-1980s.

Another difference between the first and second framing attempts by the broadcasters was in the way in which opponents were identified. Specifically, the first framing attempt did not identify specific opponents.⁶ Instead the frame focused on the interference that might arise from using the spectrum. It is obvious that the land-mobile companies did not make for compelling opponents due to the importance of their services. Unfortunately for the broadcasters, no other opponent was readily available, given the frame adopted. In the second framing attempt, however, the specific antagonist identified was the Japanese consumer electronics industry. Japanese consumer electronics firms were easy targets because in the 1980s they clearly dominated HDTV technology. In 1987, NHK, a Japanese company was the only firm to have a working HDTV system and had already begun broadcasting HDTV in Japan. The Japanese consumer electronics industry also made for an attractive antagonist because in the previous 20 years it had taken the lead from the U.S. in almost all areas of consumer electronics and had been particularly adept at commercially exploiting new technologies. For instance, while Motorola developed color television technology and Ampex invented the first videotape system, it was Japanese companies that commercially introduced these innovations and dominated these markets. Adding fuel to this fire was the widespread perception of Japanese industry as posing a competitive threat in other markets not necessarily linked to consumer electronics such as the auto industry. Targeting the Japanese consumer electronics industry as an antagonist was quite effective because it identified an opponent that would link together multiple organizational populations. Unlike the previous frame, this master frame was able to gain significant support from not only the U.S. electronics industry but also from important government officials.

The process of identifying an antagonist to promote the collective action frame is similar to the idea that conflict with an outgroup leads to greater cohesion and commitment to action on the part of the focal group (Sherif et al., 1961). Swaminathan and Wade (2001) explain the relationship between outgroup identification and collective action in business settings. They cite several examples of firms or industry associations identifying antagonists as a

means of spurring collective action by related firms. For example, microbreweries have increased cohesion in that specialist sub-population by emphasizing the differences between themselves and the mass brewers. In general, it may not be necessary for the antagonists to represent an actual threat, as long as they can be perceived as representing a threat (Swaminathan & Wade, 2001; Ingram & Inman, 1996). It is the job of the parties creating the collective action frame to convey a sense of urgency and reality to the threat.

In this case, foreign electronics companies' dominance of the U.S. consumer electronics market was less a threat than a fait accompli, as detailed in Table 3. The threat was not only from Japanese companies, as Thompson and Philips together accounted for over 34% of the market, nearly as much as the top 7 Japanese firms combined. Foreign companies quickly became the principal concern for legislators who became involved in the HDTV issue. Representative Markey, chairman of the House telecommunications subcommittee stated his fear that without active promotion of an American HDTV system, Japanese and European companies would have an insurmountable lead in the industry (Rosenblatt, 1988). The HDTV industry was presumed to lead to spillover

Table 3. Top Worldwide Color TV Manufacturers in 1988
(Estimated Production Capacity in Millions of Units and
Gross Share in Percent).

COMPANY	COUNTRY	BRAND	SHARE (%)
PHILIPS	Netherlands	6.5–7.0	7.2–7.8
MATSUSHITA	Japan	6.5–7.0	7.2–7.8
THOMSON	France	6.5–7.0	7.2–7.8
SAMSUNG	South Korea	5.5	6.1
GOLDSTAR	South Korea	5.5	6.1
SONY	Japan	4.5	5.0
SHARP	Japan	3.5–4.0	3.9–4.4
TOSHIBA	Japan	3.5–4.0	3.9–4.4
SANYO	Japan	3.0–3.5	3.3–3.9
HITACHI	Japan	3.0–3.5	3.3–3.9
GRUNDIG	Germany	2.2	2.4
NOKIA	Finland	2.2	2.4
ZENITH	United States	2.1	2.3
TOTAL	—	90.0	100.0

Source: Dupagne, M., & Steel, P. B. *High Definition Television: A Global Perspective*, p. 139. Iowa State University Press, Ames, Iowa.

effects in computers and semiconductors (Beltz, 1991), increasing the perceived damage to the economy of failing to act to reduce Japan's lead in HDTV development.

Beltz (1991) argued that the actual threat to the U.S. economy from losing out on the HDTV front was lower than people who sought to promote American HDTV efforts claimed. Her argument relied on three principal issues. First, there was a great deal of uncertainty over the adoption rate for HDTV, and if it were to be adopted slowly, its impact on the economy would be much smaller than some were claiming. Second, the spillover effects from HDTV were likely to be much smaller than those from personal computers, as the computer was already diffusing rapidly among businesses and consumers, and as the PC evolved. Finally, many of the foreign firms had production and research facilities in the United States, employing more people than the lone remaining U.S. television producer (by this point, all but one of Zenith's plants were located outside the U.S.). Thus, she argues, it was hard to define a foreign television maker.

Given the conflicting arguments described in the previous paragraphs, it is not clear whether the foreign threat to U.S. consumer electronics firms was real. What is more interesting for the purposes of this analysis is that even if the foreign threat existed, it was not aimed at the group that initiated the framing attempt. The broadcasters did not face extinction at the hands of foreign competitors, but by invoking the foreign threat in their framing efforts, they were able to generate collective action and unite Congress, the electronics manufacturers, and the FCC in the effort to preserve the broadcasters' spectrum. Effectively, the broadcasters had created an outgroup in the Japanese and European firms, which helped to unite the interested parties in the United States.

In their promotion of HDTV, then, the broadcasters created an effective frame. They diagnosed the problem as being the impending acceptance of NHK's HDTV standard in the United States, and linked it to the issue of U.S. competitiveness in electronics and more generally, the country's economic future. Their prognosis was that national competitiveness in consumer electronics and employment growth in high technology sectors of the U.S. economy depended on the creation of a separate HDTV standard for the U.S. They successfully involved Congress and created a motivational frame that incited action toward creating an American standard for HDTV, rather than relying on the Japanese standard. In turn, this frame motivated American electronics organizations such as Zenith, Sarnoff Laboratories, and eventually AT&T to work toward creating HDTV.

The success of the second set of framing actions was evident in the FCC's 1987 decision to declare a moratorium on removing spectrum from the

broadcasters, and the subsequent establishment of an HDTV competition to create a U.S. based standard. The broadcasters had succeeded in motivating the FCC to act; some of this motivation came from the broadcasters' ability to bring in other actors such as legislative bodies (members of Congress) and other organizational populations such as the U.S. consumer electronics industry represented by the American Electronics Association.

Inconsistency within Broadcasters' Collective Action Frame

Though the broadcasters experienced success by framing their desire to retain spectrum as an issue of national competitiveness, their frame was inherently inconsistent, as it was built on the premise that the broadcasters needed spectrum to provide HDTV, when many, if not all, broadcasters were opposed to HDTV. In this case, it is clear that the broadcasters were most interested in maintaining their control of the TV spectrum and had only minimal interest in HDTV because of the high costs of switching to this technology (Brinkley, 1997). In fact, a 1989 study done by Robert Ross of KYW-TV Philadelphia shows that they may have had some reasons for concern. Ross estimated that it would cost as much as 38 million dollars to convert a station in a large market so that it could broadcast HDTV. While later cost estimates were much smaller, the NAB feared that the conversion costs would be difficult for broadcasters in small local markets to absorb (Lambert, 1992). Despite their cost concerns, the broadcasters may have been willing to support HDTV because of the time lag between the desired immediate outcome of retaining the spectrum and the professed long-term goal of the development of HDTV.

By advocating HDTV the broadcasters would also be able to keep their existing spectrum. The organizations that the broadcasters attracted to their cause were uninterested in the broadcasters' desire to retain spectrum, but the lure of HDTV was powerful for these firms. The electronics companies, for example, saw HDTV as a potential means of breaking back into the consumer electronics market that was dominated by Japanese, European, and Korean companies. For these companies, the broadcasters' inconsistencies would become a major issue, as broadcasters attempted to simultaneously promote HDTV in order to keep their spectrum and impede it in order to forestall the investment it would require.

We believe that these inconsistencies are likely to arise in open systems due to the nature of the technology. Open systems involve multiple subsystems, which are often made by different firms that may be in different industries (Tushman & Rosenkopf, 1992). If a population of firms that is involved in one subsystem attempts to draw other populations into its collective action frame, it will have to create a frame that appeals to these populations, though they

may have very different interests. Our findings from this case study suggest that groups and organizations may create master frames that are inconsistent with their actual beliefs in order to achieve their collective action goals. We consider this inconsistency to be analogous to the way that organizations, facing demands from institutional forces, create structures that buffer their technical cores (Meyer & Rowan, 1977).

EVOLUTION OF HDTV BROADCASTING TECHNOLOGY

Thus far, we have attempted to illustrate that technological evolution can be affected by collective action shaped by the framing attempts of key parties. We now turn to examining the way in which technology can evolve in unexpected ways, changing the outcomes of framing efforts. In this section, we examine the developments after the HDTV race began. We find that the evolution of HDTV at the component level reduced the effectiveness of the broadcasters' framing efforts. Specifically, the development of digital transmission capabilities rendered obsolete some of the systems proposed by HDTV competitors, and increased the possibility for convergence between the television and computer industries.

HDTV technology is an open system technology that is composed of multiple subsystems. Each of these subsystems is subject to evolution that may occur independently of the changes taking place at the system level. This subsystem-level evolution can wreak havoc on organizations' attempts to create standards. Because open systems rely on standards in order to operate effectively, they are subject to strong institutional pressures as organizations, both public and private, take an interest in ensuring compatibility (see Hart, 1994, on the role of government agencies in HDTV technology evolution).

By 1987, the FCC had tacitly supported HDTV by reversing its decision to give unused television spectrum to land mobile firms. The FCC became further involved with HDTV in November 1987, with the formation of the Advisory Committee on Advanced Television Service (ACATS). This committee had representatives from a variety of sources; it was chaired by Richard Wiley, formerly the chairman of the FCC and now a powerful attorney and lobbyist for the television industry. Other members of the committee included representatives from Zenith, ABC, Quasar, CBS, Sarnoff, Bellcore, The Department of State, National Association of Broadcasters, and the Defense Advanced Research Projects Agency (DARPA). Shortly after its inception, ACATS announced that it would hold a competition in which it would test all

available HDTV systems in order to determine the optimal system for the American market.

Two technical problems and one social issue dominate the accounts of early HDTV development. The technical issues were the degree of bandwidth that could be used for HDTV and the new system's compatibility with existing sets. The social issue revolved around fears that the U.S. could cede television production to foreign (Japanese) firms if it made inappropriate choices in HDTV technology development.

Table 4 provides details on some of the systems that were proposed as HDTV standards. As Table 4 shows, there was a disparity in the amount of bandwidth that the different systems required, and in general, more bandwidth meant higher resolution. Thus, the FCC had to define what HDTV meant in the first place. This involved deciding whether enhanced-definition TV systems such as Advanced Compatible TV (ACTV) designed by a consortium of Sarnoff Labs, NBC, Thomson and Philips (hereafter referred to as ATRV, the Advanced TV Research Consortium), that produced a moderate improvement over existing broadcast capabilities, would be treated as equivalent to HDTV.

The issue of compatibility with existing television sets was complicated. NHK's system was incompatible with existing sets, but the addition of a converter would allow NTSC sets in the U.S. to receive the higher definition signal. By 1985, NHK had produced such a converter for the European market, but that did not convince Europeans to support the Japanese standard. In the United States, opinions were divided on the subject of compatibility. Joseph Flaherty of CBS declared that consumers were "undaunted by totally new technologies" and would buy new televisions if the product were shown to be superior (*Broadcasting*, October 26 1987, p. 66). Other participants felt strongly that ensuring that HDTV signals were compatible with the existing standard was important in order to protect the massive investment that consumers had already made in conventional television sets (Iredale, 1988).

The importance of HDTV for American participation in television production and consumer electronics in general was the socio-political issue that dominated newspaper accounts of HDTV in the mid-1980s. As mentioned above, by the 1980s, Zenith was the last remaining American-owned television producer and Japanese firms had a stronghold in consumer electronics in the United States. In this climate, HDTV was both a lure and a threat. Many saw HDTV as a chance for American firms to begin to participate in television production again, and predictions that HDTV would account for many billions in revenue before the year 2000 fueled these hopes.

Separating the compatibility, bandwidth and national competitiveness arguments into technical and socio-political issues as we have done above

Table 4. Specifications of Proposed HDTV Systems.⁷

Proponent (System Name)	Date Introduced ⁸	Scanning Lines	Field Rate (Hz)	Scanning Pattern	Aspect Ratio	Bandwidth Required ⁹ (MHz)	Fate ¹⁰
NHK (MUSE)	1981	1,125	60	P/I	16:9	8.4	Dropped (6/90)
RCA Labs	1985	750	60	P		8.1	Dropped (87); replaced by ACTV 1
Sarnoff Center (ACTV 1)	1987	525	59.94	P	5:3	6	Folded into ATRC Consortium
NYIT (VISTA)	1987	1,125		P/I	5:3	9	Dropped (9/89)
Philips (HDS-NA)	1988	1,250	50		16:9	6+	Folded into ATRC Consortium (01/90)
Sarnoff Center ACTV II	1988	1,050	60			6+	Folded into ATRC Consortium; Dropped after ATTC Test (09/91)
NHK (Narrow MUSE)		1,125	60	P	16:9	6+	Lost out in ATTC test
Faroujida Labs (Super NTSC)	1989	1,050	60			6	Dropped (9/90)

MIT (Channel Compatible)		787.5	59.94	P	16:9	6	Merged with GI, then GA
General Instruments (Digicipher)	1990	1,050	59.94	I	16:9	6	GA
ATRC (Sarnoff, Philips, Thomson, NBC) (Digital Simulcast)	1990	1,050	60			6+	GA
Zenith (Spectrum Compatible)	1990	787.5	59.94	P		6+	Merged with AT&T to form ATA GA
Del Rey (HD-NTSC)	1987	1035				6	Dropped (7/89)

Note: Systems in bold are those that survived to the FCC testing stage.

oversimplifies the case. Clearly, they are interrelated. A policy decision declaring compatibility to be unnecessary would have strengthened the position of the NHK system at the expense of the Sarnoff consortium's system. Conversely, forcing the systems within the existing bandwidth was advantageous to ACTV at the expense of other proposed systems.

In September 1988 the FCC decided that compatibility with existing sets was mandatory. Soon, 23 proposals by 14 firms for would-be HDTV systems arrived at ACATS. Most of these were soon withdrawn or discovered to be technically infeasible. The FCC further narrowed the field of competitors in March 1990 by declaring that only simulcast systems would be considered. Simulcast systems sent an entire signal that existing NTSC sets could receive through a single 6 MHz channel. If necessary, the system sent additional information on a separate channel and then reconciled the two signals at the receiver. Alternative proposals, known as augmented systems, required extra bandwidth for any usable signal to be received.

By the summer of 1989, testing of the various HDTV plans was less than a year away, and eleven systems representing eight proponents remained in the competition. The firms and consortia involved were: NHK (three systems), MIT (two systems), Faroudja Laboratories, Sarnoff Research Center, New York Institute of Technology, Philips, Production Services Inc., and Zenith. A few months later, NYIT dropped out of the competition, citing funding shortages, and Faroudja dropped out to pursue a different technology that would enhance existing television sets. Production Services Inc withdrew from the competition shortly before testing began, citing concerns that complying with the testing process's disclosure requirements would reveal proprietary information that could hurt its position in other markets.

Before discussing the progression of the systems any further, it is necessary to describe the subsystems of HDTV in more detail. High Definition Television consists of five key components (see Fig. 1). The *encoder* is the key piece of technology that compresses audio and video signals from sources such as 35-millimeter movies or television studio productions into a digital signal. For high definition signals, the encoder is of paramount importance because it contains the compression algorithm that allows a signal that would otherwise occupy up to 270 Mb/s to be compressed to a stream of only three Mb/s, allowing it to be broadcast over standard channels (*IEEE Spectrum*, 1995). The encoded signals are then packaged into signals ready for transmission by a *transport* device. The purpose of this device is to take the constant stream of digital information and split it into packets of information to create the flexibility for which digital television is valued. Brinkley (1997) likens this process to sending groups of soldiers into battle; whereas analog signals have to be sent out as one

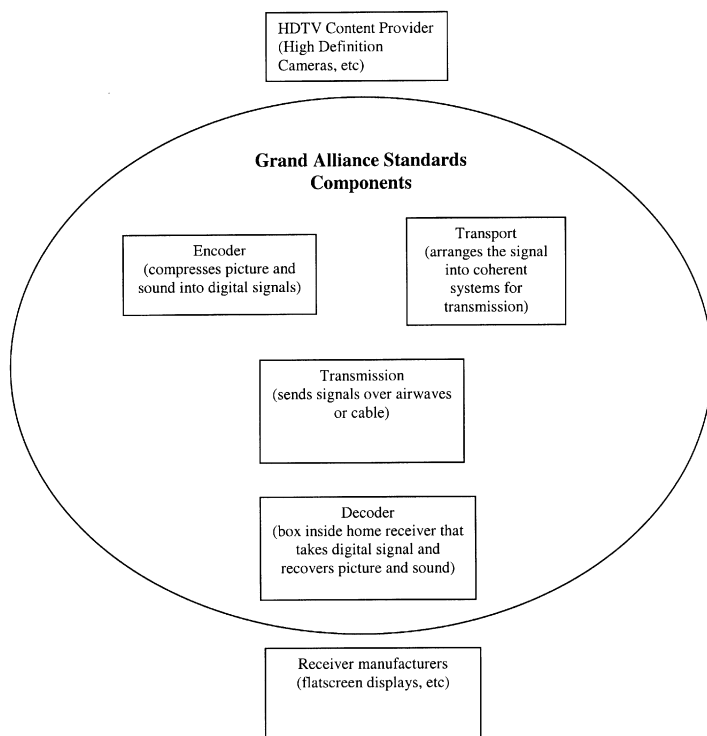


Fig. 1. Main Components of HDTV Technology.

Notes:

1. Decoder – the box inside the home receiver that takes the digital signal and turns it into picture and sound. Philips made this component.
2. Encoder – the device that compresses the picture and sound and readies it for transmission as digital signals. This is known as the heart of the HDTV system, and it was jointly produced by AT&T and General Instrument.
3. Transmission system – the method of sending the broadcast signal over the airwaves and/or cable. This standard was established through a competition between Zenith and General Instruments. Zenith won the competition.
4. Transport – the hardware that arranges the signals into coherent systems for transmission. Sarnoff (Thomson) was chosen to develop this part of the system.
5. Audio system – the digital audio standard that would be embedded in the receiver. A three-way competition was established for this component. The competitors included two Grand Alliance firms, General Instruments and Thomson, and an independent firm, Dolby. The Dolby system was chosen as the winner, placing this component outside of the Grand Alliance network.

continuous formation of soldiers, digital signals can be sent in any combination and re-assembled at the television receiver. This flexibility allows for multiple programs to be broadcast within the same bandwidth. From the transport, the signal is sent to a *transmitter*, which sends it out to antennae or through cable or satellite signals. At the receiver, a *decoder* makes sense of the signals and allows the television to create the picture and sound. Finally, the *audio system* defines the digital sound standard used.

We focus on the evolution of the encoder because it represents a core piece of the HDTV system and because changes in encoder technology were extremely important. Given the bandwidth required for early HDTV systems, the compression algorithms contained in the encoder would be vital in determining the feasibility of HDTV. The initial NHK MUSE system was broadcast over satellite signals that required 30 MHz of bandwidth, but later improvements in compression technology allowed the signal to be reduced to 8.4 MHz. Still, the signal was incompatible with existing television sets, and too wide for the six MHz allocated to each U.S. broadcast channel.

As the testing neared, the various systems represented a variety of approaches to HDTV (see Table 4). All systems shared one common attribute, however: they were based upon analog signals. Though digital audio technology was widely available, it was commonly believed that digital video signals were several years away from being viable, if indeed they would ever be seen. In fact, Joseph Flaherty from CBS remarked in early 1990 that digital television “defies the laws of physics (*The Economist*, February 12, 1990).”

Flaherty’s statement was somewhat less than prophetic. In June 1990, General Instruments announced that it would develop an all-digital system, based on technology that the company had utilized in its cable and satellite signal scrambling business. Digital signals offered the opportunity to manipulate and compress the signal much more easily, and thereby send a truly high definition signal through the existing six MHz pipeline.

The GI system, named “Digicipher,” changed the HDTV race substantially. Other entrants soon announced plans to introduce digital systems, but their lack of experience with digital technologies forced them to partner with other firms. The ATRV consortium soon followed suit in November 1990, when they introduced a digital system in addition to their analog ACTV-II system. In July 1991, the ATRV consortium invited Compression Labs to provide their video compression technology. Zenith scrapped its existing proposal in December, 1990 and joined forces with AT&T to produce a digital system. This consortium was joined in December 1991 by Scientific Atlanta, which provided compression technology. Finally, MIT formed a joint venture called American TV Alliance (ATV) with GI in January, 1991 to make the switch from a hybrid analog/digital system to an

all-digital one. GI continued developing its Digicipher technology with a different scanning format than the one used in the joint venture with MIT.

By April 1991, the testing was set to begin with four alliances entering six systems in the competition. The competition itself was drawn-out and solved little, eliminating only two systems. Narrow MUSE from NHK and ACTV-II from the ATRV consortium each performed poorly in the FCC tests. The ATRV consortium had simultaneously entered a digital system and dropped ACTV-II immediately after it was tested. NHK withdrew after all of the systems had been tested. The Japanese participation in HDTV was eliminated not through political maneuvering but rather from the evolution of a component technology, as the introduction of a digital system effectively rendered NHK's technology obsolete.

The removal of NHK and one of the ATRV consortium systems left four systems vying for the broadcasting standard. The FCC proposed a second round of testing. MIT and GI then proposed a "Grand Alliance" between the remaining contestants. On May 20, 1993, the Grand Alliance was formed. At least three reasons lie behind the companies' decision to band together at this particular point. First, the alliance allowed the firms to share the risk of developing the system. The expense of further testing was considerable, especially to MIT-GI, and Zenith, neither of which had slack resources available to finance further testing or could afford to have their system lose in future rounds, so sharing the risk was particularly attractive to these companies. Second, the alliance ended the possibility of an infinite regress developing, in which losing firms continually challenged results and prolonged the testing. The entrants with greater resources at their disposal could afford to follow such a strategy, and would likely be able to eliminate the poorer entrants. Protests had been rampant in the first round of testing, and given the nature of the technology, in which no single criteria could be used to name a superior system, the protests could go on indefinitely. Finally, Dick Rumsfeld, president of General Instrument and former and future United States Defense Secretary, indicated that the alliance would give the HDTV entrants greater bargaining power with the FCC, because the system would have no competitors.

Though the alliance ended the official race to create the HDTV standard, the standard was by no means completely developed. Major decisions remained over the system's specifications, including the sound standard to be adopted, and which company would produce which components. These decisions were gradually settled through negotiations between the alliance organizations and competitions at the component level. Table 5 shows that firms with greater technological capabilities in component HDTV technologies assumed a leadership position within the four alliances and later in the Grand Alliance in the choice of those component technologies.

Table 5. Patents Held by Grand Alliance Members in the Components of HDTV.

Firm	Encoder		Decoder		Digital Compression		Transmission	
	Until 3/1/89 ^a	3/1/89 to Grand Alliance	Until 3/1/89	3/1/89 to Grand Alliance	Until 3/1/89	3/1/89 to Grand Alliance	Until 3/1/89	3/1/89 to Grand Alliance
Zenith	4	0	12	1	0	1	2	17
AT&T	0	2	0	1	8	12	4	7
Sarnoff ^b	15	2	12	1	6	13	13	14
Philips	14	10	13	13	7	4	27	20
General Instruments	0	0	0	1	0	4	2	4
MIT	0	1	0	0	3	1	3	5
Total for Grand Alliance Members	33	15	37	17	24	35	51	67
Total Patents in area	110	53	106	50	424	266	186	167

^a This is the date on which AT&T and Zenith announced their partnership, which was the first alliance among U.S. HDTV competitors.

^b Sarnoff total includes patents assigned to RCA and Thomson.

Bold entries indicate components in which the firm took a leadership position within the Grand Alliance. Digital compression algorithms are relevant for the encoder and transport components.

EVOLUTION OF OPEN SYSTEMS TECHNOLOGY AND COLLECTIVE ACTION

As the Grand Alliance worked toward a standard for HDTV, two challenges arose, both resulting from the evolution of the encoder technology that enabled digital broadcasts. The two challenges illustrate how difficult it is to sustain collective action as open systems technologies evolve. The first challenge to HDTV proponents came from the computer industry, as the advent of digital broadcasts blurred the differences between computers and televisions, and computer firms recognized both opportunities and threats from the convergence of the two industries. The computer firms' interest illustrates an important aspect of collective action framing efforts in open systems, as evolution at the component level affected the degree to which outside populations became interested in the standard-setting process. In the language of organizational ecology, the introduction of digital encoding increased the niche overlap between the broadcasting and computing industries, which brought them into competition for resources (Hannan & Freeman, 1989).

The second challenge arose from the internal inconsistencies in the broadcasters' frames, as they began to discourage HDTV in favor of digital television (DTV). As we discuss below, this issue illustrates how changes at subsystem levels can lead to diverging interests among organizations that were initially involved in collective action. Taken together, the two developments demonstrate the frailty of framing efforts when technological change occurs; technological change can both reduce the original collective action frame's relevance for a group of actors and increase interest among members of organizational populations that were previously uninvolved.

HDTV Technology Evolution and Collective Action in the Computer Industry

Even as the Grand Alliance worked to create a standard for HDTV, the computer industry began to enter the fray, attempting to make the new broadcast standard more computer-friendly. The computer firms were concerned with one aspect of the HDTV standard in particular, which was the decision over whether interlaced or progressive scanning would be used for the system. Interlaced scanning is the format used by NTSC televisions. It relies on the fact that television pictures change much faster than the human eye can perceive. Thus, a television that receives a frame every 1/60th of a second receives more information than it really needs, and in order to preserve bandwidth, a system was devised that sent half of a frame every 1/30th of a second. The frames are

spliced at the broadcasting source, and recombined, or interlaced, at the home receiver. Though this has some minor effects on the picture quality, such as blurring lines, it works relatively well and preserves bandwidth. The small imperfections that interlaced scanning causes are very noticeable if one sits close to a television set, but from a short distance away the imperfections are rarely problematic.

Progressive scanning sends the entire frame at one time and “paints” it across the screen. It requires more bandwidth, but eliminates the problems with picture quality that interlacing causes. Progressive scanning became necessary for desktop computing because computer screens are both dependent upon fine-grained images and viewed from much closer than television sets. Thus, if the HDTV standard called for interlaced broadcasts, computers would essentially be unable to display the images, as viewers would not accept interlaced images on computer screens, given the size of the screens and the proximity from which they are normally viewed.

Thus, what seems like a relatively obscure aspect of a television system became vitally important to HDTV evolution in the mid-1990s. The Zenith-AT&T system utilized progressive scanning. Their system had to sacrifice some scanning lines compared to the interlaced systems (see Table 4), but they argued that the improvement in picture quality more than compensated for the slightly decreased detail. Still, there were proponents of interlaced scanning such as Philips and Thomson among the Grand Alliance members and throughout the television broadcasting industry, and interlace was chosen as the scanning format.

When firms in the computer industry learned that Zenith-AT&T had developed a progressive scan system, they began to consider seriously the convergence of HDTV and computers. Such a convergence had been discussed for years (see, for example, Schlender, 1989, p. 111). With the advent of progressive scanning and digital signal transmission, however, the lines between television and computing were blurred. Computers would be able to interpret and display the progressive, digital signals used in HDTV transmission, although it was impractical to display the degree of resolution present in HDTV signals on screens as small as those used for computers. Conversely, HDTV sets could be used interactively, so viewers could request customized programming and information, and the television would become an information appliance.

Previously, some firms from the computer industry had involved themselves in the HDTV debate through the American Electronics Association. This involvement had mostly been limited to trumpeting the importance of HDTV for basic technology industries that were important to both televisions and computers, such as semiconductor production. Now, however, companies such as Apple and Microsoft began to envision themselves as directly benefiting from

HDTV. The computer industry, with some help from entertainment studios, began to attack the Grand Alliance's preference for interlaced scanning. Collective action on the part of computing firms included forming a lobbying group called Americans for Better Digital TV, which pressured Congress and the FCC toward forcing a progressive scanning format on HDTV. As the computer industry weighed in on the issue of progressive v. interlaced scanning, it also advocated a lower-definition format. Microsoft, Compaq, and Intel began to advocate a 480 line progressive-scanning format at the Consumer Electronic Show. They argued that such a format would speed the convergence between television and personal computers. Coincidentally, the 480-progressive format would allow television receivers no advantage over computer screens in terms of definition, effectively nullifying years of HDTV technology development aimed at increasing screen resolution.

There was no obvious middle ground between the Grand Alliance's system and the computer lobby's proposal. The Grand Alliance chose not to decide the issue at all, instead choosing to solve the war over standards by incorporating 18 potential formats into the specifications for the system. The FCC, in turn, promoted the Grand Alliance system, leaving it to the market to determine which of the formats would persevere. Thus, both broadcasters and television manufacturers could choose which standards they would follow. This allowed HDTV to go ahead at a time when it appeared that the lack of agreement between the Grand Alliance and computer firms could delay or even cancel HDTV efforts. It also resulted, not surprisingly, in massive confusion, as there was no longer any guarantee that a particular television could receive a particular broadcaster's signal (there was no way for the Grand Alliance or the FCC to legislate how HDTV programming should be produced or received). Broadcasters have chosen different standards. For example, NBC employs a 1080 line, interlaced format, ABC has chosen a 720 line progressive format, and Fox uses a 480 line progressive format (Snider, 2002).

HDTV Technology Evolution and New Market Opportunities

While the Grand Alliance's decision to leave the display formats out of the HDTV standard may seem strange, it was done in order to bring a degree of closure to the process and avoid further attempts to delay the introduction of HDTV. As mentioned earlier, one of the attempts to delay or even abandon HDTV came from the broadcasters, as they considered other possible uses for the technology unleashed by the HDTV contest. The advent of digital signals had opened up new opportunities for broadcasters, as the flexibility of digital signals offered the possibility that broadcasters could offer interactive

programming or could send several signals over the same channel. Compression algorithms had developed to such an extent that most programming did not require the full six MHz bandwidth allowing broadcasters to offer several programs over a single channel. Sending several signals, known as multicasting, enables a single television channel to carry multiple programs at once, potentially adding greatly to the advertising revenue available per MHz of spectrum used. This extra bandwidth could also be used to offer interactive television, providing the opportunity for shopping, playing games, or enabling Internet access over the television.

Either of these options was more attractive to most broadcasters than was the prospect of investing millions per station in HDTV technology with little belief that the higher definition signals would increase their audience and advertising revenue. Broadcasters began to re-frame their arguments away from HDTV and toward DTV, in hopes that the spectrum that they had retained could be used for profitable purposes. Typical of their arguments was the one offered by Rupert Murdoch of Fox Broadcasting who maintained, "High definition is a luxury. Compared with a modern TV set, it's not that different. Why shouldn't that extra spectrum be given to you or me or anyone to put on that extra number of channels? (Brinkley, 1997, p. 304)." Similarly, John Abel of the NAB claimed, "HDTV is just for rich people. DTV is for everybody (Brinkley, 1997, p. 322)."

However, even as they argued for the freedom to use their second channel for DTV, broadcasters were still using HDTV to obtain desired benefits. For example, facing legislation that curbed energy use, the NAB argued that if they were subject to more stringent energy efficiency rules, it would delay the introduction of HDTV (Brinkley, 1997, p. 316). This two-faced approach to the HDTV issue may have been rational for broadcasters, but it created animosity toward them in two groups that they had attracted in their initial attempt to retain spectrum several years earlier. Both the organizations in the Grand Alliance, which had spent several years and millions of dollars creating HDTV technology, and the government that had protected the broadcasters' spectrum, were unwilling to see HDTV abandoned.

Congress reacted particularly negatively to the broadcasters' intention to use their spectrum for anything other than high definition broadcasts and threatened to auction off the spectrum to the highest bidder. This threat was credible, as spectrum auctions had generated over \$9 billion in revenue from cell phone companies (*Broadcasting and Cable*, April 10, 1995). In general, the broadcasters had little chance of generating a coherent frame around DTV; there were no obvious enemies to combat and the broadcasters were unable to attract other organizations to this frame. Pressure from Congress and the manufacturers in the Grand Alliance forced the broadcasters to capitulate, and by 1998,

the networks were pledging their support for HDTV (Pope, 1998). At present, each television broadcaster is allotted two channels, one of which is to be used for analog standard-definition broadcasts, while the other is to simulcast programs in HDTV. By May 2002, all television stations must have a digital signal; the channel that carries the simulcast analog signal must be returned to the FCC by 2006, but only if digital television has reached 85% penetration by that time. There is little likelihood that penetration will be 85% within eight years of HDTV's debut; color television required 22 years to reach 85% (*Broadcasting and Cable*, November 16, 1998). As of January, 2002, only 229 of the 1200 TV broadcast stations had converted to digital transmission with the number expected to go up to 800 by the end of the year (Snider, 2002).

At present, the industry faces considerable uncertainty. Though some television studios have migrated to digital signals ahead of the schedule mandated by the FCC (*Broadcasting and Cable*, 2001), the amount of DTV and HDTV content available is still low. Sales of television receivers capable of receiving the updated signals are much slower than industry forecasts had predicted. Some estimates suggest that 625,000 HDTV sets were sold in 2000 and 715,000 between January and August 2001 (Husted, 2001). Recently, *Business Week* referred to the effort to create HDTV as "perhaps the biggest blunder of 20th century communications policy (Carney, 2001)." While that statement is open to debate, it is clear that the development of HDTV has been difficult and costly, and even with more than 15 years and untold millions of dollars invested, the degree to which HDTV will diffuse among television viewers is still unclear.

CONCLUSION

In this final section of this paper, we draw on the HDTV example in order to develop a conceptual model of the role of collective action in shaping technological change. Figure 2 depicts the influence of sociopolitical processes on technological change as described by Tushman and Rosenkopf (1992). They argue that sociopolitical processes assume greater importance during eras of ferment when many technological alternatives are available. These processes are of greater consequence in the evolution of open systems technologies, especially that of core subsystems that link component technologies.

Figure 3 provides greater detail on how sociopolitical processes may play out in the evolution of open systems technologies such as HDTV. The identification of an external threat and the discovery of new market opportunities motivate attempts by industry participants to respond collectively. Mobilization of collective action requires the creation of a collective action frame, one that is credible and that resonates with industry participants.

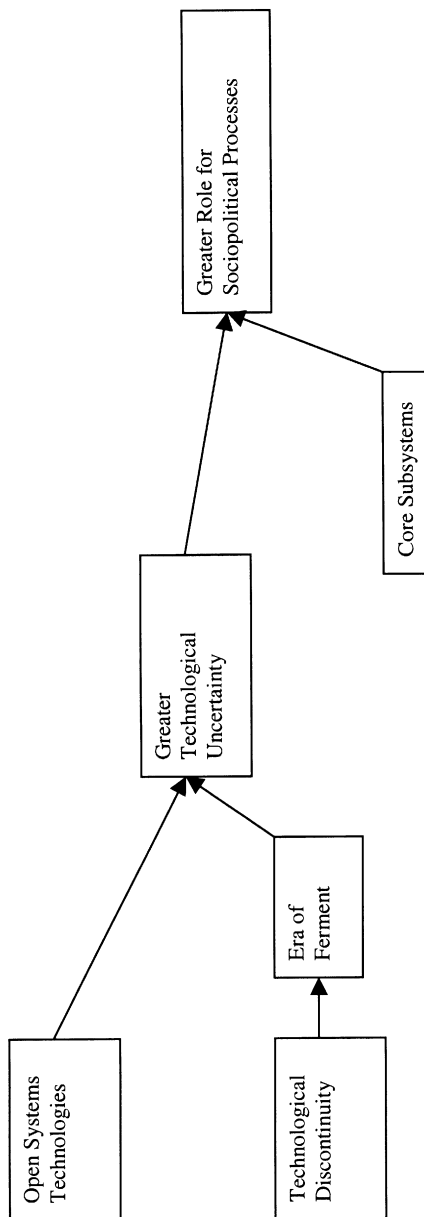


Fig. 2. Technological Change and Sociopolitical Processes.

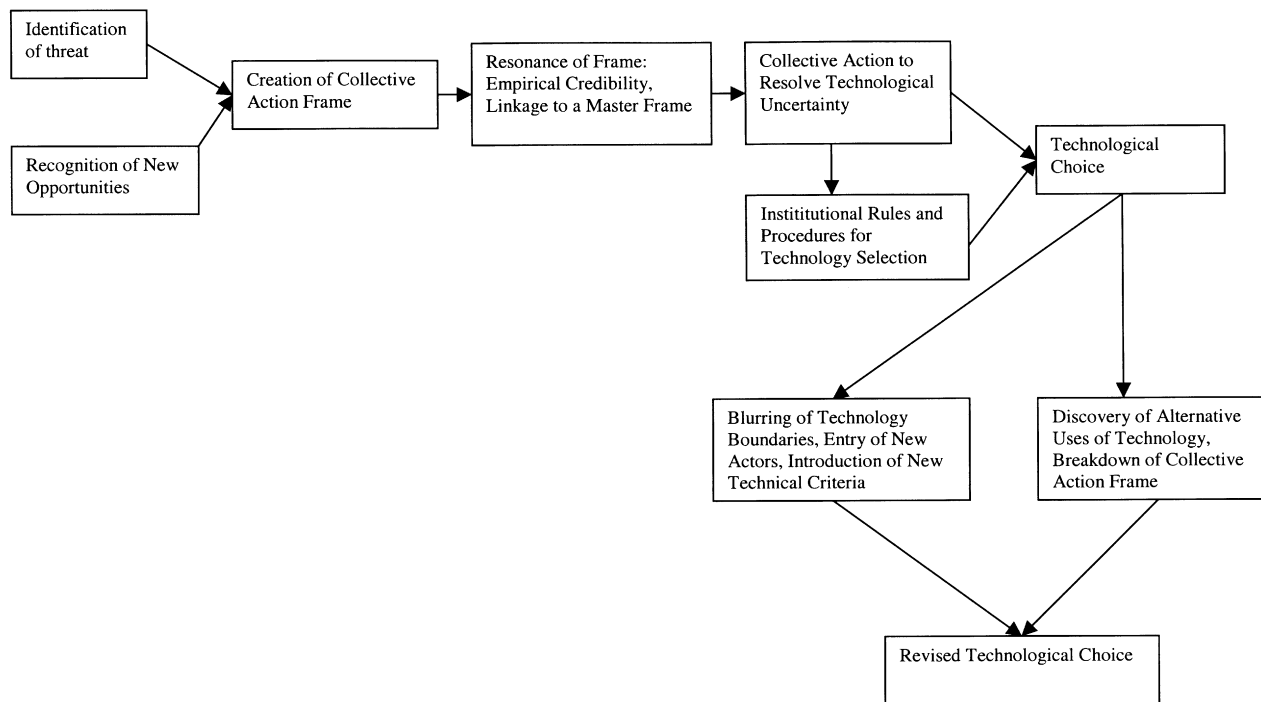


Fig. 3. Sociopolitical Processes and the Evolution of Open Systems Technologies.

Institutional rules designed by industry participants and regulatory agencies are then used to select one among several technological alternatives. But this initial technological choice may itself lead to a pressure for further change from at least two sources. First, components of open systems technologies may evolve such that technologies converge across multiple industries. This is likely to lead to the involvement of firms from other industries that introduce new technical criteria into the technology selection process. Second, industry participants may discover alternative uses for the technology, leading to a breakdown of the initial collective action frame. Both of these pressures will likely lead to further modifications in the technology that is adopted by industry participants.

The framing efforts undertaken by U.S. broadcasters illustrate the way in which technological trajectories can be affected by social forces. At the same time, the HDTV example demonstrates how framing efforts are susceptible to becoming obsolete due to changes in subsystem technology. We believe that obsolescence of collective action frames is particularly likely in the case of open systems, which are characterized by several interdependent subsystems, each of which may evolve due to developments inside or outside of the larger system. HDTV technology development was thus affected by changes that were largely out of the control of individual participants in the technology selection process.

NOTES

1. Collective action frames are interpretive frames used by organizations to generate collective action in an organizational field. Scholars of technology and organization have also used the frame concept in a different sense – the interpretive frame that individual members of organization use to understand technology within organizations (Orlikowski & Gash, 1994).

2. NTSC stands for National Television Systems Committee. The NTSC standard was set for black and white televisions in 1941 and was revised for color televisions in 1953.

3. Frames are related to, but are distinct from, ideologies. Specifically, Benford and Snow (2000) argue that ideologies represent coherent and pervasive world views that affect perceptions of everyday life, while frames are actions that seek to amplify or shatter existing ideologies.

4. As Benford and Snow (2000) outline, the frame's resonance depends upon several factors, including the credibility of the organization that proffers the frames. We concentrate here on the empirical credibility and consistency of the frames; as we demonstrate in our discussion of HDTV, these are the most salient for this setting.

5. A less comprehensive version of this legislation had been enacted in 1972, when some spectrum had been put aside for land mobile use in the ten largest U.S. cities. The 1985 legislation sought to set aside additional spectrum.

6. Brinkley (1997) reports that the NAB did try to target Motorola as an opponent in their initial battle against land mobile. The NAB lobbyists found evidence of Motorola's poor environmental practices and alleged abuse of government contracts. The attempt to produce animosity toward Motorola was entirely unsuccessful.

7. Other systems for which we do not have technical information include those proposed by NHK (Muse 9 and Muse E), Osborne, Hi-Resolution Sciences, Dolby, Japanese BTA, Scientific Atlanta, Production Services Inc., and MIT (2nd system).

8. We use the date that the product's specifications were announced as the introduction date. This is more appropriate than the date that the system itself debuted for two reasons. First, it is the announcement of specifications that often seems to spur competing firms/alliances into action, or affects policy decisions. Second, many systems that never evolved into working prototypes, such as RCA's initial attempt to build HDTV, clearly had impact on other actors' decisions.

9. 6+ indicates that the system required more than the six MHz allotted to conventional broadcasts, but that the extra spectrum could be broadcast on a separate channel and then re-combined at the receiver. Early systems such as the initial NHK entry could not accomplish this, and thus required contiguous spectrum.

10. GA indicates that the system became part of the "Grand Alliance."

ACKNOWLEDGMENTS

We would like to thank Beth Bechky, Andy Hargadon, Paul Ingram and participants at the 2nd Annual Davis Conference on Qualitative Research for their insightful comments on earlier versions of this paper.

REFERENCES

- Abramovitz, M. (1987). Broadcasters woo with Crisper Shots; Special TV Show part of Effort to Preserve UHF Frequencies. *Washington Post* (January 4).
- Aldrich, H. E. (1999). *Organizations Evolving*. London: Sage Publications.
- Aldrich, H. E., & Fiol, M. C. (1994). Fools Rush in? The Institutional Context of Industry Creation. *Academy of Management Review*, 19(4), 645-670.
- Babb, S. (1996). 'A true American system of finance': Frame resonance in the U.S. labor movement, 1866 to 1886. *American Sociological Review*, 61, 1033-1052.
- Barley, S. R., Freeman, J., & Hybels, R. C. (1992). Strategic Alliances in Commercial Biotechnology. In: N. Nohria & R. G. Eccles (Eds), *Networks in Organizations: Structure, Form, and Action* (pp. 311-347). Cambridge, MA: Harvard University Press.
- Beltz, C. A. (1991). *High Tech Maneuvers: Industrial Policy Lessons of HDTV*. Washington, D.C.: The AEI Press.
- Benford, R. D. (1987). *Framing Activity, Meaning, and Social Movement Participation: The Nuclear Disarmament Movement*. Ph.D. Thesis, University of Texas.
- Benford, R. D., & Hunt, S. A. (1992). Dramaturgy and Social Movements: The Social Construction and Communication of Power. *Sociological Inquiry*, 62, 36-55.
- Benford R. D., & Snow, D. A. (2000). Framing Processes and Social Movements: An Overview and Assessment. *Annual Review of Sociology*, 26, 611-639.

- Brinkley, J. (1997). *Defining Vision: The Battle for the Future of Television*. New York, NY: Harcourt Brace.
- Broadcasting (later, *Broadcasting and Cable*), various issues.
- Cable, S., & Shriver, T. (1995). Production and Extrapolation of Meaning in the Environmental Justice Movement. *Sociology Spectrum*, 15, 419–442.
- Carney, D. (2001). HDTV: Don't Blame the FCC for Tuning Out. *Business Week* (February 5), 52.
- Davies, S. (1999). From Moral Duty to Cultural Rights: A Case Study of Political Framing in Education. *Sociology of Education*, 72, 1–22.
- Dupagne, M., & Seel, P. B. (1998). *High-Definition Television: A Global Perspective*. Ames, IA: Iowa University Press.
- Garud, R. (1994). Cooperative And Competitive Behaviors During The Process Of Creative Destruction. *Research Policy*, 23, 385–394.
- Garud, R., & Rappa, M. A. (1994). A socio-cognitive model of technology evolution: The case of cochlear implants. *Organization Science*, 3, 344–362.
- Hafner, K., & Lyon, M. (1996). Casting the net. *Sciences*, 36, 32–37.
- Hannan, M. T., & Freeman, J. (1989). *Organizational Ecology*. Cambridge, MA: Harvard University Press.
- Hart, J. A., & Tyson, L. (1989). Responding to the Challenge of HDTV. *California Management Review* (Summer), 132–145.
- Hart, J. A. (1994). The Politics of HDTV in the United States. *Policy Studies Journal*, 22(2), 213–228.
- Holburn, G., & Vanden Bergh, R. Policy and Process: A Framework for the Design of Non-Market Strategy (this volume).
- Husted, B. (2001). High-Definition Television: To Buy or not to Buy. *The Atlanta Journal and Constitution* (November 1), 8BE.
- IEEE Spectrum* (1995). HDTV and the New Digital Television. (April), 34–81.
- Ingram, P., & Clay, K. (2000). The Choice-Within-Constraints New Institutionalism and Implications for Sociology. *Annual Review of Sociology*, 26, 525–546.
- Ingram, P., & Inman, C. (1996). Institutions, Intergroup Competition and the Evolution of Hotel Populations around Niagara Falls. *Administrative Science Quarterly*, 41, 629–658.
- Iredale, R. J. (1988). Clearing the HDTV Picture. *Electronic Engineering Times* (August 8), 38.
- Katz, M., & Shapiro, C. (1985). Network Externalities, Competition, and Compatibility. *American Economic Review*, 75, 424–440.
- Klandermans, B. (1997). *The Social Psychology of Protest*. London: Blackwell.
- Lambert, P. (1992). HDTV Push Feels Like Shove to Broadcasters. *Broadcasting* (Jun 29), 122(27), 3, 13.
- Lieberman, M. B., & Montgomery, D. B. (1998). First-mover (dis)advantages: Retrospective and link with the resource-based view. *Strategic Management Journal*, 19, 1111–1126.
- McKendrick, D. G., & Carroll, G. R. (2001). On the Genesis of Organizational Forms: Evidence from the Market for Disk Arrays. *Organization Science*, 12, 661–682.
- Meyer, J. W., & Rowan, B. (1977). Institutionalized organizations: Formal structure as myth and ceremony. *American Journal of Sociology*, 83, 340–363.
- Morris, A. D. (1992). Political Consciousness and Collective Action. In: A. D. Morris & C. M. Mueller (Eds), *Frontiers in Social Movement Theory* (pp. 351–373). New Haven, CT: Yale University Press.
- Mueller, M. (1983). Spectrum Fees vs. Spectrum Liberation. *Regulation* (May/June), 21–27.
- Nelson, R. R., & Winter, S. G. (1982). *An Evolutionary Theory of Economic Change*. Cambridge, MA: Harvard University Press.

- Orlikowski, W. J., & Gash, D. C. (1994). Technological frames: Making sense of information technology in organizations. *ACM Transactions on Information Systems*, 12, 174–207.
- Pinch, T. J., & Bijker, W. E. (1987). The social construction of facts and artifacts: Or how the sociology of science and the sociology of technology might benefit each other. In: W. E. Bijker, T. P. Hughes & T. Pinch (Eds), *The Social Construction of Technological Systems* (pp. 17–50). Cambridge, MA: MIT Press.
- Pope, K. (1988). TV Plans to Use New Digital Capacity to Improve Picture, Not Add Channels. *Wall Street Journal* (April 3), A3.
- Rao, H., Morrill, C., & Zald, M. N. (2000). Power Plays: How Social Movements and Collective Action Create New Organizational Forms. In: B. M. Staw & R. I. Sutton (Eds), *Research in Organizational Behavior*, 22, 239–282. Greenwich, CT: Elsevier.
- Schlender, B. R. (1989). Couch Potatoes! Now it's Smart TV. *Fortune* (November 20), 111–115.
- Sherif, M., Harvey, O. J., White, B. J., Hood, W. R., & Sherif, C. W. (1961). *Intergroup Conflict and Cooperation: The Robbers Cave Experiment*. Norman, OK: Institute of Intergroup Relations.
- Snider, M. (2002). The Line on the Super Bowl: Digital but not HDTV. *USA Today* (January 15), 6D.
- Snow D. A., & Benford, R. D. (1988). Ideology, Frame Resonance and Participative Mobilization. *International Social Movement Research*, 1, 197–218.
- Stinchcombe, A. L. (1965). Social structure and organizations. In: J. G. March (Ed.), *Handbook of Organizations*. Chicago: Rand-McNally.
- Swaminathan A., & Wade, J. B. (2001). Social Movement Theory and the Evolution of New Organizational Forms. In: C. B. Schoonhoven & E. Romanelli (Eds), *The Entrepreneurship Dynamic in Industry Evolution* (pp. 286–313). Palo Alto, CA: Stanford University Press.
- The Economist* (1994). From Idiot Box to Information Appliance (February 12), SS5.
- Tushman, M., & Rosenkopf, L. (1992). On the Organizational Determinants of Technological Change: Towards a Sociology of Technological Evolution. In: B. Staw & L. Cummings (Eds), *Research in Organization Behavior* (Vol. 14, pp. 311–347). Greenwich, CT: JAI Press.
- Valocchi, S. (1996) The Emergence of the Integrationist Ideology in the Civil Rights Movement. *Sociology Problems*, 43, 116–130.
- Vanhaverbeke, W., & Noorderhaven, N. G. (2001). Competition Between Alliance Blocks: The Case of the RISC Microprocessor Technology. *Organization Studies*, 22, 1–30.
- Wade, J. (1995). Dynamics of Organizational Communities and Technological Bandwagons: An Empirical Investigation of Community Evolution in the Microprocessor Market. *Strategic Management Journal*, 16, 111–133.
- Zuo, J., & Benford, R. D. (1995). Mobilization Processes and the 1989 Chinese Democracy Movement. *Sociological Quarterly*, 36, 131–156.

This Page Intentionally Left Blank

THE EVOLUTION OF UNIVERSITY PATENTING AND LICENSING PROCEDURES: AN EMPIRICAL STUDY OF INSTITUTIONAL CHANGE

Bhaven N. Sampat and Richard R. Nelson

ABSTRACT

In a recent paper (Nelson & Sampat, 2001) we proposed that it is fruitful to conceptualize institutions as “social technologies” that are standard among economic actors in particular contexts. This paper extends the social technology concept to study institutionalization and institutional change, based on a case study of the history of social technologies used by universities to manage their patenting and licensing activities. While at the beginning of the twentieth century, universities avoided patenting and licensing activities, today all research universities have “technology transfer offices” to patent and market faculty inventions. That is, this social technology has become an institution. Based on historical narrative, we argue that the social technologies orientation highlights several important aspects of institutional change that are not prominent in the mainstream institutionalist literatures. Moreover, the evolution of social technologies has interesting parallels to the evolution of physical technologies.

1. INTRODUCTION AND BACKGROUND

There has recently been a resurgence of interest in “institutions” in the economics literature, and across the social sciences more generally. In a recent paper (Nelson & Sampat, 2001) we showed that even within the economics literature the term “institution” means different things to different authors. We argued that an important consequence of this variety of “institutionalisms” is that a broad and cumulative empirical research program on institutions and institutional change has yet to emerge.

However our survey of the literature also suggested an important commonality across the institutions concepts in the literature: Veblen’s “general habits of action and thought,” Schotter’s (1981) “how the game is played,” Commons’ (1924) “working rules,” North’s (1990) “rules of the game,” and Williamson’s (1975, 1985) “governance structures,” all have a family resemblance. Specifically, each of these economists have aimed to call attention to standardized modes of transaction and interaction employed by actors to get things done, in contexts where effective performance requires coordinated action. However, authors have differed in the extent to which their focus is on the standardized behavior patterns per se, or the factors that mold and support standardized behavior. And among those focused on the reasons for standardized behavior patterns, there are differences in factors viewed as central in “institutionalizing” behavior.

Our view of what is common across the various concepts of institutions, and what is divergent, led us to propose that the most promising focus for a theoretical and empirical research program, which has the promise of being cumulative, would be on the generally employed behavior patterns per se. This focus has two major advantages. First, it orients the analysis, right from the start, towards the phenomena of interest. Further, the object of such a focus is describable in reasonably straightforward ways, and can be studied empirically through a variety of different routes. Second, the analysis is open regarding the factors and forces that lie behind the standardized behavior in question. In some cases there might be well-recognized and sharp “rules of the game,” perhaps but not necessarily articulated in a body of formal law. Particular organizational forms might be involved in an essential way, or they might not be. Strong norms might or might not be associated with the body of practice. What lies behind, what causes the emergence of, and what holds in place, standardized patterns of human and organizational transacting and interacting should be an open research question.

An emphasis on the behavior patterns led us towards conceptualizing institutions as “social technologies” that are standardized in particular contexts.

Social technologies are routines for inter-organizational or intra-organizational interaction. When they are standard in particular contexts, i.e. employed by most relevant actors in those contexts, they become institutions. In modern capitalist societies “using the market” and “respecting property rights” are important institutions, i.e. social technologies that are standard across most actors. Some social technologies are standard only in particular contexts. Thus the set of routines associated with an “industrial research laboratory” and the “multidivisional form of governance” are standard in some, but not all, industries. The social technology “accept cigarettes in return for goods and services” is standard in prisons, but not in most other contexts. The focus of this essay will be on a particular context: research universities. In particular, we focus on social technologies used by universities to manage the patenting and licensing of their inventions, and the dynamics of the standardization of these routines.

Before proceeding, it is worth clarifying the concept of an organizational “routine,” as it is central to the remainder of this paper. As developed in Nelson and Winter (1982), a routine involves a collection of procedures which, taken together, result in a predictable and specifiable outcome, like a recipe. Complex routines, of the sort associated with the production of goods and services, almost always can be analytically broken down into a collection of subroutines. The program built into a routine generally involves two different aspects: a recipe that is anonymous regarding any division of labor or coordination, and a division of labor plus a mode of coordination. In Nelson and Sampat (2001), we proposed that the former is what scholars often have in mind when they think of “technology” in the conventional sense; this aspect of a routine is the “physical technology” involved. The latter aspect of the routine, that which involves the coordination of human action and interaction, is the “social technology” involved.¹

This paper explores the social technologies concept further, and extends it to study the process of institutional change. Most writings in the various institutionalist literatures agree that institutional change is poorly understood. Douglass North made this point forcefully in his 1993 Nobel Address, and in a recent paper on institutionalism entitled “Where Have We Been and Where Are We Going?” suggests that “to proceed, we must understand what institutions are and how they evolve” (p. 5). In particular, North emphasizes the need to understand the processes by which new institutions come into existence, the role of learning in institutional evolution, and why inefficient institutions can persist. We show below that the social technologies conceptualization sheds light on each of these. Moreover, North (1993) also suggests the need for “integrating technological and institutional analysis” in the analysis

of economic change (p. 352), a task to which the social technologies concept, with its natural analogs to physical technologies, is well-suited to do.

In the remainder of this paper, we argue that the social technologies orientation offers four major advantages for examining the process of institutional change, based on a the history of universities' routines for management of patenting and licensing. First, it highlights the processes by which search for new social technologies (candidates for new institutions) occurs, which bears upon the "optimality" of institutions. Second, this institutions concept equates the process of institutionalization with the diffusion of social technologies. This allows for detailed examination of diffusion processes, and comparison with those for physical technologies. This is interesting not only in its own right, but also because the physical technology analogy sheds light on important unresolved issues on how institutions emerge, and in particular the role of learning in institutionalization. Third, and relatedly, the concept of institutions as standardized social technologies provides a novel explanation of why institutions may have certain economic efficiency attributes (i.e. low "transaction costs") as is often suggested in the literature. Fourth, by orienting the analysis around the social technologies themselves, this perspective allows that a wide variety of factors can induce institutional change and support particular institutions.

In Section 2, we present the historical narrative, oriented around the social technologies used by universities to manage these activities. Social technologies for management of patents and licenses include two type of routines. First, the routines for dealing with faculty members who develop patentable inventions, embodied in and supported by *de facto* or *de jure* patent policies.² Second, the routines specifying how patentable inventions are handled.³ In this narrative in Section 2, we focus primarily on the latter set of routines, though we will see that universities' patent policies themselves often shaped these routines, and thus the distinctions are blurry. In Section 3, we draw on the history to explore and elaborate themes in the evolution of social technologies, and relate to the literature on institutional change. Section 4 concludes.

2. THE HISTORY OF SOCIAL TECHNOLOGIES USED BY UNIVERSITIES TO MANAGE PATENTING AND LICENSING: A CASE STUDY OF INSTITUTIONAL CHANGE

Most educational institutions make every effort to avoid becoming directly involved in the intricate legal and commercial aspects of patent management.

Archie Palmer (1947)

Almost all research universities in the United States have technology licensing operations.

Lita Nelsen (1998)

2.1 Pre-World War II: Few Universities Have Routines for Patenting and Licensing

The “American research university” is itself a complex bundle of routines, the broad outlines of which became standardized by the first decade of the twentieth century (Geiger, 1986; Veysey, 1965). By that time, research universities had adopted a package of routines governing how research is done and research results are disseminated that would later be characterized as the “institutions of open science” (see e.g. David, 1998, p. 16). Under this mode of behavior, scientists typically are committed to disclosure and wide dissemination of their research results (Merton, 1973), and other scientists are presumed free to use these results in their own research. Accordingly, before World War I, no American research universities had a well-defined set of policies or procedures for handling patentable inventions.

The growth of industrial research in the late nineteenth and early twentieth century was a primary factor leading to the development of such routines. Since much of the early growth of U.S. industrial research occurred in “science-based” industries, such as electricity and chemicals, growth in industrial research increased industrial demand for university research, which itself had expanded considerably by the early 20th century (Geiger, 1986). Moreover, many industrial research laboratories counted among their primary responsibilities the monitoring of technological and scientific developments outside of their parent firms, which included academic research advances (Mowery, 1981). As a result, the probability that industry would be interested in university research outputs increased during the first third of the 20th century.

Related to these developments, after World War I, industry increasingly funded research at universities (Geiger, 1986; Noble, 1979). In these cases, universities that previously had paid little attention to patents were forced to confront the issue of who should receive the patent rights to inventions that resulted from sponsored research. Financial difficulties created by the Great Depression also appear to have been important in inducing universities to consider patenting and licensing. McKusick (1948) notes:

Universities had, of course, considered the possibility of patent profits before, but they had usually spurned the proposition, believing that with the bad publicity that might accompany patent exploitation, the university endowment might in the end actually suffer financially. This last fear, though still effective in the thirties, was overpowered by compelling needs

for new sources of income. University administrators turned to working out methods of patent management that would not be detrimental to their public relations (p. 213).

Consideration of routines for patenting and licensing was shaped by strong norms condemning university patenting or university involvement in patent management that were present at the time, and are suggested by the passage above. The conventional wisdom was that patenting by universities is unnecessary to fulfill the mission of the university to advance knowledge; that it could threaten “open science”; and that it would undermine the public’s trust in universities as institutions committed to advancing knowledge for its own sake. (See AAAS, 1934, for a contemporary survey.) Opposition to medical patents was particularly widespread, based on the argument that patents restricted the use of new discoveries and therefore had no place in the medical community (Weiner, 1986). Opponents of medical patenting by universities also expressed concern over public perceptions of university profiteering at public expense in the field of public health (McCusick, 1948).

As universities had little experience in these arenas, their search for new social technologies was conducted somewhat blindly. President Elliot of Purdue noted at a 1935 Conference on university patents that “[w]e are feeling our way, and no one of us knows just how to do this” (NRC, 1935, p. 49). Consequently, the results of early experiments with social technologies by a handful of prominent universities proved important to shaping the directions that others would follow.

The University of Wisconsin initiated the most important experiment in 1924. Scientist Harry Steenbock demonstrated a method of increasing the vitamin D content of food and drugs via the process of irradiation. Steenbock, despite the criticism of his colleagues at the University and many in the medical community, wished to patent his findings. In particular, he argued that in this case patenting was necessary for quality control, i.e. to prevent the unsuccessful or even harmful exploitation of the invention by unqualified individuals or firms. He believed that incompetent exploitation of the process, which might discredit the research results and possibly the university, could be avoided by patenting the process (Apple, 1996).^{4, 5}

Once the decision to acquire the patent had been made, the question of how to administer it remained. Steenbock offered to assign the patent to Wisconsin for management. However, the University was not convinced that creation of an administrative organ to handle patents was worth the necessary political and financial risk (Apple, 1996). Thus, a different social technology was developed. Steenbock convinced several alumni to create the Wisconsin Alumni Research Foundation (WARF), a university affiliated but legally separate foundation that

would accept assignment of patents from University faculty, would license these patents, and would return part of the proceeds to the inventor and the University. According to Apple (1996) the idea was that “[w]ith this structure, business matters would not concern or distract the university from its educational mandate; yet academe could reap the rewards from a well-managed patent whose royalties would pay for other scientific work” (p. 42). Other universities were anxious to see whether this new social technology worked. A contemporary article in the *Journal of the American Association of University Professors* noted that:

The experiment which is being tried out at Wisconsin is an interesting educational procedure. Its course will, doubtless, be watched with considerable interest by other educational institutions that will find themselves compelled to give more or less attention to problems of this character (Russell, 1929, pp. 272–273).

Seemingly at odds with the expressed purpose for taking out the patents, the Steenbock patents were licensed exclusively to Quaker Oats for use of the process in cereal products, and to a limited number of pharmaceutical licensees for the development of vitamin D supplements.⁶ Other well known food and pharmaceutical licensees were denied licenses to the process, causing a number of observers to question WARF’s devotion to the public good. There was some speculation that irradiated products would be considerably less expensive if manufacturers did not have to pay royalties to WARF. Undoubtedly, the Foundation’s high profits were the trigger for some of the concern: by 1936, the patents had earned substantial income for the University (Spencer, 1939, p. 27). As we discuss below, these concerns about narrow licensing and profiteering would affect the course taken by other universities over the next decades.

The results of the WARF “experiment” were thus mixed. On one hand, WARF was highly profitable, which whetted the appetites of other university administrators and trustees (Fishman, 1996). As a result, a few universities (mostly confined to the “land grant” schools) developed social technologies similar to that embodied in the “research foundation” model: creation of a loosely affiliated structure to voluntarily accept patents from university inventors and license them to firms, returning part of the profits to the universities. On the other hand, even this limited involvement in patents tarnished Wisconsin’s reputation, a fact which other universities, including the Massachusetts Institute of Technology (MIT), would note.

MIT began to consider development of routines for management of patenting and licensing only following the rise of university-industry interaction during the 1930s (Fishman 1996). In considering development of a social technology

to handle patenting, it relied heavily on feedback from experiments of other institutions. In McCusick's (1948) study of the development of MIT's patent policy and procedures, he notes that "through the years the men responsible for the patent policy of the Institute constantly exchanged experiences with other universities. Thus, the development of MIT patent policy was an outgrowth of the general university experience" (p. 206). In particular, the WARF model and WARF's experiences appear to have had considerable impact on the formulation of MIT's policies and procedures. While the Institute welcomed the possibility of generating revenues from its patents, it was wary of the scrutiny and criticism that WARF had attracted (Fishman, 1996). To insulate itself, MIT opted to work with a "third party" technology transfer agent, the Research Corporation.

The Research Corporation was founded in 1912 by Berkeley chemist, Frederick Cottrell.⁷ In the early 1900s, Cottrell developed a pollution control device, the electrostatic precipitator; did not wish to personally profit from the invention, but rather to use the royalties from it to fund scientific work. He thought of offering the patents to the University of California, but decided not to, for fears that administration of patents would interfere with the norms and routines of academe.⁸ Instead, in 1912, he formed a special organization, the Research Corporation, to administer the patents. For the first two decades of its existence, this was the principal activity of the organization. However, Cottrell had long harbored the idea that the Research Corporation might also administer patents for other academics who, like him, were interested in donating all or part of the royalties for research. In Cottrell's view, the purpose of the Research Corporation "was not merely to produce revenue for scientific research, but to act as a sort of laboratory of patent economics and to conduct experiments in patent administration" (as cited in McKusick, 1948, p. 208).

This vision was fulfilled in 1937, when MIT signed the first "invention administration agreement" with Research Corporation. Under the terms of the agreement, MIT would disclose to Research Corporation inventions that it deemed potentially patentable. Research Corporation agreed "to use its best efforts to secure patents on inventions so assigned to it and to bring these inventions into use and derive a reasonable income therefrom" and further to "use its best efforts to protect these said inventions from misuse and to take such steps against infringers as [it] may deem for the best interest of the parties hereto, but with the general policy of avoiding litigation wherever practicable." All services were provided at the expense of Research Corporation. Any license income net of expenses were to be divided according to a formula by which MIT split net royalties with Research Corporation on a 60/40 basis. Research Corporation was to use its portion of the earnings to support its grants activities.

Research Corporation's agreement with MIT buttressed the former's position as a major player in this field, and set an example to which other universities turned. It was widely viewed as being a more compelling way to dispose of patents and licenses than the WARF model, as it allowed for significantly more insularity from the business aspects of patent management, and was costless to the university. Princeton and Columbia were among the other universities that signed invention administration agreements with Research Corporation in the late 1930s.

However the cases above are exceptions. Though a handful of institutions used the WARF model or the Research Corporation model for patent management before World War II, most universities did not change their de facto routine, "don't get involved with patents," and continued to leave patenting and licensing decisions up to individual faculty members. Some went further, prohibiting patenting by faculty members, and some had policies that "they will patent nothing in medicine and leave other patentable discoveries to the disposition of the discoverer or inventor" (Spencer, 1939, p. 23). That is to say, there was no standard social technology before World War II. But activity was beginning to bubble. Spencer, writing in 1939, noted that:

During the last decade, the number of universities which have adopted well defined policies in regard to inventions made by their employees, or with the aid of university equipment, has steadily increased. Additionally, there are today many other universities actively considering the adoption of a 'Patent Policy' and it is likely that another ten years will see such a policy in every school in the country that offers scientific or engineering courses (Spencer, 1939, p. 1).

This early history already suggests some themes that are important to the evolution of social technologies used by universities to manage their patenting and licensing. First, they are responsive to costs and benefits. Thus search behavior began in response to "exceptional circumstances" and the additional need for income because of the Depression. These costs and benefits are shaped by rules of the game, in this case norms about the appropriateness of university involvement in patenting, and changes in relationships with industry. Second, the decision to adopt new social technology (in whole or part) was based on collective learning, and in particular on feedback from the successful and unsuccessful experiments of others. Each of these dynamics would continue in the postwar era.

2.2 1950–1969: The Research Corporation Model; Scattered Internal Patent Management

U.S. involvement in World War II and the Cold War transformed the structure of the U.S. national innovation system (Mowery & Rosenberg, 1998). Nowhere

was this transformation more dramatic than in U.S. universities. Academic research experienced a surge of federal funding and the prewar weakness in academic science that typified the vast majority of U.S. universities by comparison with their European counterparts gradually was replaced by excellence in a broad array of scientific fields. The expansion in federal support for academic research during and after World War II intensified university involvement in patent issues.

Increased federal funding of university research strengthened the motives for university involvement in patenting, just as the growth in university-industry links had done during the 1920s and 1930s. The expanded scale of the academic research enterprise increased the probability of a patentable invention. In addition, many federal research sponsors required the development of a formal patent policy. Summarizing the survey of U.S. universities that he conducted during the late 1950s, Palmer (1962) found that 85 universities had adopted or revised explicit patent policies during the 1940–1955 period; more than half of these new or revised policies were announced during 1946–1955. McKusick (1948) suggests that by the late 1940s, virtually all major U.S. universities had developed patent policies, an estimate that contrasts sharply with the pre-1940 situation. Palmer (1947) noted, however, that the routines for patent management were not standardized at this time:

There is no common pattern of policy statement, administrative procedure, recognition of the inventor determination of equities, assignment requirement, patent management plan, distribution of proceeds, and protection of the public interest. Nor is there any convenient grouping according to the size of an institution, complexity of university organization, or kinds of research undertaken (p. 648).

Over the decades that followed, there was some convergence across one dimension: the use of Research Corporation to evaluate inventions and to manage licensure of patents. In 1946, the Research Corporation established a Patent Management Division. The idea was to extend the “third party technology transfer” agreement that it had with MIT, Columbia, and Princeton to other universities. Under the terms of such agreements with Research Corporation, university faculty submitted invention disclosures to the Patent Management Division, which bore all costs of invention evaluation, patent prosecution, and licensing. Firms wishing to license inventions would negotiate directly with Research Corporation. The invention administration agreements were fairly standard across institutions, save for slight differences in the “sharing rules” between Research Corporation, the university, and the inventor.

In addition to the “insulation” rationale discussed above, there was a “division of labor” advantage for using the Research Corporation. Universities generally were not keen to get involved in the intricacies of patent management,

reflecting the costs and complexity of this activity. Significant fixed costs are associated with establishing an office, creating an administrative structure, developing a record keeping system, staffing up, developing academic and industrial contacts, and the like. These fixed costs, combined with the fact that most institutions could only expect a handful of patents annually, meant that it was generally inefficient to operate a patent management organization on a small scale (Palmer, 1948). Thus, there were both political and practical reasons that universities chose not to get involved with these activities.

A number of factors made this social technology increasingly attractive to universities. First and perhaps most importantly, it appeared to work satisfactorily for the early adopters. Second, the fact that a number of important universities had already adopted this model meant that other relevant actors had developed their own routines to interact with Research Corporation. Thus, various government agencies became used to working with Research Corporation to establish their rights in particular inventions, and potential licensee firms (particularly in the electronics and pharmaceutical industries) became accustomed to dealing with Research Corporation as well. Third, universities could learn from one another (and from Research Corporation) on how to use this routine, e.g. on how to streamline the disclosure submission process. Each of these fueled the diffusion of the Research Corporation model.

As a result, while there remained considerable variance in actual patent policies (Palmer, 1962) – e.g. whether a policy existed, whether disclosure was voluntary or mandatory, whether there were restrictions on patenting medicine – universities increasingly used Research Corporation for patent management. Figure 1, which shows the number and proportion of “research universities” (as classified by the Carnegie Commission in its 1973 report)⁹ that had Invention Administration Agreements with Research Corporation over the 1950–1980 period, shows the rate and extent of the diffusion of this model over the post-war era. Note that by the 1970, fully 70% of Carnegie research universities had adopted the Research Corporation model.

Undoubtedly, this is not because using Research Corporation was the “optimal” social technology, in any sophisticated sense of the term. Indeed, as Mowery and Sampat (2001) report, there were substantial inefficiencies with this model. One source of these inefficiencies resulted from agency problems. The Research Corporation was primarily interested in maximizing licensing income, while client universities were in cases interested in nurturing long-term relationships with industry. A prominent illustration of these tensions is the dispute between MIT and Research Corporation over licensure of Jay Forrester’s magnetic core memory patents, which led in 1963 to the cancellation of the

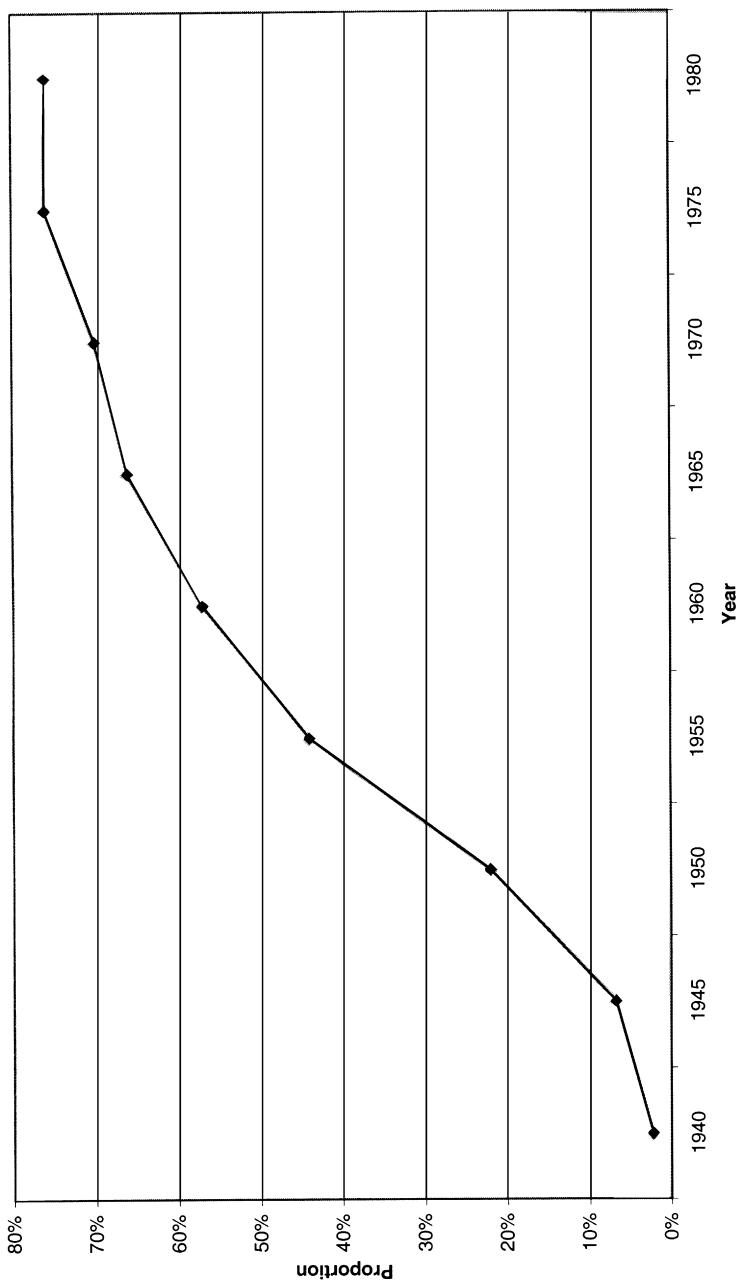


Fig. 1. Proportion of Carnegie Research Universities with IAA's with Research Corporation: 1940–1980.

MIT's invention administration agreement with Research Corporation. (See Mowery & Sampat, 2001, for full description.)

In addition, by the late-1960s, it became clear that Research Corporation could not effectively manage the entire technology transfer process from screening to licensing, for each of the more than 200 institutions with which it had invention administration agreements (Mowery & Sampat, 2001). In particular, faculty involvement was essential in invention screening, marketing, and commercialization, but this level of interaction was difficult under Research Corporation's "central broker" model, leaving many client universities unsatisfied with the Corporation's success in commercializing their inventions. In response, a number of client universities began to bypass Research Corporation, managing patents by themselves. This became increasingly so in the early 1970s (Mowery & Sampat, 2001).

In addition to these inefficiencies with Research Corporation, changes in federal patent policy also induced universities to consider bringing parts of patent management operation in-house. As a growing number of U.S. universities began to seek patents for faculty inventions and to manage their patenting and licensing activities themselves, federal agencies began to consider requests from universities for title to specific inventions on a case-by-case basis (Mowery et al., 2001; Mowery & Sampat, 2001). In response to university complaints over the cumbersome nature of this process, however, federal agencies involved in the support of academic R&D began to negotiate Institutional Patent Agreements (IPAs) with universities in the late 1960s. IPAs eliminated the need for case-by-case reviews of the disposition of individual academic inventions and facilitated licensing of such inventions on an exclusive or nonexclusive basis. The justification for these IPAs was based on logic similar to that earlier voiced by Cottrell and that would later become the justification for the passage of the Bayh-Dole act: that absent some patent protection and the possibility of exclusive licensing, firms would lack incentives to commercialize the fruits of university research. Though this argument was based on little evidence, implicit in it was the endorsement of the position that university patenting was in the public interest. It thus represented the beginning of a weakening of the long-standing norms against university patenting.

The evolution of social technologies employed over the 1950–1969 period illustrates several general themes. As in the prewar period, "search routines" were first invoked in response to changes in complementary routines, this time the increased reliance of universities on federal government funds. Routines employed were responsive to costs and benefits. Research Corporation was a low cost option for most universities because it insulated them from the political dangers of involvement in patenting, and freed them from the necessity

of developing in-house routines in an arena where they had little experience. Again, social learning was important. The Research Corporation model was increasingly adopted because it appeared to “work” (at least satisfactorily) for the early universities that employed it. However, MIT’s “divorce” from Research Corporation also illustrates that in cases where costs of a social technology are too high, new routines are sought out. Moreover, gradual learning about the inefficiencies of the Research Corporation model led to increased in-house management by its clients. The latter change was also facilitated by changes in “rules of the game,” and in particular changes in norms towards university patenting.

2.3 1970–1980: The Early Diffusion of the “Technology Transfer Office” Model

In the late 1960s, a new social technology, the “technology transfer office” model, was “invented” by Neils Reimers at Stanford University. Under this model, invention screening, the patent application process, and licensure was managed by the university, rather than an outside agent. Moreover, the model differed from the routines of the few universities that had already begun handling patents in-house (like Wisconsin, other land grants with research foundations, and MIT) insofar as it was run by businessmen with experience in marketing inventions and negotiating deals, rather than attorneys and academic administrators.¹⁰ That is, under the new social technology, inventions were aggressively marketed (a variation on the Research Corporation model) via in-house personnel (a variation on the WARF model) who specialized in technology transfer.

The inefficiency of the Research Corporation model provided the immediate catalyst for the development of the “new model” (Stanford, 2000). In a recent oral history, Reimers (1998) recalls that shortly after he was hired as associate director of Stanford’s Sponsored Projects Office in 1968:

I looked up the income we had from Research Corporation from ’54 to ’67 and it was something like \$4,500. I thought we could do a lot better licensing directly, so I proposed a technology licensing program.

The program began in 1968 on a pilot basis. When in its first year it returned \$55,000, the experiment was judged a success, and Stanford formally developed what is commonly thought to be the first “technology transfer office” (Matkin, 1990; Stanford, 2000). The early successes of this social technology attracted the attention of other universities.

At the same time, Research Corporation, believing that it may have entered into a region of “diseconomies of scope” in patent management (Mowery & Sampat, 2001) began persuading its university partners to develop their own expertise in the screening and evaluation parts of the operation. That is, it planned to specialize in patent prosecution and licensing, leaving the early stages of the technology transfer process to the universities themselves. In an attempt to assist them, it began an outreach program aimed at teaching universities about managing the early stages of academic technology transfer. Paradoxically, these missionary activities appear to have helped universities to develop full-scale technology licensing offices like Stanford, eventually abandoning reliance on Research Corporation altogether.

Learning about the new social technology was, again, a collective process. By the mid-1970s there were a number of conferences on university-industry technology transfer and setting up technology transfer offices, where representatives of universities with experience in technology transfer, like Stanford and Wisconsin, shared collective experiences among themselves and with others who had little experience with patenting. The Society of University Patent Administrators (SUPA) was founded in 1974 as a vehicle to exchange best practice amongst institutions.

Other universities began to mimic the “technology transfer office” in the 1970s, and interest in the new social technology was further fueled by the rise of a new field of science where academic research outputs were of immediate commercial interest, biotechnology. The fact that a key early discovery in this field, the Cohen-Boyer technique for genetic engineering, was patented and licensed by Stanford and the University of California and seemed very likely to generate significant income, was important in focusing collective university attention on the possibilities of generating revenues via patenting and licensing, and in the technology transfer office model in particular. (Of course, the WARF Vitamin D patents had a similar effect earlier in the century.) At the same time, the norms militating against patenting in biomedicine were weakening (Weiner, 1986), which reduced the costs of university patenting in this increasingly lucrative area.

A number of universities were responsive to these changes in costs, benefits, and opportunities. As early as 1974, Research Corporation noted in its Annual Report that every major institution was considering setting up a technology transfer office. Before 1980, 20 such technology transfer offices existed, and many other universities were considering setting one up (Mowery et al., 2001).

In considering social technologies for patent management, universities once again looked to one another's experiences. A 1976 letter from Columbia

University's Provost to its President suggests "Before Columbia steps into this quagmire . . . we will inform ourselves of the practices of our sister institutions." In 1979, Harvard surveyed the practices of research universities, and reported that while "history and precedent have frozen most sizable universities into using inefficient structures," Stanford's new model is "a relatively clear-headed approach."¹¹

Thus, in the 1970s, the "invention" of the new social technology was responsive to changes in costs and benefits, in particular the revealed inefficiencies of the Research Corporation. Its diffusion was fueled by some evidence of its success, as well as the rise of new fields of science. The diffusion process was magnified and accelerated by a change in formal rules of the game, as discussed below.

2.4 1981–2000: The Standardization of the "Technology Transfer Office" Model

In the late 1970s, there remained some uncertainty about whether this active university involvement in patenting and licensing was appropriate, given the long-standing norms that militated against it. Moreover, uncertainty remained about federal policy toward academic patenting and licensing of federally funded research advances. Most federal agencies continued to require cumbersome case-by-case petitioning for approval, and some appeared to be reconsidering whether to allow universities to issue exclusive licenses.

The Bayh-Dole act, passed in 1980, resolved both of these uncertainties. The Act provided blanket permission for performers of federally funded research to file for patents on the results of such research and to grant licenses for these patents, including exclusive licenses, to other parties. These provisions facilitated university patenting and licensing in at least two ways. First, it replaced the web of IPAs that had been negotiated between individual universities and federal agencies with a uniform policy. Second, and more importantly, the Act's provisions expressed Congressional support for the negotiation of exclusive licenses between universities and industrial firms for the results of federally funded research. Important to the passage of Bayh-Dole was "evidence" that absent patent protection, technology transfer would not occur. (This evidence was later shown to lack merit, as discussed in Eisenberg, 1996). In the same year that Bayh-Dole was passed, the U.S. Supreme Court ruled in *Diamond v. Chakrabarty* that genetically modified organisms are patentable, concurrently reducing uncertainties about the legality of patenting in biotechnology, the field where university patenting was expected to be most profitable. Together, the legislative and judicial actions magnified and accelerated university development

of intellectual property policies which asserted rights to faculty inventions, and technology transfer offices which evaluated, patented, and marketed these inventions. In his history of the American research university after World War II, Geiger (1986) notes that:

University behavior changed most broadly during the 1980s in the area of patenting. Almost every research university re-evaluated its policies . . . and established procedures to enhance the potential flow of patents . . . One after another, universities became decidedly more aggressive in the three phases of the patenting process – encouraging faculty to disclose inventions, filing for patents, and marketing the licenses to those patents (p. 317).

During the 1980s, the rate of diffusion of the “technology transfer office” model increased dramatically, so that by the mid-1990s nearly 200 universities, and all major research universities, had adopted the model (AUTM, 1996; Nelsen, 1998).

Unfortunately, annual data on adoption is not available, but several sources of data are informative. Figure 2 shows the proportion of Carnegie research universities that had assigned at least 0.5 FTE (full time equivalent) employees to “technology transfer” activities, over the 1965–1995 period. Note that the trend towards involvement in technology transfer begins in the 1970s, but takes off after Bayh-Dole. This is also seen in Fig. 3, a plot of the proportion of Carnegie research universities assigned at least one patent (since 1965), over the 1965–1995 period.

In addition to diffusion and standardization of administrative, patent prosecution, and marketing routines, patent policies became standardized as well. Thus nearly all research universities today assert rights to faculty research outputs developed using university resources, and none have restrictions on biomedical inventions. Nearly all offer some share of royalty revenues to the faculty inventor and his department, though the precise sharing rules vary. The results of the diffusion of these policies and the adoption of the “technology transfer office” model are reflected in Fig. 4, which plots the number of patents issued to Carnegie Universities over the 1965–1995 period.

Here again, the new social technology spread and was made more efficient over time via a cumulative social learning process. In the early 1980s, several groups organized conferences to share expertise and draft model documents; these proved important vehicles through which universities converged on a standard and learned from one another’s experiences (Bowie, 1994). Neils Reimers was hired as a consultant at several universities to assist them in setting up technology transfer offices. This social learning process continues to this day. The publications and conferences of AUTM, the Association of University Technology Managers, a society which assists members in the management and

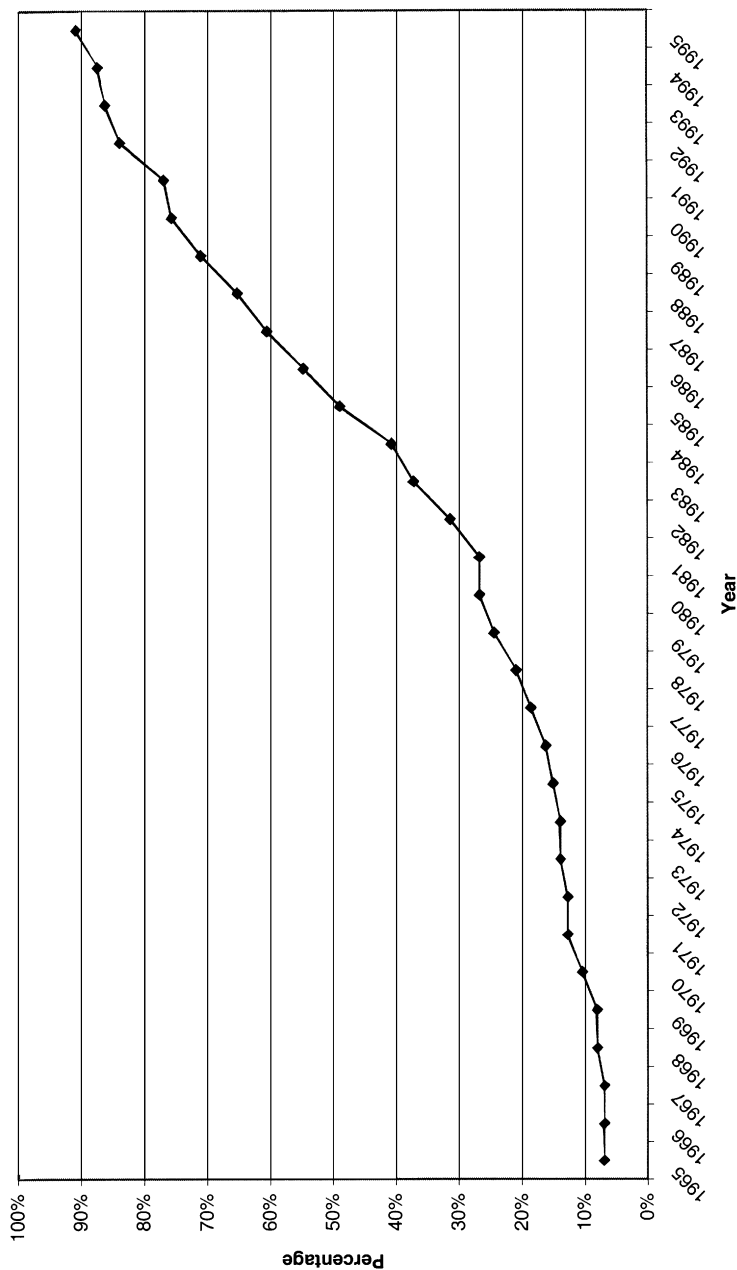


Fig. 2. Proportion of Carnegie RUIs with > 0.5 FTE "Technology Transfer" Personnel, 1965–1995.

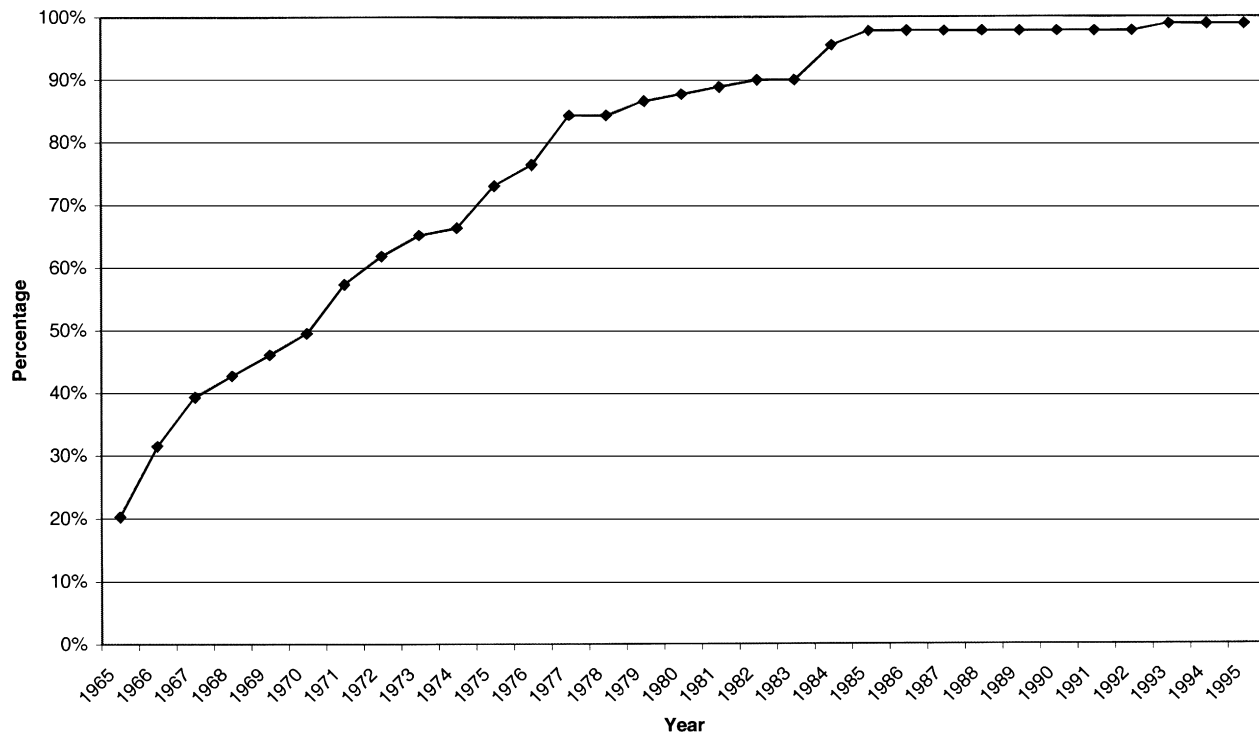


Fig. 3. Proportion of Carnegie Research Universities with at least One Patent Since 1965, 1965–1995.

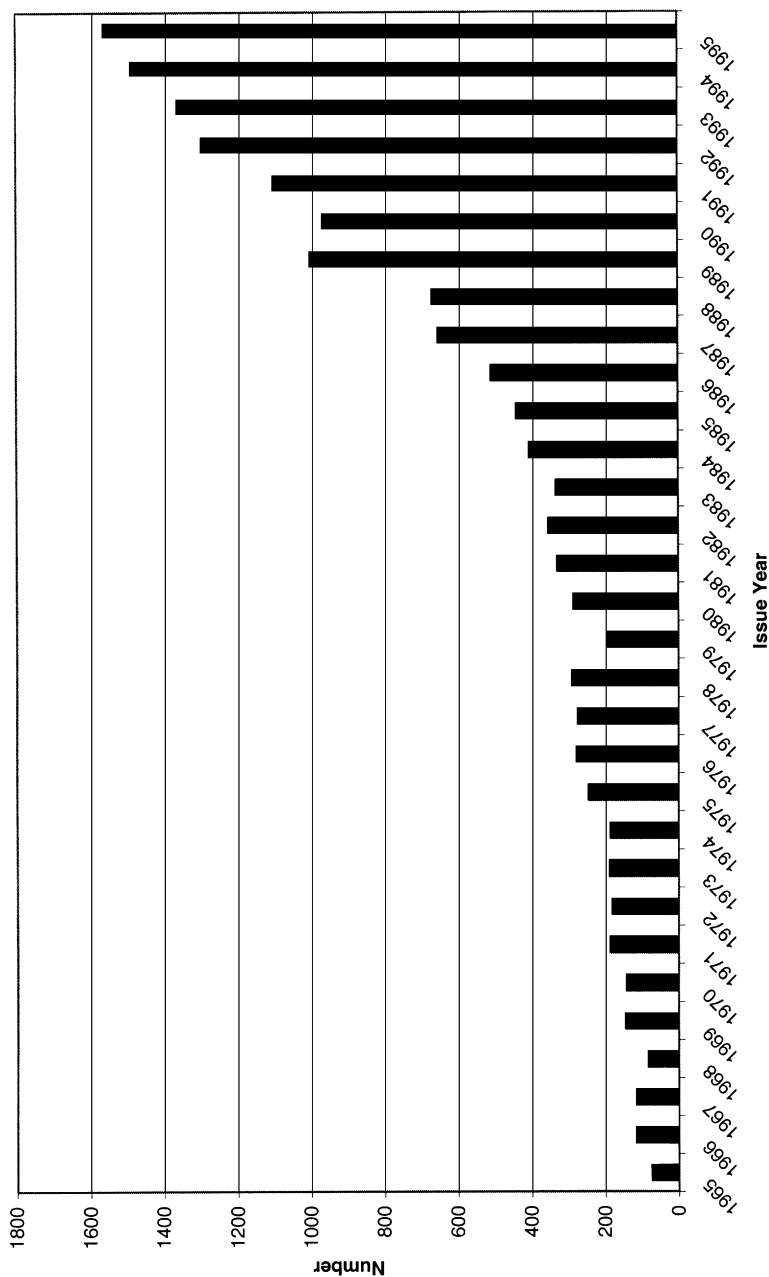


Fig. 4. Number of Patents Issued to Carnegie Research Universities by Year, 1965–1995.

promotion of academic technology, are particularly important vehicles.¹² There also exists an active internet mailing list, *techno-l*, where university technology transfer officers exchange experiences about particular licensees and how to manage particular types of deals. The dimensions of this social learning can be seen by perusing the active internet mailing list for technology transfer officers, *techno-l*, or the agenda from recent meetings of the Association of University Technology Managers (AUTM).¹³ Importantly, such interchange of experience is facilitated by the fact that the basic model and policies used by the universities are standard.

By the mid-1990s, the technology transfer office was “institutionalized,” in the sense that all relevant actors used basically the same model. This is not to say that the routines followed by these universities are identical (Berkowitz et al., 2000). Despite some differences, most observers would agree that the routines followed by universities are of the same family: the technology transfer office is an institution. Indeed, universities without such institutions today would have trouble getting grants from certain government agencies, and perhaps even recruiting faculty members in some departments.

Though nearly all universities now employ this model, this is not to say that the new institution is necessarily an optimal response to extant costs and benefits. A recent study estimates that fewer than 50% of universities generate enough licensing income to cover their costs (Trune & Goslin, 1997). Even if the utility functions of universities include parameters other than profits, as they undoubtedly do, there is little evidence that the technology transfer office model is serving these ends (Mowery et al., 2001). There is no strong evidence that the new model facilitates “technology transfer.” This is in fact difficult to assess, given a number of intervening factors (such as the rise of the biotechnology and computer industries) confound any causal statements about the presence of the technology transfer offices and industry utilization of university technologies.

The uncertainty about the effectiveness of the new institution is particularly interesting because it is supported by a body of law (Bayh-Dole) that was passed under the (largely unfounded) assumption that lack of active university involvement in technology transfer and lack of patent protection on university technologies were inhibiting technology transfer from universities to industry. There has been little empirical analysis of the social benefits and costs of the new social technology. Typically, university administrators and groups like AUTM point to growing level of patents, licenses, and revenues as evidence that the new regime is working, with the implicit (and perhaps heroic) assumption that these metrics are somehow related to gains in social welfare. Interest groups have formed – like AUTM, university administrators

and technology transfer officials, and pharmaceutical companies – that have a vested interest in its survival. A perusal of the *techno-l* mailing list or leafing through an AUTM publication suggests that a norm has developed among these parties that this is the “right” (not only “expected”) way things are done. That is, social technologies can take on lives of their own.

Despite a lack of clear evidence, policymakers and university administrators appear to believe that the new institution works (see e.g. United States Government Accounting Office, 1998; AUTM, 1999). Even smaller universities, who likely do not have the research base to make it profitable – are increasingly employing this social technology. In addition, foreign governments (e.g. in Canada and Japan) are looking to the new American “system” of university-industry technology transfer as an example. Thus this social technology may well diffuse further.

3. SOCIAL TECHNOLOGIES, INSTITUTIONS, AND INSTITUTIONAL CHANGE

Empirical work on institutions, institutionalization, and institutional change is still in its early, and pre-paradigmatic stages (Alston et al., 1997). In this section, we discuss several advantages of viewing this episode of institutional change through the lens of “social technologies,” and relate to broader themes in the economic literature on institutions. From this vantage point, a prerequisite for the creation of a new institution is the development of a new social technology. One advantage of the social technologies formulation is that it unpacks the “institutionalization” process into the emergence of a new social technology and its spread, allowing for fine-grained consideration of the sources of new institutions and their efficiency attributes.

First, let us consider origins of new social technologies. A widely held assumption in the literature is that the search for new institutions is responsive to costs and benefits, e.g. changes in relative factor prices (North, 1981; Williamson, 1975). In the case discussed above, we saw several such innovations, including the WARF model, the Research Corporation model, and the technology transfer office model. In each of these cases, the creation of the “new” models was indeed responsive to costs and benefits. Each of the first two were created in response to the need for some vehicles to manage patents on particularly profitable inventions, the need for new sources of revenues, as well as in response to changes in routines to manage interaction with industry more generally. The third was created in response to inefficiencies in the Research Corporation model, though the growth of federal support and

potentially patentable research, as well as changes in routines for interaction with federal research sponsors, also played a role.

In this case search was indeed responsive to changes in costs and benefits. But what is more interesting is that one source of these changes were changes in higher order complementary routines, i.e. routines for interaction with industry sponsors, routines for interaction with federal sponsors. In this case, as in others, particular social technologies are typically nested within a broader constellation of routines, and changes in complementary routines can induce search behavior. There are strong similarities in the literature on physical technologies. In particular, Rosenberg (1969) has noted that when physical technologies are parts of technological systems, that changes in complementary technologies can serve as important “inducements” to search, or focus exploratory activity in particular directions.

Another characteristic of the search process for new social technologies was that it drew significantly on the experiences of others. Thus in its decision to rely on Research Corporation, MIT drew upon WARF’s experiences and the WARF model. Similarly, Reimers’ technology transfer office model is clearly a variant of the Research Corporation model. Part of this reliance on others’ experiences is that for social technologies, more so than physical technologies, “offline” testing is typically not feasible: the main way to establish if something works is to try it, or to get feedback from others’ experiences with similar (or the same) social technologies. One result of this is that search for new social technologies remains local. This topographical characteristic of the search space is also present in several discussions of search for physical technologies, including Nelson and Winter (1982), who emphasize searching “in the neighborhood” of current physical technologies. This is not to say that social technologies (or physical technologies) cannot be “new” – that is obviously false – but rather that what has already been seen exerts an important influence on the search domain. This presents some inertia in the evolution of social technologies, which limits the degree to which any observed set of institutions can be considered “optimal,” at least in the naive sense of the term.

The social technologies concept thus has the advantage of focusing attention on the origins of new modes of behavior, ultimately the sources of new institutions. It also invites examination of the process by which social technologies spread and become institutions (what the sociologists call “institutional isomorphism,” see Powell & DiMaggio, 1991) which is analogous in many ways to the “diffusion” of physical technologies. This is important because it illuminates the sources of prevailing institutions and their efficiency attributes, and sheds light on several questions that are rarely posed in mainstream institutional analysis.

In this case history, the two social technologies that diffused to the point where they could be considered standardized were the Research Corporation model, which spread in the 1950s, 1960s, and 1970s, and the technology transfer office model, which spread over the late 1970s, 1980s, and 1990s. Clearly, perceived costs and benefits were important to universities' decisions to adopt these social technologies, as is typically the case with the diffusion of new physical technologies (Griliches, 1957; Rosenberg, 1972). In addition, each of these social technologies were subject to various types of dynamic increasing returns, which fueled the diffusion and standardization process.

In the case of the Research Corporation model, one source of dynamic increasing returns was that as more and more universities began to use it, actors in the patenting and licensing process (university research administrators, faculty, industry members, government sponsors) grew increasingly used to dealing with the latter (indeed expected to do so), and developed their own routines to mesh with universities' practices. This obviously made the model that much more attractive to the next round of potential adopters. In addition, there were some efficiency gains over time as universities learned from one another how to motivate faculty members to turn over inventions to Research Corporation, how to screen inventions for patentability, and how to decide what types of inventions to turn over to Research Corporation. Similarly, over the 1980s other actors' expectations and routines began to mesh with the technology transfer office model, and there was considerable sharing of experiences amongst universities that increased the efficiency of the model. Each of these acted as positive feedback mechanisms that made the social technologies increasingly attractive over time.

There are several similarities in the dynamics of diffusion and standardization of physical technologies. First, it is well known that when physical technologies are embedded in larger systems of physical technologies and complementary routines, once a sufficient number of actors (or a sufficiently important set of actors) have adopted a particular technology, there is a tendency for convergence towards a *de facto* standard. Second, as with social technologies, "learning by using" is an important source of improvements in physical technologies (Rosenberg, 1982). Moreover, in many contexts, channels exist for shared social learning about particular physical technologies, creating another sort of positive feedback mechanism that can lead to the emergence of standards.

Thus, the social technologies concept naturally focuses attention on the process by which idiosyncratic practices can become institutions, and the importance of various positive feedback mechanisms in that process. Relatedly, a third advantage of the viewing institutions as social technologies is that it

provides a novel insight on the efficiency of institutions. The transaction cost economics strand of the new institutional economics suggests that institutions are chosen to reduce transaction costs. The social technologies orientation, with its emphasis on techniques and learning and its analogies to physical technologies, provides a new twist to this story: particular patterns of behavior may be low transaction cost *because* they are institutions, i.e. standardized. A first appealing characteristic of using standard modes of behavior is that they mesh with the expectations and routines of other actors. Part of this argument was made earlier by Kindleberger (1983), who suggested that standards themselves can lower transaction costs by facilitating expected modes of behavior and interaction, i.e. by reducing uncertainty. Another benefit of standards is that they have been honed via a process of cumulative social learning, and thus actors need not rely only upon own “learning by using” for productivity improvements.

Of course, this is not to say that institutions are optimal. From the literature on physical technologies, we know that the same factors that make standards attractive – various forms of network effects – can slow down movement to improved technologies (Stoneman, 1987). Moreover, in the context of network effects, a relatively inefficient technological alternative can become “locked in” if chosen early on by chance historical events. The classic example of this “path dependent” evolution is the QWERTY keyboard, discussed by David (1986). The same dynamics can hold for institutions, as discussed by North (1991), and can limit the efficiency of institutions.

Above, we have made the point that the search for social technologies is local, limited to variations on what has been tried or seen in practice, suggesting that there can always be better options “out there” that are yet undiscovered. Less trivially, limited feedback on the fitness of social technologies also limits the efficiency of institutions. As this case illustrates, it is difficult to determine whether a social technology works: for example, the various inefficiencies of the Research Corporation model were only learned over time. In other words, incorrect beliefs that a social technology works can persist in the absence of sharp feedback. This is also true of some physical technologies (Constant, 1999). However, as Nelson (2000) suggests, the problem may be more pronounced for social technologies, insofar as they typically are less amenable to offline testing (i.e. they can only be tested by being used), they are less modular (i.e. it is difficult to isolate and test the performance of particular parts of social technologies), and relatedly, their workings typically are not illuminated by strong bodies of science. Of course, this is not uniformly true: as with physical technologies (Nelson, 2000), some “types” of social technologies may be more likely to generate clear performance signals than others.

Finally, the social technologies concept is advantageous because, by focusing on behavior patterns themselves, it is flexible as to the sources of standardization. In the discussion above, changes in rules of the game, norms, and relative efficiencies of various arrangements were all important to institutional evolution and standardization. The social technologies concept admits all of these, and hence is especially suitable for organizing long term historical analyses of institutional change. Thus, in this case study, examination of only changes in rules of the game, norms, scientific opportunities or transaction costs would have provided an incomplete account of the institutional changes that occurred over the twentieth century.

4. CONCLUSIONS

Nelson and Sampat (2001) proposed that it is fruitful to think about institutions as standardized social technologies. This paper suggests that the social technologies framework is useful for the analysis of institutional change. In particular, because it is based in a routine-theoretic setup that focuses attention on how particular activities get done, the social technologies framework highlights the dynamic processes by which particular routines become standardized, a dimension that is neglected in the mainstream economics literature on institutions.

But of course, this is a single case study, and real proof of the utility of the social technologies framework will come from future empirical studies of institutional change. Moreover, just as the dynamics of the evolution of physical technologies varies across types of technologies and industries, this is also likely to be true for social technologies. For example, there may be important differences between the level of, and the channels for, social learning involved in the standardization of the M-form and the standardization of technology transfer offices. Laws and public policies (social technologies employed by government actors), may evolve and become standardized through mechanisms different from those employed by decentralized private actors. Examining differences in the evolutionary dynamics of different types of social technologies remains an important task for future research.

It is widely recognized that economic growth is a result of the co-evolution of technology and institutions (Rosenberg & Birdzell, 1986; Nelson, 1994). This process is difficult to describe, let alone model. It is hoped, ultimately, that the institutions as social technologies conceptualization can help orient

appreciative and formal models of this process.

NOTES

1. Upon some reflection, it is clear that most economic actors typically (routinely!) employ both physical and social technologies. Thus the physical technology for making a peanut butter and jelly sandwich includes steps like: get two slices of bread, arbitrarily index each slice by A and B, arbitrarily index the sides of each slice by 1 and 2, spread peanut butter on side 1 of slice A, spread jelly on side 1 of slice B, and put together slices A and B with sides 1 facing inwards. Of course, “producing” such a sandwich also includes social technologies for acquiring the necessary inputs (e.g. go to the supermarket, driving on the right side of the road, wait in a line, exchange certain types of paper for the ingredients) and for coordinating labor (e.g. dad agrees to manage the production process if mom agrees to wash the dishes afterwards).

2. For example, as we describe below, early in the twentieth century, one routine commonly employed was “if a faculty member develops a patentable invention, do nothing” whereas by the end of the century a routine including the decision rule “if a faculty member develops a patentable invention and does not assign it to the university, admonish or punish her” became common.

3. Examples include “place the patent in the public domain, “assign the invention to a patent licensing agent,” or “market the invention to potential licensees.”

4. Apple (1989) also notes that another more or less unstated reason that Steenbock wanted to patent the invention was to prevent margarine producers from acquiring the process, thus protecting the region’s dairy interests.

5. Blumenthal et al. (1996) conjecture that Steenbock was also motivated by the experiences of his colleague Dr. Stephan Babcock. Babcock had developed a method for determining the butterfat content of milk, and, as was then standard practice, did not file for patent protection on the invention. However, lack of patent protection limited the degree to which Babcock could prevent low-quality producers from flooding the market with “Babcock testers,” eventually discrediting the method altogether. See also Apple (1996).

6. In contemporary legal jargon, the invention was co-exclusively licensed for the pharmaceutical field of use.

7. This discussion of the Research Corporation draws largely from Mowery and Sampat (2001).

8. Specifically, he noted that “a danger was involved, especially should the experiment prove highly profitable to the university and lead to a general emulation of the plan. University trustees are continually seeking for funds and in direct proportion to the success of our experiment its repetition might be expected elsewhere . . . the danger thus suggested was the possibility of growing commercialism and competition between institutions and an accompanying tendency for secrecy in scientific work (Cottrell, 1932, as quoted in Mowery & Sampat, 2001).

9. In its 1973 report, the Carnegie Commission on Higher Education classified the nation’s 173 doctorate granting institutions as Research Universities and Doctoral Universities. Institutions that awarded at least 50 doctorates in 1969–1970 and were among the 50 leading recipients of federal financial support in at least two of the three years 1968–1969, 1969–1970, 1970–1971 were classified as “Research University I”

(RU1). Institutions that awarded at least 50 doctorates in 1969–1970 and ranked in between 50th and 100th in federal financial support in two of the three years were classified as “Research University 2” (RU2). We treat the union of the RU1s and RU2s as “Carnegie Research Universities.”

10. Matkin (1990) suggests that at the handful of universities that handled patents in house before the late-1960s (like MIT and the University of California), “[t]hese activities were generally carried out at a low level and were concerned with the legal aspects of patenting rather than with marketing” (p. 66). He later notes that Stanford’s program was “the first, and for a long time the only university patent administration office that had marketing as its primary purpose” (p. 71).

11. Both documents were found in the “Patents” folder of the Columbiana Library, Columbia University. They are available from the authors upon request.

12. AUTM grew out of the aforementioned Society of University Patent Administrators.

13. Evidence of the importance of social learning is provided by Mowery, Sampat, and Ziedonis (2002), who show that the quality of “entrant” university patents (as measured by the number of times they are cited in later patents) converged with those of “incumbent” universities over the 1981–1992 period, and this learning process cannot be explained by university specific learning curves alone.

ACKNOWLEDGMENTS

Thrainn Eggertsson provided useful comments on a previous draft. This research would not have been possible without the assistance and enthusiasm of archivists at the National Research Council, the Smithsonian Institution, the National Archives, Research Corporation, the Massachusetts Institute of Technology, Harvard University, and Columbia University. We gratefully acknowledge the generous financial support of the Andrew Mellon Foundation and Columbia University’s Office of the Executive Vice Provost.

REFERENCES

- (AUTM) Association of University Technology Managers (1996). *Licensing Survey*. Norwalk, CT.
- Alston, L. J., Eggertsson P., & North, D. C. (1996). *Empirical Studies in Institutional Change*. Cambridge; New York: Cambridge University Press.
- American Association for the Advancement of Science (1934). The Protection By Patents of Scientific Discoveries. *Science*, 79.
- Apple, R. (1996). *Vitamania: Vitamins in American Culture*. New Brunswick: Rutgers University Press.
- Berckowitz, J., Feldman, M., & Feller, I. (2000). Equity and the Technology Transfer Strategies of Universities. Draft Manuscript.
- Blumenthal, D., & Epstein, S. (1986). Commercializing University Research: Lessons From the History of the Wisconsin Alumni Research Foundation. *New England Journal of Medicine*, 314, 1621–1626.

- Carnegie Commission on Higher Education (1973). *A Classification of Institutions of Higher Education: A Technical Report*. Berkeley, California.
- Commons, J. R. (1924). *Legal Foundations of Capitalism*. New York: Macmillan.
- Constant, E. (1999). Reliable Knowledge and Unreliable Stuff. *Technology and Culture*, 40, 324–357.
- Corporation Research (1974–1975). *Annual Report*.
- Cottrell, F. (1932). Patent Experience of the Research Corporation. *Transactions of the American Institute of Chemical Engineers*.
- Dasgupta, P., & David, P. (1994). Towards a New Economics of Science. *Research Policy*, 23, 487–521.
- David, P. (1998). Common Agency Contracting and the Emergence of ‘Open Science’ Institutions. *American Economic Review*, 88, 15–21.
- David, P. A. (1986). Understanding the Economics of QWERTY: The Necessity of History. In: W. N. Parker (Ed.), *Economic History and the Modern Economist* (pp. 30–49). Oxford and New York: Blackwell.
- Eisenberg, R. (1996). Public research and private development: patents and technology transfer in government-sponsored research. *Virginia Law Review*, 83, 1663–1727.
- Fishman, E. (1996). *MIT Patent Policy 1932–1946: Historical Precedents in University-Industry Technology Transfer*. Philadelphia: University of Pennsylvania.
- Geiger, R. L. (1986). *To Advance Knowledge: The Growth of American Research Universities, 1900–1940*. New York: Oxford University Press.
- Geiger, R. L. (1993). *Research and Relevant Knowledge: American Research Universities Since World War II*. New York: Oxford University Press.
- Griliches, Z. (1957). Hybrid Corn: An Exploration in the Economics of Technical Change. *Econometrica*, 48, 501–522.
- Kindleberger, C. P. (1983). Standards as Public, Collective and Private Goods. *Kyklos*, 36, 377–396.
- Matkin, G. W. (1990). *Technology Transfer and the University*. New York: National University Continuing Education Association: American Council on Education: Macmillan.
- McKusick, V. L. (1948). A Study of Patent Policies in Educational Institutions, Giving Specific Attention to the Massachusetts Institute of Technology. *Journal of the Franklin Institute*, 245, 193–225, 271–300.
- Merton, R. K. (1973). *The Sociology of Science: Theoretical and Empirical Investigations*. Chicago: University of Chicago Press.
- Mowery, D., Nelson, R., Sampat, B., & Ziedonis, A. (2001). The Growth of Patenting and Licensing by U.S. Universities: An Assessment of the Effects of the Bayh-Dole Act of 1980. *Research Policy*, 30, 99–119.
- Mowery, D., & Sampat, B. (2001). Patenting and Licensing University Inventions: Lessons from the History of the Research Corporation. *Industrial and Corporate Change*, 10, 317–355.
- Mowery, D., Sampat, B., & Ziedonis, A. (2002). Learning to Patent? Institutional Experience and the Quality of University Patents. *Management Science*, 48, 73–89.
- Mowery, D. C., & Rosenberg, N. (1998). *Paths of Innovation: Technological Change in 20th Century America*. Cambridge, U.K.; New York: Cambridge University Press.
- National Research Council (U.S.), and Archie MacInnes Palmer (1948). *Survey of University Patent Policies, Preliminary Report*, by Archie M. Palmer, director of survey. Washington.
- Nelsen, L. (1998). The Rise of Intellectual Property Protection in The American University. *Science*, 1460–1461.
- Nelson, R. (2000). On The Uneven Evolution of Human Know-How. Draft Manuscript.

- Nelson, R. R., & Sampat, B. N. (2001). Making Sense of Institutions as a Factor Shaping Economic Performance. *Journal of Economic Behavior and Organization*, 44, 31–54.
- Nelson, R. R., & Winter, S. G. (1982). *An Evolutionary Theory of Economic Change*. Cambridge, Mass.: Belknap Press of Harvard University Press.
- Noble, D. F. (1979). *America by Design: Science, Technology, and the Rise of Corporate Capitalism*. Oxford: Oxford University Press.
- North, D. (1990). *Institutions, Institutional Change, and Economic Performance*. Cambridge: Cambridge University Press.
- North, D. (1993). Where Have We Been And Where Are We Going? Draft Manuscript.
- North, D. C. (1981). *Structure and Change in Economic History*. New York: Norton.
- North, D. C. (1993). Toward a Theory of Institutional Change. In: W. A. Barnett, M. J. Hinich & N. J. Schofield (Eds), *Political Economy: Institutions, Competition, and Representation: Proceedings of the Seventh International Symposium in Economic Theory and Econometrics.. International Symposia in Economic Theory and Econometrics* (pp. 61–69). Cambridge; New York and Melbourne: Cambridge University Press.
- Palmer, A. (1947). Patents and University Research. *Law and Contemporary Problems*.
- Palmer, A. M. (1962). *University Research and Patent Policies, Practices, and Procedures*. Washington: National Academy of Sciences-National Research Council.
- Powell, W. W., & DiMaggio, P. (1991). *The New Institutionalism in Organizational Analysis*. Chicago: University of Chicago Press.
- Reimers, N. (1998). Stanford's Office of Technology Licensing and the Cohen/Boyer Cloning Patents. *Oral History Collection: University of California, Berkeley*.
- Rosenberg, N. (1969). The Direction of Technological Change: Inducement Mechanisms and Focusing Devices. *Economic Development and Cultural Change*, 18, 1–24.
- Rosenberg, N. (1972). Factors Affecting the Diffusion of Technology. *Explorations in Economic History*, 10, 3–33.
- Rosenberg, N. (1976). On Technological Expectations. *Economic Journal*, 86, 523–535.
- Rosenberg, N. (1976). Technological Change in the Machine Tool Industry. In: *Perspectives on Technology*. Cambridge: Cambridge University Press.
- Rosenberg, N., & Birdzell, L. E. (1986). *How the West Grew Rich: The Economic Transformation of the Industrial World*. New York: Basic Books.
- Schotter, A. (1981). *The Economic Theory of Social Institutions*. Cambridge: Cambridge University Press.
- Spencer, R. (1939). *University Patent Policies*. Chicago: Northwestern University Law School.
- Stoneman, P. (1987). *The Economic Analysis of Technology Policy*. Oxford; New York; Toronto and Melbourne: Oxford University Press.
- Trune, D., & Goslin, L. (1998). University Technology Transfer Programs: A Profit/Loss Analysis. *Technological Forecasting and Social Change*, 57.
- United States Government Accounting Office (1998). *Technology Transfer: Administration of the Bayh-Dole Act by Research Universities*. Washington, D.C.: U.S. Government Printing Office.
- Veysey, L. R. (1965). *The Emergence of the American University*. Chicago: University of Chicago Press.
- Weiner, C. (1986). Universities, Professors, and Patents: A Continuing Controversy. *Technology Review*, 1, 33–43.
- Williamson, O. (1975). *Markets and Hierarchies: Analysis and Antitrust Implications*. New York: Free Press.
- Williamson, O. E. (1985). *The Economic Institutions of Capitalism: Firms, Markets, Relational Contracting*. New York London: Free Press.

PART 2:
WHY ARE FIRMS DIFFERENT?

This Page Intentionally Left Blank

COMPETITION, CONTINGENCY, AND THE EXTERNAL STRUCTURE OF MARKETS

Ronald S. Burt, Miguel Guilarte,
Holly J. Raider and Yuki Yasuda

ABSTRACT

This paper is in three parts about the market factor in contingency theory: (1) We focus on the dual structure of markets; the internal structure of relations among producers vs. the external structure of buying and selling with other markets. We use a network model to describe the association between performance and the dual structure of American markets from 1963 to 1992. (2) We reverse-engineer the network model to infer the “effective” level of competition among producers in each market. Effective competition, a measure of competitive intensity, is inferred from observed market profits predicted by the market network of dependence on other sectors of the economy. Producers with profit margins higher than expected from observed market structure must face an “effective” level of competition lower than the level implied by the observed structure. Instead of predicting performance from internal and external market structure, we use data on performance and external structure (the more reliable and detailed data) to infer internal structure. (3) We demonstrate the research value of the effective competition variable for its reliability (illustrated by automatic adjustment for the

exogenous shock of imports in 1982), its accuracy (illustrated by revealing the contingent value of a strong corporate culture in Kotter & Heskett's, 1992, study), and as a market factor integrating case with comparative research. We close discussing the market conditions measured by effective competition, which, as an unobserved variable, is more subject than observed variables to misinterpretation.

INTRODUCTION

Consider an example piece of organization research; an example to which we return later, an example about economic performance linked to corporate culture. Corporate culture is to a corporation what it is to any other social system, a set of beliefs, myths, and practices shared by people such that they feel invested in, and part of, one another. There is a rich literature describing how the cultures of organizations differ and the concept of culture developed in organization studies (e.g. Barley, Meyer & Gash, 1988; Ott, 1989; Martin, 1992; Schein, 1996; Pfeffer, 1997, pp. 120–126; Scott, 1998, pp. 133–136, 311–313; Hirsch & Levin, 1999, p. 209), but it is sufficient for the purposes here to put aside the specific beliefs that employees share and focus on culture strength. The culture of an organization is strong when employees share beliefs, myths, and practices so as to feel invested in, and part of, one another. Culture is weak when employees hold widely different, even contradictory, beliefs so as to feel distinct from one another (where the individual, as Durkheim, 1897, p. 157, so nicely put it in his analysis of social integration created by shared beliefs, is “far more the author of his faith”).

In theory, a strong corporate culture can enhance corporate economic performance by reducing costs. One factor is lower monitoring costs. The shared beliefs, myths, and practices that define a corporate culture are an informal control mechanism that coordinates employee effort (e.g. O'Reilly, 1989; Kotter & Heskett, 1992, Chap. 2; Barker, 1993). Employees deviating from accepted practice can be detected and admonished faster and less visibly by friends than by the boss. The firm's goals and practices are more clear, which lessens employee uncertainty about the risk of taking inappropriate action so they can respond more quickly to events. New employees are more effectively brought into coordination with established employees because they are less likely to hear conflicting accounts of the firm's goals and practices. Moreover, the control of corporate culture is less imposed on employees than it is socially constructed by them, so employee motivation and morale should be higher than when control is exercised by a superior through formal lines of authority. In addition, there are labor savings. For reasons of social pressure from peers, the attraction of

pursuing a transcendental goal that makes one's job significant beyond pay, or the exclusion of employees who do not fit the corporate culture, employees work harder and for longer hours in an organization with a strong corporate culture (e.g. Kotter & Heskett, 1992, Chap. 2). The savings from lower monitoring costs and free, quality labor mean that firms with stronger corporate cultures can be expected to enjoy higher levels of economic performance. Whatever the magnitude of the economic enhancement, call it the "culture effect."

There is evidence of the culture effect, but the evidence also shows that the effect is contingent on market environment as illustrated in Fig. 1. The graphs in Fig. 1 are adapted from Burt, Gabbay, Holt and Moran's (1994) analysis of Kotter and Heskett's (1992) data on corporate culture and economic performance. We take a closer look at these data later in the paper. For the moment, allow that we have – from Kotter and Heskett for 180 firms in 19 markets – a measure of the strength of a firm's corporate culture, and a measure of the firm's economic performance based on a decade of returns to invested capital. To make comparisons across markets, subtract from each firm's score the average in its market (Burt et al., 1994, pp. 348–350). As evidence of the culture effect, economic performance has a significant positive association with culture strength across the 180 firms (0.51 correlation, 7.4 *t*-test with 18 dummy variables adjusting for market means; Burt et al., 1994, Fig. 2).

The results in Fig. 1 show how the culture effect varies between markets (cf. Burt et al., 1994, Figs 3 and 5). The graph in the bottom-right corner of Fig. 1 shows the culture effect for 36 sample firms in the four most "effectively" competitive markets in Kotter and Heskett's study (airlines, apparel, motor vehicles, textiles; "effective" defined below). The culture-performance correlation (CPr) is 0.72 with a 5.8 *t*-test, showing that firms with stronger cultures have significantly higher returns to invested capital. At the other extreme, the graph in the bottom-left corner of Fig. 1 shows a very different culture effect for the 30 sample firms in the four least effectively competitive markets (beverages, communications, personal care, pharmaceuticals). There is no association in these markets between culture strength and economic performance (0.06 CPr, 0.3 *t*-test).

Each dot in the graph at the top of Fig. 1 is a market in Kotter and Heskett's study, positioned vertically by the strength of culture effect within the market (CPr), and horizontally by the effective level of competition within the market. Causality could run either or neither way in the association between culture strength and performance (see Burt et al., 1994, 365ff.), but either way it is clear that the association increases in proportion to competitive pressure in a market. For the purposes of this paper, we assume the causal order of the culture

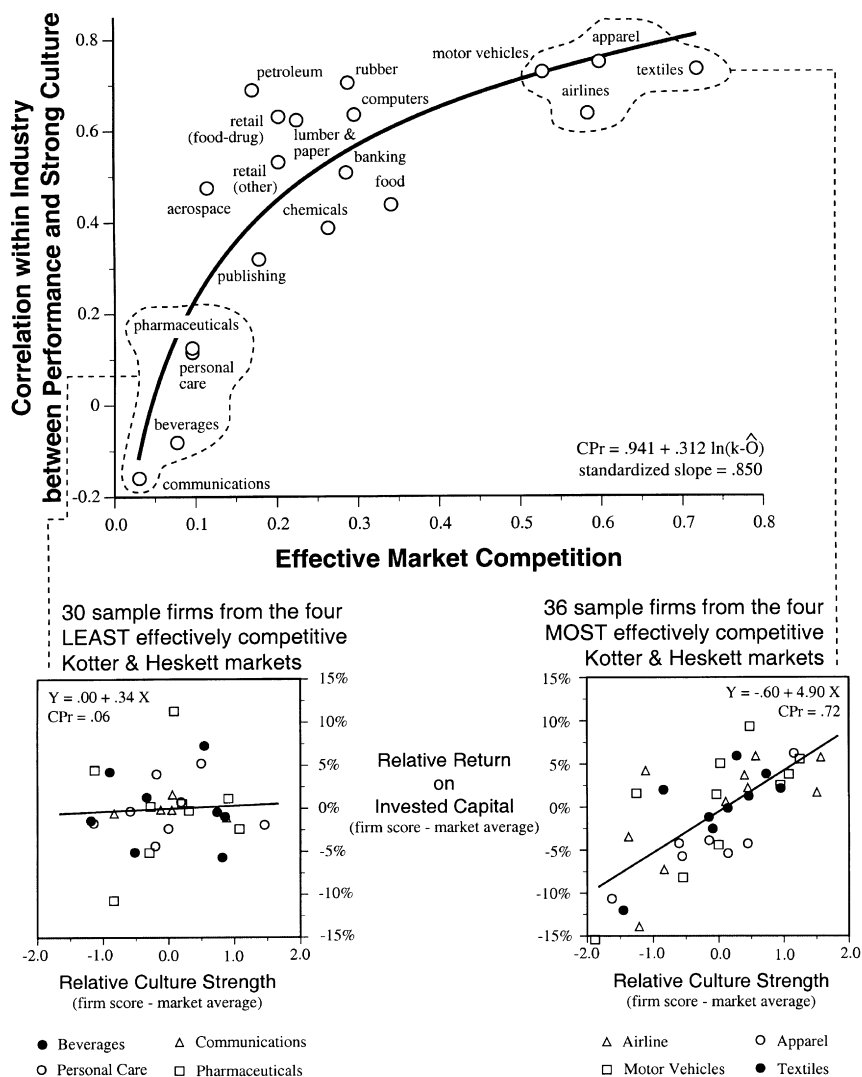


Fig. 1. Value of a Strong Corporate Culture is Contingent on Market Competition.

effect and infer from Fig. 1 that strong cultures are more valuable to firms in more competitive markets.

The horizontal axis – effective competition – is key to the observed contingency and we have not yet defined what it means for a market to be “effectively” competitive. Nor did Burt et al. (1994) define it. Readers were referred for explanation to an early draft of the paper you are now reading. More specifically, for example, it seems odd to say that the beverages market to the left in Fig. 1 is not effectively competitive given rivalry such as between Pepsi and Coke. The communications market is among the least effectively competitive of markets in the graph, but the many alternative communication producers surely create a healthy level of competition within the market (though these data describe the late 1970s and early 1980s, when there was less rivalry in communications than there is today). Common-sense questions notwithstanding, whatever it is that the measure of effective competition captures, it reveals the contingent value of strong corporate culture in Fig. 1, and it will be shown to be productive in other ways as well. In fact, we will show that more familiar measures of market competition do not as clearly reveal the contingency so obvious in Fig. 1.

Effective competition is the central concept in this paper. We derive the concept from the dual network structure of markets, internal and external. The tradition is to discuss and measure competition in terms of internal structure. Competition is between producers within a market, so producers are a natural frame of reference for thinking about competition (e.g. Swedberg, 1994; Lie, 1997, on the sociology of markets). Markets are sorted in terms of internal structure such that competition in one can be said to be some degree more intense than in another. Examples are competition in terms of the number of producers (e.g. Hannan & Freeman, 1989), their relative share of the market (e.g. Caves, 1982, pp. 8–16; Burt, 1983, Chap. 2; Weiss, 1989; Schmalensee, 1989, pp. 966–967), their recognition of one another as a frame of reference (e.g. White, 1981; DiMaggio & Powell, 1983; Burt, 1992, pp. 197–208; Han, 1994), or their network of relations with one another (e.g. Baker, 1984; Podolny, 1993; Podolny, Stuart & Hannan, 1996; Powell, Koput & Smith-Doerr, 1996).

Markets also have an external structure defined by the network of producer buying and selling in other markets. External structure is familiar from contingency theories of optimum organization (Lawrence & Lorsch, 1967), resource dependence theories of organization ties across markets (Pfeffer & Salancik, 1978; Burt, 1983, 1992; Finkelstein, 1997; even law suits across markets, Gersen, 1999; see Pfeffer, 1997, for review), institutional accounts in which producer legitimacy in a market depends on recognized affiliation with symbolic forms beyond the market (e.g. Meyer & Rowan, 1977; DiMaggio & Powell, 1983, 1991; Zuckerman, 1999; see Scott, 1998, for review), and strategy

frameworks such as the five-forces ideograph so widely used in business schools to discuss market competition (Porter, 1980). In a sense, strategic thinking has moved beyond market structure to re-emphasize the resources created down an experience curve (Stern & Stalk, 1998), and consider resources more broadly in terms of corporate deployment of assets and processes (e.g. Foss, 1997), but the value of a resource continues to be a function of the advantage it confers on an organization in a specific market. The value of a strategy for deploying resources remains contingent on the market in which the strategy is applied.

Our goal is to get another handle on the market factor in contingency theory, to produce graphs like Fig. 1 describing the contingent value of specific organizational resources. Our handle on the market factor comes from using the external structure of a market, combined with observed market profits, to infer the internal condition of the market; much as an entomologist uses the exoskeleton of a bug to determine phylum and genus. Effective competition, a measure of competitive intensity, is inferred from observed market profits predicted by the market network of dependence on other sectors of the economy. Producers with profit margins higher than expected from observed market structure must face an “effective” level of competition lower than the level implied by the observed structure. Effective competition as a concept involves its own debatable assumptions, which we discuss at the end of the paper, but it also introduces an alternative and demonstrably productive perspective on the market factor in contingency theory.

We proceed in three sections. We establish the functional form of the empirical association between performance and the network structure of aggregate American markets. We then reverse-engineer the network model to infer effective levels of competition within the markets, and demonstrate the research value of effective competition for contingency theory.

THE DUAL STRUCTURE OF MARKET COMPETITION

We use network theory to juxtaposition the performance effects of internal and external market structure (a “stylized fact” following Schmalensee’s, 1989, review of market structure research). Relying on review elsewhere (Burt, 1992), we can be brief. Structural autonomy, A , measures the extent to which producers in a market are free from the pressures of competitive pricing. A multiplier function defines structural autonomy in terms of constraint implicit in the internal and external structure of a market:

$$A = \alpha (k-O)^{\beta} C^{\gamma}, \quad (1)$$

where O measures producer coordination, k is a constant equal to maximum coordination ($k-O$) measures the internal constraint of producer competition, and C measures the external constraint of dependence on coordinated suppliers and customers (defined below). *Ceteris paribus*, performance should decrease with the internal constraint of competitive producers (β negative), and decrease with the external constraint of dependence on coordinated suppliers and customers (γ negative).

Aggregate American Markets

Every five years, the Department of Commerce publishes benchmark input-output tables of the American economy reporting the dollar value of goods exchanged between sectors of the economy. The benchmark tables are computed from a census of buying and selling in the economy. The data collection and processing involved result in benchmark tables appearing several years after the benchmark year. We use the most recent thirty years of benchmark tables; 1963, 1967, 1972, 1977, 1982, 1987, and 1992. The first four were described in Burt (1988). The later three have been available on diskette from the U.S. Department of Commerce, Bureau of Economic Analysis.

The tables contain sales and cost data with which performance can be compared across markets. We use price-cost margins, a profit measure of net income to sales introduced by Collins and Preston (1969) and widely used in market structure research: P equals dollars of value added minus labor costs, quantity divided by sales (see Burt, 1988, p. 371ff., on price-cost margins computed from input-output tables vs. the *Census of Manufactures*). For example, apparel producers had sales in 1987 of \$64,184 million, of which \$27,003 million was value added beyond the cost of supplies, of which \$17,503 million was employee compensation – leaving apparel producers with a profit margin of 15¢ on a dollar of sales (0.148 price-cost margin). Communications producers did bigger and better; \$161,127 million in 1987 sales, of which \$94,949 million was value added beyond the cost of supplies, of which \$36,761 million was employee compensation, leaving a 36¢ profit margin (0.361 price-cost margin).

The tables distinguish manufacturing and nonmanufacturing markets (“sectors” in input-output terminology) defined at a detailed level roughly corresponding to four-digit Standard Industrial Classification (SIC) categories. For example, there are 528 producer sectors in the 1982 table, from “poultry

and eggs” within the aggregate “livestock” market, to “motion pictures” within the “amusements” market. The detailed categories combine to define 77 aggregate markets that are the units of analysis in this paper.

We use aggregate markets for two reasons: First, the aggregate markets are more likely to encompass the operations of the medium and large firms so often studied in organization research. Some of the aggregate markets are too broad for organization research (e.g. beverages and food processing are combined in the aggregate food market), but many are defined at a level appropriate for organization research (e.g. tobacco, metal containers, household appliances, motor vehicles). Where aggregate markets are too broad for a research project they can be disaggregated to an appropriate level with the data on detailed categories (as we did to match input-output data to Kotter & Heskitt’s, 1992, market categories, see footnote 7). Second, we can compare aggregate markets over the thirty years for which we have data. Detailed categories can change substantially between benchmark tables, but the aggregate categories are comparable in the sense that changes occur within, rather than between, aggregate categories. We have a total of 537 market observations in seven panels, a panel for each benchmark input-output table.¹

Internal Market Constraint (β effect)

Internal market constraint refers to the competition between producers in the same market seeking the same business. Competition is inversely related to the level of coordination among producers. This is the traditional axis of market competition, varying from the lack of competition in monopoly markets (maximum producer coordination) to the intense competition of commodity markets where opportunistic undercutting of one another’s prices drives market price to the minimum possible and prevents any one producer from rising above market price (minimum producer coordination). A fitting summary is Stigler’s (1957, p. 262) conclusion to his review of market competition in economic analysis; “If we were free to redefine competition at this late date, a persuasive case could be made that it should be restricted to meaning the absence of monopoly power in a market.” The empirical task is to sort markets in terms of internal structure such that competition in one can be said to be some degree more intense than competition in the other. Of alternatives, market share measures discussed as concentration ratios are the most widely used.

Per standard practice, we use four-firm concentration ratios (market share of the four largest firms). Each aggregate market corresponds to a set of four-digit SIC categories (categories assigned to each input-output sector are published with each benchmark table). We measure producer coordination (O) in an

aggregate market by the average concentration within market segments: $O = (\sum_q S_q CR_q) / (\sum S_q)$, where S_q is the dollars of sales by establishments in SIC category q , and CR_q is the four-firm concentration ratio for SIC category q . Concentration ratios for manufacturing markets are taken from the *Census of Manufactures* and ratios in non-manufacturing are approximated with sales data in other census publications from the U.S. Department of Commerce (Burt, 1988, p. 370, 1992, pp. 89–91). Varying from 0 to 1, concentration (O) indicates the extent to which a small number of producers hold a large share of their market.

The presumption is that more concentration indicates more coordination among producers, which means less intense competition, so producers can obtain higher profit margins. For example, apparel is less concentrated than communications (respective concentration ratios of 0.262 and 0.447) with the corresponding, above-mentioned difference in profit margins (15¢ on the dollar in apparel, 36¢ in communications), and the communications margin decreases over time with decreasing producer coordination (52¢ on a dollar of sales in 1963 when AT&T held a virtual monopoly on the market, to 45¢ in 1977, to 39¢ in 1982, and down to 36¢ in 1987 with regional companies and independent producers growing over the years). The correlation between concentration and profit margin is statistically significant across markets more generally, but weak in magnitude (Schmalensee, 1989, pp. 973–976). For example, concentration and price-cost margin have a 0.31 correlation across our 364 observations in manufacturing (6.3 t -test).

External Market Constraint (γ effect)

Weak performance-concentration correlations are to be expected, according to Eq. (1), if the correlations are computed without controls for variation in the external structure of markets (e.g. see Raider, 1998, on corporate innovation). Markets are not independent production sites. The mix of goods purchased from supplier sectors is determined by production technology, which ensures a network of variable dependence among production markets. Car producers, for example, can purchase steel from one or another company, but they must purchase steel somewhere. External market constraint is the competitive disadvantage associated with being dependent on coordinated suppliers and customers. Analogous to the metaphor of countervailing power (Galbraith, 1952, for the metaphor; Lustgarten, 1975, for an illustrative effort to operationalize the metaphor), the network concept of external constraint is grounded in the sociology of Simmel (1922) and Merton (1957) describing the autonomy created by conflicting affiliations (see Burt, 1983, 1992, for application to product markets).

Measures of external market constraint begin with resource dependence. Producers in a market are dependent on another market to the extent that a large portion of producer buying and selling directly or indirectly involves the other market. There is a network of asymmetric dependence weights implicit in the input-output table of buying and selling: $w_{ij} = (p_{ij} + \sum_q p_{iq}p_{qj})^2$, $i \neq q \neq j$, where p_{ij} is the proportion of producer i business that directly involves market j and the sum is the proportion of producer business that indirectly involves market j . Proportion p_{ij} is dollars of buying and selling between markets i and j , $(z_{ij} + z_{ji})$, divided by the sum of all producer buying and selling in other markets; $\sum_j (z_{ij} + z_{ji})$, $i \neq j$, where z_{ij} is dollars of sales from market i to j in the input-output table. Dependence weight w_{ij} varies from 0 to 1 with the extent to which producer buying and selling is directly (p_{ij}) or indirectly ($\sum_q p_{iq}p_{qj}$) with establishments in market j (see Burt, 1992, pp. 54–62, for other specifications and connections with laboratory results on exchange networks).

Dependence is constraint when a buyer or supplier market contains few independent competitors. Transaction-specific constraint score c_{ij} measures the extent to which producers in market i are constrained in their transactions with market j , and the sum of transaction-specific scores measures the aggregate buyer-supplier constraint on producers in market i : $C = \sum_j c_{ij} = \sum_j w_{ij}O_j$, $i \neq j$, where O_j is the coordination of producers in sector j , which we measure with concentration in sector j (as described above).

Figure 2 is a simplified market network (useful in the next section to illustrate effective competition and link it to the descriptive results in this section). Each dot in Fig. 2 represents a producer. Lines indicate coordinating ties within markets as well as aggregate buying and selling between markets. Relations are on or off for this illustration. Markets are distinguished by circles around substitutable producers. The sociogram is a fragment of the trade network around the four producers in the gray circle buying supplies (in markets A, E, and F), and selling their output (in markets B, C, and D).

The table at the bottom of Fig. 2 shows the constraint on gray-circle producer transactions with each other sector. In this simplified network, the p_{ij} are one over the number of a market's ties. Producer coordination (O) is high to the extent that a few disconnected producers are responsible for a large proportion of market output. Let the producers in Fig. 2 be the same size, so each producer in the gray circle has a 25% market share, the coordinated producers in market A together hold a 100% market share, and each producer in disorganized market C has a 12.5% market share. Transaction-specific constraint score c_{ij} (bottom row of table) is the product of producer dependence on market j (weights w_{ij} in first row) times producer coordination in market j (O_j in second row).

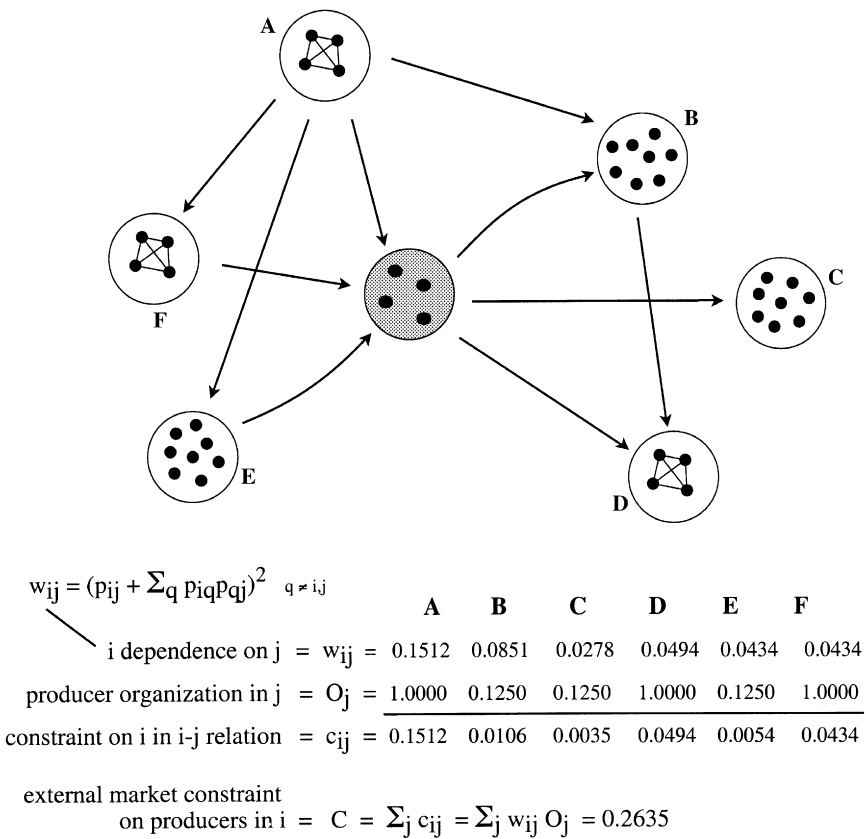


Fig. 2. Illustrative Network Fragment.

Producer transactions with market C are least constrained because C is disorganized and does no business in the other markets in this network. Producer transactions with market A are most constrained. Market A producers are tied to one another to operate as a single organization and do business in three of the other supplier-customer markets. Aggregate buyer-supplier constraint on the gray-circle producers is the sum of the six transaction-specific constraints ($C = 0.26$). We have a buyer-supplier constraint score for each market in each benchmark input-output table.

Association with Performance

Now to the empirical regularity captured by Eq. (1). Results in Table 1 connect our measures with prior work, extend the work into the 1990s, and explain our choice of functional form for a baseline model.

The four models at the top of the table are alternative functional forms linking performance with internal and external market constraint. We have several hundred observations, but they are repeated observations of the same 77 markets, and market structure at this aggregate level is quite stable over the thirty years (see Table 3 later). Therefore, we estimated each model twice. At the extreme of maximum autocorrelation, our data on each market are one observation repeated over time, which yields the estimates in the “a” column for each model (computed from data averaged over time so there is one observation per market, $N = 77$). At the other extreme of minimum autocorrelation, the five years between benchmark tables could be sufficient interval to treat each observation as independent, which yields the estimates in the “b” column for each model (computed from seven observations on 76 markets and five on the restaurant market, $N = 537$). The “b” column includes a control for trend and the 1982 drop in margins (discussed below in Table 4). Between the extremes of assuming maximum and minimum autocorrelation are various statistical methods for dealing with autocorrelation (e.g. Burt, 1988, presents similar results estimated with pooled cross-section controls for autocorrelation). We present estimates at the two extremes of autocorrelation because we reach the same substantive conclusions either way, and so ignore autocorrelation for the purposes of this paper. The summary point is that estimates of β and γ are significantly negative for the respective effects of internal and external market constraint (least clearly in linear Model I and most clearly in nonlinear Model IV, which is the model in Eq. 1).

Model I defines a linear form for the prediction. Burt (1983) described this association for 1967 with profits in American manufacturing markets defined at broad and detailed levels of aggregation, and extended the results into non-manufacturing through the 1960s and 1970s (Burt, 1988). The results are replicated in Table 1 for aggregate American markets into the 1990s; margins decline linearly with internal and external market constraint. The control for non-manufacturing adjusts for higher margins in non-manufacturing. Producer competition ($k-O$) is measured with constant k set to 1 (maximum concentration possible and slightly higher than the highest observed score of 0.963 for iron ore mining in 1977). Burt (1988) separates out the effect of constrained business with government sectors, which we have combined for the purposes of this paper with other producer buying and selling in the external constraint measure C . Similar

Table 1. Performance and Market Structure Across Three Decades.

	Linear Model $A = \alpha + \beta (k-O) + \gamma C + \delta X$		Nonlinear Model $A = \alpha (k-O)^{\beta} C^{\gamma} X^{\delta}$					
	Ia	Ib	IIa	IIb	IIIa	IIIb	IVa	IVb
Criterion Variable	P		P		P		ϵ^P	
R ²	0.337	0.306	0.409	0.322	0.456	0.344	0.649	0.538
Intercept (α)	0.110	0.100	0.052	0.053	0.041	0.045	0.898	0.943
Internal Market Constraint (one minus producer concentration; β effect)	-0.110 [-0.216] (-2.1)	-0.101 [-0.186] (-4.8)	-0.268 [-0.230] (-2.5)	-0.223 [-0.175] (-4.8)	-0.366 [-0.315] (-3.5)	-0.310 [-0.244] (-6.7)	-0.082 [-0.331] (-4.6)	-0.071 [-0.302] (-9.9)
External Market Constraint (buyer- supplier constraint index C; γ effect)	-0.497 [-0.202] (-2.0)	-0.354 [-0.158] (-4.1)	-0.211 [-0.249] (-2.7)	-0.169 [-0.183] (-5.0)	-0.309 [-0.363] (-4.0)	-0.247 [-0.267] (-7.3)	-0.068 [-0.375] (-5.1)	-0.049 [-0.289] (-9.3)
Controls (δ adjustments): Nonmanufacturing Land	0.156 (6.1)	0.154 (14.8)	0.705 (6.5)	0.707 (13.2)	—	—	—	—
1982	—	—	—	—	1.578 (7.3)	1.459 (14.1)	0.397 (11.3)	0.371 (23.2)
Year	—	-0.049 (-3.6)	—	-0.437 (-6.0)	—	-0.426 (-5.9)	—	-0.042 (-3.7)
	—	0.003 (1.4)	—	0.032 (2.5)	—	0.036 (2.9)	—	0.005 (2.5)

Note: These are ordinary least-square estimates predicting market price-cost margins. Standardized coefficients are in [brackets] and routine *t*-tests are in (parentheses). The (a) column for each model contains estimates from data averaged over time ($N = 77$). The (b) column contains estimates from each observation of each market ($N = 537$).

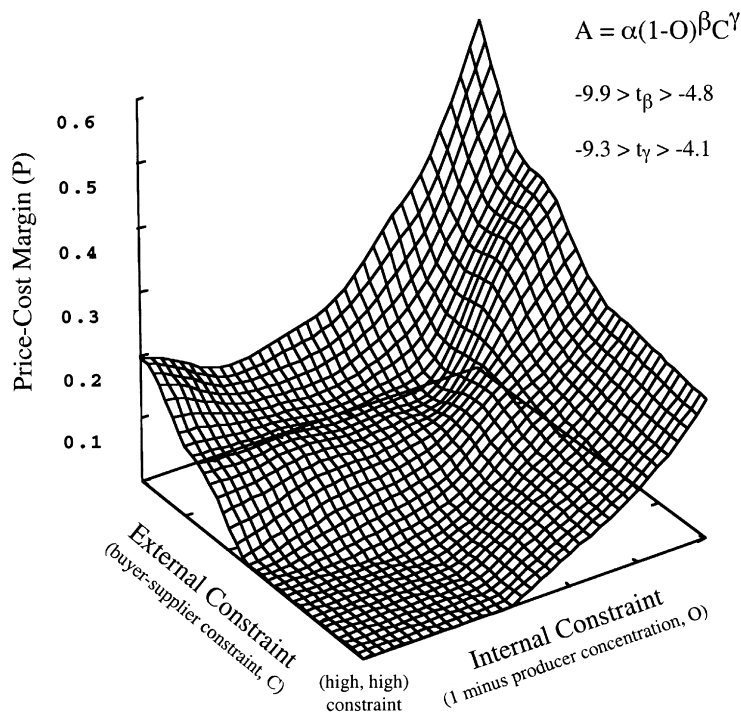


Fig. 3. Producer Performance by Internal and External Market Constraint (509 observations of aggregate American markets between 1963 and 1992).

results have been observed in other countries; Germany during the 1970s and 1980s (Ziegler, 1982; Burt & Freeman, 1994), Israel during the 1970s (Talmud, 1994), Japan in the 1980s (Yasuda, 1996), and Korea in the 1980s (Jang, 1997).

The other models in Table 1 have the nonlinear form in Eq. (1) needed to capture the steeper effects of market conditions approaching monopoly. Model II replicates the results in Burt (1992, pp. 92–100) over our broader time period. Model III is the same as II but the nonmanufacturing dummy is replaced with a land-market dummy. Margins are not uniformly higher outside manufacturing so much as they are high in five nonmanufacturing markets concerned with political control over land. The five “land” markets return much higher profits than expected from their market structure: farming (sector 2, which excludes livestock and dairy products), forestry and fishery (sector 3), coal mining (sector

7), crude petroleum and natural gas (sector 8, not to be confused with petroleum refining), and real estate (input-output sector 71). We are not concerned in this paper with the specifics of these markets, so we control for their difference from other markets and see the predicted market structure effects more clearly in the sense of stronger *t*-tests in Model III. Finally, trial and error with alternative forms led us to Model IV for the clearest association with market structure. Estimates come from regressing price-cost margins over $\ln(A)$ rather than regressing $\ln(P)$ over $\ln(A)$ as in Models III and II. Model IV is our baseline model for the remainder of the paper.²

Figure 3 summarizes the Model IV market structure effects in a visual display. The two dimensions of market structure define the floor of the three-dimensional graph. The front corner of the floor is maximum internal and external constraint ($O = 0, C = 1$). The back corner of the floor is maximum structural autonomy ($O = 1, C = 0$). The wire-mesh performance surface is based on 509 observations (all markets excluding the four distinguished by the land dummy in Table 1). The surface is high over combinations of internal and external market constraint where price-cost margins are high (using a distance-weighted least-squares smoothing to average adjacent price-cost margins).

Four points are illustrated in Fig. 3. First, the surface slopes down from maximum structural autonomy at the back of the graph to maximum constraint at the front of the graph (i.e. more competitive markets yield lower profit margins). Second, the surface slopes downward more steeply at the back of the graph showing the stronger effect of market structure on performance for more autonomous producers. Third, the back edge of the surface slopes down more steeply to the left than to the right showing the stronger effect of internal market constraint. Fourth, the smooth surface implies continuous performance effect across mixtures of internal and external market constraint.

EFFECTIVE COMPETITION

Producers with profit margins lower than expected from observed market structure must face an “effective” level of competition higher than the level implied by the observed structure.

Given a functional form in Model IV, Table 1, for the association between performance and market structure, we can reverse-engineer the model to infer the effective level of producer coordination from observed performance and producer dependence on other markets. Begin with the logarithm of Model IV:

$$\hat{P}_i = \ln(A_i) = \ln(\alpha) + \beta[\ln(k - O_i)] + \gamma[\ln(\sum_j w_{ij} O_j)], i \neq j \quad (2)$$

where the equation is written for producers in market i doing business with supplier and customer markets j , buyer-supplier constraint C is replaced with its definition, and $\hat{P}$ is the price-cost margin predicted by market structure (observed margin minus a residual not predicted by market structure). The residual includes effects of the control variables in Table 1 plus unknown other effects presumed random. This is an equation for testing hypotheses about the effects. We know, from past work and the results in Table 1, that performance varies across market structures as illustrated in Fig. 3 and that β and γ are significantly negative.

Re-write Eq. (2) to infer the effective level of producer coordination from observed performance and producer dependence on other markets:

$$P_i = \ln(A_i) = \ln(\alpha) + \beta[\ln(k - \hat{O}_i)] + \gamma[\ln(\sum_j w_{ij} \hat{O}_j)], i \neq j \quad (3a)$$

which is the regression model for the following network model obtained by taking the antilog of Eq. (3a):

$$A_i = \alpha [(k - \hat{O}_i)]^r + (\sum_j w_{ij} \hat{O}_j)^\gamma, i \neq j \quad (3b)$$

where r is the effect ratio of internal vs. external market constraint (β/γ , discussed below), and the concentration measure of producer coordination (O) in Eq. (2) has been replaced by an unobservable “true” level of coordination $\hat{O}$.

Error is the key difference between Eqs. (2) and (3). Errors in Eq. (2) occur in the performance variable (thus the predicted level of performance, $\hat{P}$). There is no error in the Eq. (3) performance variable. Profit is the log of structural autonomy, $P = \ln(A)$, and structural autonomy is the exponential of profit, $A = e^P$. Errors in Eq. (3) occur in the producer-coordination variable (thus the predicted level of producer coordination, $\hat{O}$). Instead of asking what value of outcome Y can be expected from known causal variable X , we ask in Eq. (3) what value of X would be necessary to generate the known outcome Y . To obtain their known profit margin, operating from their known network position in the economy, how well coordinated must producers be?

Variable $\hat{O}$ measures the “effective” coordination of producers in the sense that coordination is inferred from the market structure effect on performance. Effective often means “good,” but it should be clear that such is not our use of the word. Effective coordination is the level of producer coordination implicit in the observed level of producer performance (P) and the observed network of producer dependence relations with other markets (w_{ij}). Given $\hat{O}$ as the effective level of coordination among producers, $(k - \hat{O})$ is the effective level of

competition within their market – which we propose to use as the market context factor in contingency theory.³

Identification and Computation

Equation (3) defines for N markets a system of N equations containing $N + 2$ unknowns (intercept α , slope γ , and an effective level of coordination within each of N markets). To define a unique solution, we gain a degree of freedom by normalizing scores to the maximum in any market ($\hat{O} = \hat{O}/\hat{O}_{\max}$), and gain one or more by fixing scores in one or more markets to equal observed concentration ratios, or fixing the minimum score at zero ($\hat{O} = (\hat{O} - \hat{O}_{\min})/(\hat{O}_{\max} - \hat{O}_{\min})$).

We use a Newton-Raphson algorithm to solve for the unknowns (see Appendix for details and analogy to network eigenvector models). The algorithm adjusts observed levels of producer coordination to improve the match between observed and expected performance. Where observed performance is higher than expected from market structure, increase the effective coordination of producers and decrease the effective coordination within key supplier or customer markets. The process is illustrated in Table 2 below. Constant k is set to a value slightly higher than the maximum possible coordination score of 1 to ensure that producer competition is always a positive fraction so variation in buyer-supplier constraint can affect performance in even the most organized markets (we set k to 1.001). Replacing β with ratio r allows us to preserve the effect balance between internal and external constraint when inferring $\hat{O}$ ($r = \beta/\gamma = 0.071/0.049 = 1.45$ for the aggregate American markets in Table 1, cf. Table 4).

Residual Coordination

The difference between concentration and effective coordination measures the extent to which concentration understates the level of coordination within a market: $\hat{O} = \mu + O + \phi$, so;

$$\phi = \hat{O} - O - \mu \quad (4)$$

where μ is an intercept (mean $\hat{O}$ – mean O), $\hat{O}$ is the effective coordination of producers, O is the level of coordination indicated by concentration, and ϕ is an error term measuring residual coordination. Residual coordination varies from positive to negative scores. A positive score indicates producers better

coordinated than they appear to be. Zero indicates producers effectively as coordinated as they appear to be. A negative score indicates producers effectively less coordinated than they appear to be.

Numerical Illustration

Table 2 contains effective coordination scores for the network fragment in Fig. 2. Three panels of data are displayed. The first contains concentration and dependence weights as already discussed for the illustrative network.

The second describes coordination when performance is determined by observed market structure. Performance differences between the markets follow from the earlier discussion of Fig. 2. Performance is high in market D, for

Table 2. Effective Coordination in the Fig. 2 Network Fragment.

	A	B	C	D	E	F	Gray Circle
Observed Market							
Structure (Fig. 2)	—	0.085	0.000	0.000	0.085	0.085	0.340
	0.151	—	0.000	0.151	0.000	0.000	0.340
Interdependence	0.000	0.000	—	0.000	0.000	0.000	1.000
Weights (w_{ij})	0.000	0.340	0.000	—	0.000	0.000	0.444
	0.340	0.000	0.000	0.000	—	0.000	0.391
	0.340	0.000	0.000	0.000	0.000	—	0.391
	0.151	0.085	0.028	0.049	0.043	0.043	—
Concentration (O)	1.00	0.125	0.125	1.00	0.125	1.00	0.250
Network Constraint (C)	0.191	0.388	0.250	0.154	0.438	0.438	0.264
P Determined by Observed Structure							
Performance (P)	0.856	0.108	0.152	0.878	0.096	0.773	0.162
Effective Coordination ($\hat{O}$)	1.00	0.125	0.125	1.00	0.125	1.00	0.250
Residual Coordination (ϕ)	0.00	0.00	0.00	—	0.00	0.00	0.00
P Not Determined by Observed Structure							
Performance (P)	0.086	0.108	0.152	0.878	0.096	0.773	0.162
Effective Coordination ($\hat{O}$)	0.148	0.414	0.661	1.00	0.000	0.997	0.351
Residual Coordination (ϕ)	-0.844	0.297	0.543	—	-0.117	0.005	0.109

Note: Performance in the first solution is determined by the observed market structure; $P = [(k-O)C]^{-0.1}$, with constant k set to 1.001. Effective coordination in market D is fixed to its observed value (1.0). Parameters for the first solution are $\alpha = 1.000$ and $\gamma = -0.100$ with a 1.000 multiple correlation. For the second solution, $\alpha = 0.899$ and $\gamma = -0.121$ with a 1.000 multiple correlation.

example, because producers are completely coordinated facing minimum constraint from other markets. Markets E and F face strong external constraint, but F is better organized to handle it and so shows the higher performance. Residual coordination is zero in every market. Effective coordination ($\hat{O}$) equals observed coordination (O) when performance is determined by observed market structure.

The bottom panel describes what happens when performance is not predicted by observed market structure. Performance is the same as in the first panel – except in market A. Instead of the 86¢ profit expected on a dollar of sales, we lowered the observed margin to 9¢. Market performance in the second panel is therefore not determined by market structure. Now effective coordination differs from concentration. For example, observed concentration of 1.0 in market A is adjusted down to 0.148 effective coordination. If market A is under-performing, then producers must be less coordinated than they appear to be. Also, key supplier or customer markets must be better coordinated than they appear to be. Market A is most dependent on the gray-circle market ($w_{17} = 0.340$) and the observed 0.25 concentration in the gray-circle market is adjusted up to 0.351 effective coordination. If market A producers are less coordinated than they appear to be, then the constraint on their customer markets is less. Since performance in those markets has not changed, producers in the customer markets must be less coordinated than they appear to be. The low observed concentration in E is adjusted down to minimum effective coordination. Changing coordination within the gray-circle market changes the external constraint on every other market because every other market depends on it (seventh column of w_{ij} in Table 2 are all substantially over zero). The most affected is market C. Market C is exclusively and entirely dependent on the gray-circle market (w_{37} is the maximum possible, 1.0). Since market C performance is unchanged and it now faces more severe supplier constraint, producers in market C must be better coordinated than they appear to be. The concentration of 0.125 in market C is adjusted up to 0.661 effective coordination.

This story can be told various ways to make the point: Effective coordination traces inconsistencies between performance and producer coordination through the network of market dependencies to find levels of effective coordination consistent with the observed performance differences between markets.

WHY EFFECTIVE COMPETITION?

Effective competition can be a productive market variable for contingency theory. In this final section, we demonstrate the point in three ways: more

reliable measurement robust to ad hoc adjustments and exogenous shocks to markets, more accurate measurement revealing contingency, and a market factor integrating case with comparative research.

Reliable Measurement

The presumption in predicting performance from market structure is that prediction errors are due to error in measuring the dependent variable, performance.

But which of the three elements in the baseline model – performance, internal structure, and external structure – is most subject to measurement error? Price-cost margins (P) and the network of market dependencies (the w_{ij} used to define buyer-supplier constraint C) are observable with census precision down to the detailed level of about 500 production markets in the American economy and they measure what they are presumed to measure. Price-cost margins measure the extent to which producers receive income above production costs. Market dependency w_{ij} measures the extent to which buying and selling by producers in market i directly or indirectly involves producers in some other market j .

In contrast, producer coordination (O) measured by concentration data involves substantial assumption and guesswork. The assumption is that intensity of competition among producers decreases with the market share of the largest firms. Economists have long been troubled by the ambiguous connection in theory between concentration and competition (e.g. Schmalensee, 1989, p. 966). No competition is clear at the extreme of monopoly – one company controlling all producers means no alternative offers except as the company allows them. It is not clear in theory how competition increases with the introduction of additional companies, though there are empirical results on profits decreasing most with the presence of third and subsequent large competitors (Kwoka, 1979; Burt, 1983, pp. 16–32), and there are abundant empirical results on mortality correlates of the number of companies in a market (e.g. Hannan & Freeman, 1989). Causal direction is also ambiguous. Concentration is assumed to measure a lack of competition, which allows producers to increase profit margins. An alternative is to argue that superior firms earn high profits and large market share, which means that correlation between concentration and profit is spurious (Peltzman, 1977; Ravenscraft, 1983; Weiss, 1989, pp. 7–10).

There would be problems with concentration measures of coordination even if there were a linear connection between competition and concentration. Weiss (1989, pp. 1–10) offers a succinct review of issues. Local competition is an issue we find especially troubling since concentration ratios are computed

for the national economy. Some services, such as hotels, restaurants, and government-sanctioned utilities, are typically sold locally rather than nationally (e.g. Baum & Mezas, 1992, on local crowding among hotels; Ingram and Roberts, 2000, on hotel performance and competitor friendship networks). The cost of shipping can also protect local producers from distant producers. Stone, concrete, and the like are obvious examples, but in the history of almost all industries there is an element of local competition due to transportation barriers (e.g. Bigelow, Carroll, Seidel & Tsai, 1997, on regional competition in the automobile industry). Even in national markets, however, there can be a preference for personal ties between supplier and customer such that competition is more local than national. Romo and Schwartz (1995) describe companies in inter-dependent markets moving to the geographical area around a core firm to facilitate flexible, personal ties (see Romo & Schwartz, 1993, on the growth of regional service economies; cf. Sorenson & Stuart, 2001, on geographic concentration and social networks). Whether for reasons of transportation costs, government regulation, or the importance of personal ties, competition in certain markets is more local than national. The largest firms in a locally competitive market can have a small share of the national market, even if each dominates their local market. In such a market, concentration at the national level understates the level of competition, so the profit margin determined by competition will be higher than expected from the national level of concentration. And concentration ratios in manufacturing are the better data! Concentration in nonmanufacturing is more obviously affected by measurement error since ratios are approximated from sales at the level of whole firms that span multiple markets (Burt, 1988, p. 370, 1992, pp. 89–91).

The problems with concentration measures of producer coordination are acceptable in the sense that they have precedent in published empirical research and continue to generate expected market structure effects.

However, the ambiguous measurement they represent is troubling, and unnecessary if error is more accurately handled in the network model of effective competition. The presumption in effective competition is that performance can be observed and measured more accurately than producer coordination. Coordination is treated as an unobserved market condition to be inferred from performance and the constraint implicit in a market's external structure of transactions with supplier and customer markets. An attractive consequence is that effective competition adjusts automatically for exogenous shocks to market competition that affect producer performance. We can illustrate this point with the effect of imports on American markets in 1982.

Market Structure Stability

Internal and external market constraint are dramatically stable across the thirty years on which we have data. Table 3 lists correlations for the repeated measures; producer concentration (O) in the upper diagonal, buyer-supplier constraint (C) in the lower diagonal. There are ups and downs in individual markets (e.g. concentration and price-cost margin increase from 1972 to 1977 within oil and gas drilling, then drop back down in 1982 to their level before the energy crisis), and trend in some markets (e.g. the decreasing concentration and margins in communications mentioned earlier), but the primary feature of the table is stability (see Burt, 1988, for detailed results on stability through the 1960s and 1970s). Average scores do not change over time ($F_{6,530}$ is 0.07 for concentration and 0.13 for buyer-supplier constraint, giving the constant-mean hypothesis a 0.99 probability of being true). A single principal component accounts for 96% of the concentration variance in the seven panels, and 89% of the buyer-supplier constraint variance in the seven panels (and going beyond Table 3, 93% of the variance in the seven vectors of dependence weights w_{ij} used to compute buyer-supplier constraint).

We estimated two covariance models to better understand stability at this aggregate level. A single-factor model presumes complete stability: market structure has a true value Y that is constant over time but observed with error in panel t ($y_t = d_t Y + e_t$). A simplex model says that stability is more short term than long-term: the structure observed in panel t is an incremental change from structure in the preceding panel ($y_t = d_t y_{t-1} + e_t$). The strong correlations in Table 3 and extensive variance described by a single principal component

Table 3. Observed Market Structure Across Three Decades.

	1963	1967	1972	1977	1982	1992	Mean	S.D.
1963	0.992	0.964	0.934	0.922	0.907	0.952	0.348	0.245
1967	0.961	0.977	0.943	0.933	0.920	0.964	0.350	0.242
1972	0.871	0.933	0.959	0.934	0.951	0.987	0.357	0.254
1977	0.825	0.888	0.926	0.932	0.940	0.984	0.354	0.249
1982	0.724	0.796	0.843	0.870	0.934	0.949	0.335	0.212
1987	0.792	0.850	0.899	0.910	0.972	0.978	0.342	0.223
1992	0.696	0.770	0.835	0.881	0.972	0.965	0.351	0.238
Mean	0.062	0.058	0.060	0.062	0.057	0.059	0.057	
S.D.	0.049	0.045	0.050	0.053	0.046	0.052	0.053	

Note: These are correlations, means, and standard deviations for the aggregate markets, producer concentration (O) in the upper diagonal and buyer-supplier constraint (C) in the lower diagonal. With pairwise deletion, there are 76 markets during the 1960s and 77 markets thereafter.

suggests a single-factor model, but there is also evidence of simplex structure in which panels further apart in time are less correlated. Buyer-supplier constraint in adjacent panels is correlated 0.94 on average ($0.94 = [0.961 + 0.933 + 0.926 + 0.870 + 0.972 + 0.965]/6$), which decreases to 0.90 between panels two time periods apart, 0.85 for panels three time periods apart, 0.80 for panels four time periods apart, 0.78 for panels five time periods apart, and 0.70 between the first and last panels which are six time periods apart. Neither model, however, describes the data.⁴ We also get rejections beyond a 0.001 level of confidence with log scores, and models fit to the more comparable five panels after 1970. Aggregate market structure is clearly stable across the seven panels, but there is more than a single-factor or simplex process responsible for the stability.

Disconnect from Market Performance in 1982

The association between performance and market structure, however, is not stable. Table 4 shows what happens when our baseline Model IV in Table 1 is estimated for each panel separately. Through the 1960s and 1970s, then again in 1987 and 1992, the association is as depicted in Fig. 3 – price-cost margins decrease with the internal market constraint of disorganized producers (t -tests of -2.9 to -5.4) and the external market constraint of dependence on coordinated suppliers and customers (t -tests of -3.2 to -4.8).

In 1982, however, margins are independent of producer concentration (-1.7 t -test). There is evidence of buyer-supplier constraint lowering performance (-3.2 t -test), but the evidence is hardly reassuring. If negotiating with coordinated suppliers and customers erodes performance, why doesn't producer coordination enhance performance?

Imports an Exogenous Shock in 1982

The market model's failure to predict performance in 1982 can be traced to an exogenous shock. Except for their significant drop in 1982 (-2.9 t -test, $P < 0.01$), margins are stable across the panels preceding 1982 ($F_{3,302} = 0.30$, 0.82 probability of no difference) and again after 1982 ($F_{5,454} = 0.82$, 0.53 probability of no difference before and after 1982). The 1982 exogenous shock cannot be traced to the market structure variables. The results in Table 3 show that market structure continued through the 1980s as it was during the 1960s and 1970s, and the 1982 failure remains if we exclude the four markets showing unusual change in the preceding decade (-0.8 t -test for concentration association with performance in 73 markets excluding oil and gas drilling, ferrous ores mining, nonferrous ores mining, and chemical mining). Further, the exogenous shock in 1982 was a shock to certain markets more than a shock to the whole

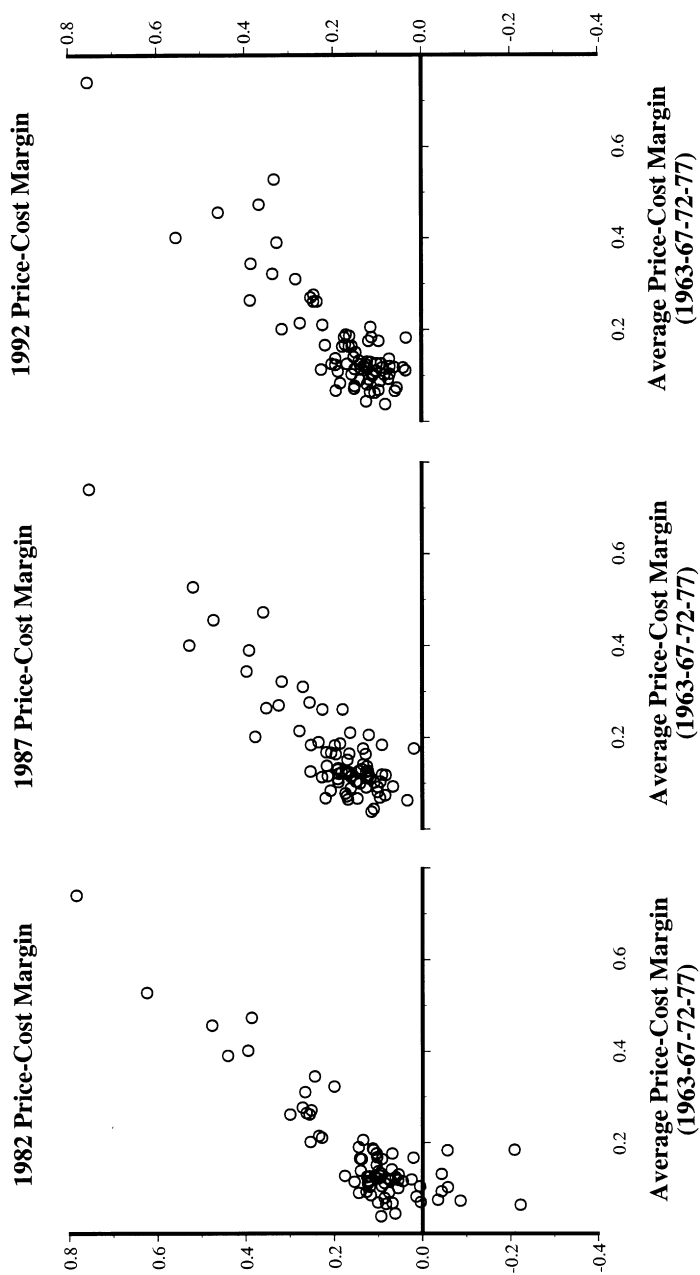


Fig. 4. Price-Cost Margins Over Time.

economy. Figure 4 shows how average price-cost margins in the 1960s and 1970s are associated with later margins. Negative margins in 1982 are the most apparent difference between the graphs. No aggregate market at any other time has a negative margin; only in 1982, when eight markets are in the red. Margins in the other 69 markets are similar in 1982 to their averages across the preceding panels (0.93 correlation). Also, the association lost in 1982 between performance and market structure is regained if we put aside the eight markets with negative profit margins (-2.5 t -test for the β effect of internal market constraint in the remaining 69 markets, and -2.7 t -test for the γ effect of external market constraint).

Further study (not reported here) of the markets with negative margins led us to imports as the exogenous shock responsible for the 1982 failure. The search is an interesting story in its own right, but it is sufficient for the purposes of this paper to show the result. The last five columns in Table 4 contain estimates of our baseline model with a control added for the market share held by imports in each market.⁵ Imports have a negative effect on profit margins (see Schmalensee, 1989, p. 976, for similar results in other countries). The point is that the baseline association between performance and market structure is again apparent when imports are held constant. In the third to the last column in Table 4, 1982 margins decrease significantly with the internal market constraint of disorganized producers (-2.5 t -test) and the external constraint of dependence on coordinated suppliers and customers (-3.1 t -test). Imports continue to increase their share of American markets after 1982 (the average market share of imports is 6%, 7%, 8%, 12%, and 13% in 1972, 1977, 1982, 1987, and 1992 respectively), but they do not have in 1987 or 1992 the disruptive effect that they had in 1982.

Effective Competition Adjusts for Exogenous Shock

Thus, the market model failed in 1982 because we measured producer coordination in terms of domestic producers. The market share of the four largest domestic producers overstated the effective level of producer coordination when there are foreign competitors in the market.

The significant point with respect to effective competition is that the exogenous shock of imports in 1982 is captured by effective competition without us having to know about the shock. When 1982 margins were less than expected in certain markets, the network model adjusted down the effective coordination of producers in the markets and so adjusted up the effective level of competition within the markets. Negative residual coordination (ϕ in Eq. 4) identifies markets in which the effective coordination of producers is less than implied by concentration.

Table 4. Performance, Market Structure, and Imports.

	1963	1967	1972	1977	1982	1987	1992	1972	1977	1982	1987	1992
R ²	0.495	0.505	0.629	0.646	0.520	0.577	0.445	0.657	0.689	0.583	0.590	0.535
Effect Ratio of Internal to External Constraint ($r = \beta/\gamma$)	1.45	1.35	1.17	1.07	0.62	1.78	1.02	1.45	1.42	1.00	1.99	1.49
Intercept (α)	0.919	0.937	0.881	0.938	0.850	0.957	0.955	0.830	0.882	0.779	0.931	0.890
Internal Market Constraint (One minus Producer Concentration; β effect)	-0.084 [-0.349] (-4.0)	-0.077 [-0.323] (-3.8)	-0.083 [-0.417] (-5.4)	-0.065 [-0.360] (-4.9)	-0.048 [-0.138] (-1.7)	-0.098 [-0.398] (-5.0)	-0.058 [-0.257] (-2.9)	-0.094 [-0.472] (-6.0)	-0.075 [-0.419] (-5.8)	-0.069 [-0.199] (-2.5)	-0.106 [-0.431] (-5.3)	-0.076 [-0.339] (-4.0)
External Market Constraint (buyer-supplier Constraint index C; γ effect)	-0.058 [-0.308] (-3.6)	-0.057 [-0.301] (-3.5)	-0.071 [-0.374] (-4.8)	-0.061 [-0.308] (-4.0)	-0.078 [-0.284] (-3.2)	-0.055 [-0.280] (-3.4)	-0.057 [-0.293] (-3.2)	-0.065 [-0.342] (-4.4)	-0.053 [-0.266] (-3.6)	-0.070 [-0.254] (-3.1)	-0.053 [-0.270] (-3.3)	-0.051 [-0.265] (-3.1)
Land (δ adjustment)	0.329 (7.8)	0.338 (8.1)	0.410 (10.6)	0.387 (10.6)	0.490 (8.9)	0.375 (9.5)	0.344 (7.2)	0.411 (11.0)	0.383 (11.1)	0.483 (9.3)	0.371 (9.4)	0.351 (8.0)
Market Share Of Imports								-0.020 [-0.181] (-2.4)	-0.026 [-0.222] (-3.1)	-0.036 [-0.262] (-3.3)	-0.012 [-0.117] (-1.5)	-0.030 [-0.313] (-3.7)

Note: These are ordinary least-squares estimates for model IV in Table 1 predicting price-cost margins computed from input-output data on the 77 markets. Standardized coefficients are in [brackets] and routine *t*-tests are in (parentheses). Market share of imports is added to the estimation equation as $\ln(0.01 + F)$, where *F* is the ratio of imported goods sold over total goods sold. Import data are not available in the 1963 or 1967 input-output tables.

It is not surprising to learn that the residual coordination of producers in 1982 is correlated with imports. A quarter of the variance in residual-coordination adjustments to concentration in 1982 can be predicted from imports (-0.56 correlation with log imports, -5.8 t -test). The negative association shows that as the market share of foreign-made goods increased, domestic producers were effectively less coordinated than concentration implied. *With its automatic adjustment for exogenous change, effective competition is more reliable than observed market structure as the context variable in contingency theory.*

Moreover, effective competition makes its adjustments simultaneously in supplier and customer markets up and down the production chain that runs through the market directly affected by an exogenous variable such as imports. For example, the price-cost margin in iron ore mining decreased from 21¢ in 1972, to 18¢ in 1977, and 6¢ in 1982. Organization within the market is part of the story. Concentration was above 90% through the 1970s, then dropped to 69% in 1982. Direct imports are part of the story. Imported iron ore held about a third of the market through the 1970s, which increased slightly to 37% in 1982. But a systematic change that mirrors the lost profit margin is downstream of the producers. Iron mining depends on sales to steel companies; 87% of their sales in 1982, and imports take an increasing share of the steel market in the preceding years (9% in 1972, 11% in 1977, 16% in 1982). By the end of 1983, iron ore mining had virtually ceased in the western half of the country, and federal legislation had been introduced in both houses of the Congress to limit imports to a proportion of domestic production (Klinger, 1983).

Leather shows a more complex production-chain effect. The price-cost margin for leather tanning and preparation decreased from 8¢ during the 1970s to -3 ¢ in 1982. Downstream imports help explain leather's poor performance in 1982. Shoe manufactures accounted for 72% of domestic leather sales in 1982, and domestic shoe production was down. The market share of imported shoes increased dramatically to record levels in 1982, primarily because of the removal of the Orderly Marketing Agreements with Korea and Taiwan, and the strength of the U.S. dollar against foreign currencies. The 19% market share of imports in 1972 grew to 38% in 1982. Leather producers were simultaneously oppressed by exogenous change upstream. Supplier exports increased the price of hides. In the early 1980s, some countries such as Argentina, Brazil and India used embargoes and export taxes to restrict hide exports to encourage the growth of their own leather markets. World demand shifted to American hide suppliers, raising the price of hides, and so increasing the cost of supplies for American leather producers. In addition, domestic customers were buying more imported leather. Imported leather increased from 13% of the market in the 1970s to 19% in 1982. As the Department of Commerce report summarized the situation

(Byron, 1983, p. 4); "The five-year outlook for the U.S. leather tanning and finishing industry remains poor."

Contingency Revealed

Our second point on the value of effective competition is its ability to make contingency more apparent. Return to the "culture effect" discussed at the beginning of the paper – a strong corporate culture can improve the economic performance of a firm. Kotter and Heskett (1992) offer a rare opportunity to test the culture effect with data on performance and strength of corporate culture for a selection of firms in a variety of broad markets analogous to the market categories in *Fortune* magazine. We use the 180 firms in 19 markets summarized in Fig. 1.

To measure the relative strength of culture, Kotter and Heskett mailed questionnaires to the top six officers in each sample company, asking them to rate the strength of corporate culture in the other firm selected for study in their market. Respondents were given three indicators of a "strong" culture (Kotter & Heskett, 1992, pp. 159–162): (1) managers in the firm commonly speak of their company's "style" or way of doing things; (2) the firm has made its values known through a creed or credo and has made a serious attempt to encourage managers to follow them; and (3) the firm has been managed according to long-standing policies and practices other than those of just the incumbent CEO. Responses were averaged to define the strength of a firm's corporate culture, and we remove negligible market differences in culture strength by subtracting from each firm's score the average for all sample firms in the market (Burt et al., 1994, p. 347; horizontal axis of the graphs at the bottom of Fig. 1).

Kotter and Heskett (1992, pp. 166–174) list three measures of economic performance: net income growth from 1977 to 1988, average return on invested capital from 1977 to 1988, and average yearly increases in stock prices from 1977 to 1988. The three performance measures are reported to have correlations of 0.46, 0.31, and 0.26 respectively with culture strength (Kotter & Heskett, 1992, p. 189). We use average return on invested capital to measure performance because it is most similar to the price-cost margins we have from the input-output tables, and its reported .31 correlation with culture strength is intermediate between the alternative performance measures. Almost half of the performance variance between firms can be predicted from the market in which a firm operates.⁶ To study performance net of market differences, we subtract from each firm's performance score the average score for all sample firms in the market (vertical axis of the graphs at the bottom of Fig. 1).

Figure 1 shows that the culture effect – strong on average across firms – varies from dramatic in some markets to nothing in others. The -0.85 correlation in the graph at the top of Fig. 1 shows how closely the culture effect varies with market competition (analysis with effective competition computed for more detailed market categories yields the same contingency function, Burt et al., 1994, p. 368).⁷ The nonlinear regression line in the graph is a contingency function in the sense of describing how the culture effect is a function of market competition. For any specific level of market competition, the contingency function defines an expected strength of correlation between culture strength and producer performance.⁸

The significant point with respect to effective competition is that the contingency visible in Fig. 1 is less visible, in fact, virtually invisible, if markets are ordered on the horizontal axis of Fig. 1 by their observed structure. Note the regression equation in the corner of the graph at the top of Fig. 1 – the contingent value of a strong corporate culture (measured by the correlation between culture and performance in a market, CPr) increases with the effective level of competition in the market ($k-\hat{O}$). The 0.85 correlation across markets is 0.46 for producer concentration ($1-O$), or 0.20 if we replace effective competition with buyer-supplier constraint (C). If we predict CPr with all three market measures, only effective competition is associated with the contingent value of a strong corporate culture; 5.0 t -test for ($k-\hat{O}$), 0.4 t -test for ($1-O$), 1.5 t -test for C .

Results in Table 5 make the point at the organization level. Firm performance is predicted from culture strength and a contingency function keyed to market competition. The positive association between performance and culture strength increases significantly with the effective level of competition in a firm's market (4.3 t -test in the second column of Table 5). There is no association with either of the two measures of observed market structure (-0.2 and 1.1 t -tests for internal and external constraint).

A reviewer suggested that we add other concentration measures to the table, measures adjusted for the problems discussed in the preceding section; localized competition, government regulation, imports, and so on. The suggestion takes us back to the preceding section. The problem with such tests is the lack of a definitive adjustment against which effective competition can be tested. Reasonable adjustments to concentration are a list of variably ad hoc alternatives, with the significance of any one more a function of data than theory (e.g. adjusting concentration for imports is in theory reasonable at any time, but the results in Table 4 show that the adjustment is only in 1982 critical to seeing the expected association between performance and producer concentration).

Table 5. Contingent Culture Effect at the Level of Individual Firms.

Squared Multiple Correlation	0.255	0.334	0.255	0.260
Intercept	0.000	0.000	0.000	0.000
Corporate Culture, Relative Strength (firm score-market average)	3.053 (7.4)	9.451 (6.2)	2.937 (3.9)	2.369 (3.1)
Interaction between Corporate Culture and Market Competition:				
Effective Competition ($k-\hat{O}$)	—	9.365 (4.3)	—	—
Observed Internal Market Constraint (one minus producer concentration, $1-O$)	—	—	-0.348 (-0.2)	
Observed External Market Constraint (buyer-supplier constraint index, C)	—	—	—	22.42 (1.1)

Note: These are ordinary least-squares estimates predicting a firm's relative performance within its market (vertical axes in graphs as the bottom of Fig. 1) from the relative strength of its corporate culture (horizontal axes in graphs at the bottom of Fig. 1), with an adjustment for stronger culture performance association in more competitive markets. Effects were estimated with market performance differences held constant using 18 dummy variables to distinguish the 19 markets. Slope adjustments with log market structure variables are similar; 4.3, 0.5, and 1.5 *t*-tests respectively for the three interaction terms in the table.

Moreover, if one managed to capture all relevant adjustments, adjusted concentration would equal the effective coordination of producers ($\hat{O}$), and so duplicate our results in Fig. 1 and Table 5 with effective competition. The effective competition model saves us the labor that would otherwise be spent on learning what adjustments are needed in any one year. Concentration is adjusted automatically to balance producer performance with the constraint implicit in producer buying and selling in other sectors of the economy.

Our results do not reject producer concentration or buyer-supplier constraint as useful measures. Effective competition might or might not work as well at revealing the contingent value of other aspects of organizational form. That is a task for future research. What we do know from the above results is that for at least one often-discussed aspect of organizational form – the value of having a strong corporate culture – unobserved effective competition reveals contingency better than familiar, observable conditions of producer concentration or buyer-supplier constraint.

Integrating Case and Comparative Research

Our third point on the research value of effective competition is the bridge it provides for integrating case and comparative research. To generalize the culture effect to firms not in Kotter and Heskett's study, we have to know how their firms constitute a sample of organizations. Such knowledge rarely exists in organization research because firms are almost never selected for study by probability criteria (e.g. Kotter & Heskett's, 1992, p. 19, sampling frame is defined in the book by its goal to; "get a large and diverse sample of companies"), and even the best efforts are based on quota-sampling frames that do not compare across studies (size in sales or assets, market categories, geographical regions, etc.). Standard operating procedure is to publish case studies as if they represented organizations more generally, and convenience samples as if they were probability samples, whereupon routine statistical inference can be used to guide generalizations. Given the general acceptance of contingency theory, and the sophisticated sampling frames possible with current technology, it is surprising to see so little attention given to sampling organizations from which population inferences can be made (for an exception, see Kalleberg, Marsden, Aldrich & Cassell, 1990; Kalleberg, Knoke, Marsden & Spaeth, 1996, esp. Chap. 2, on strategies for sampling organizations).

Effective competition is a bridge for inferring population parameters from sample statistics. Given a contingency function (as in Fig. 1), and census network data available on the population of markets (input-output tables), apply the contingency function to the census data to make inferences about the economy as a population of organizations. For example, the expected culture effect $E(CPr)$ in Kotter and Heskett's sample of firms equals the sum across markets i of the proportion $P(i)$ of sample firms drawn from market i times the culture-performance correlation $F(i)$ predicted by the contingency function for the market; $E[CPr] = \sum_i P(i) F(i) = 0.49$. The 0.49 correlation expected between culture strength and performance in the Kotter and Heskett data from the contingency function and the proportion of sample firms drawn from each study market is close to the 0.51 correlation observed across the sample firms. This is the methodology that Burt et al. (1994, pp. 359–365) use to draw inferences about the probable strength of the culture effect for American firms throughout the economy.

Here we focus on selecting organizations and markets for generalizable case study. The strength of the culture effect in a research design can be predicted from two things a priori to conducting research: the contingency function, and the relative number of firms selected for study from markets at known points on the contingency function. To study the processes by which a strong

corporate culture (or similar coordination-enhancing forms of organization) enhance performance, select firms for study from effectively competitive markets (e.g. airlines, apparel, motor vehicles, or textiles at the right in Fig. 1), where a strong corporate culture is known to be a competitive advantage. It won't be surprising to find the culture effect, but the goal of the research is to describe the social processes responsible for the effect. At the other extreme, to study processes that substitute for the culture effect on performance, study firms in the markets where producers are effectively coordinated (e.g. communications, beverages, or pharmaceuticals at the left in Fig. 1). There will be no evidence of the culture effect, but the goal of the research is to describe how other social mechanisms such as interpersonal networks coordinate employees in lieu of culture to enhance performance.

Consider a hypothetical case of two students designing research on the performance effects of a strong corporate culture. One selects 10 beverage firms for in-depth case analysis because he worked in the industry and so has good personal contacts there. The other student selects 10 apparel firms for the same reason. Two reasonable and interesting projects with a relatively large number of firms for case analysis.

There is no need to do the research. The first student has selected a market in which producers are effectively coordinated (score of 0.916 for beverages in 1982), so effective competition is low (beverages are to the far left in Figure 1), and a strong corporate culture offers no competitive advantage. This student will find no evidence of higher performance in strong-culture firms, will generalize his results to conclude that there is no culture effect, and later advise client firms against wasting resources on institutionalizing a strong corporate culture. The second student has selected markets at the other extreme of the contingency function. Effective coordination is low within apparel ($\hat{O}$ score of 0.396), so effective competition is high (apparel is to the far right in Fig. 1), and a strong corporate culture is a competitive advantage. This student will find evidence of higher performance in strong-culture firms, will generalize her results to conclude that performance depends on developing a strong corporate culture, and later advise client firms to concentrate on institutionalizing a strong corporate culture. Meta-analysts will later average the significant-results project with the negligible-results project to conclude that evidence is mixed on whether or not a culture effect exists. Within the scope of the individual studies, all three conclusions are reasonable. Nevertheless, all three are wrong; simplistic in their ignorance of the culture effect's contingency function.

How essential is effective competition to this integration? We are merely using census data on markets to integrate case and comparative research. The same method has been used in organization research for decades with

Table 6. Extremes of Residual Producer Coordination (ϕ).

O	ϕ	P	Imports	
0.34	0.00	0.17	11%	means scores ($N = 77$)
Effective Producer Coordination Much Higher than Concentration Implies:				
0.01	0.69	0.77	0%	Real Estate & Rental (71)
0.04	0.63	0.47	4%	OtherAgriculture (farming, 2)
0.03	0.60	0.39	26%	Forestry &Fish (3)
0.03	0.52	0.12	1%	Livestock (1)
0.09	0.44	0.24	0%	Hotels, Personal & Repair Services (not auto, 72)
0.04	0.44	0.24	0%	Business Services (73)
0.08	0.44	0.26	0%	AutomobileRepair& Services (75)
0.20	0.42	0.26	0%	Coal Mining (7)
0.06	0.39	0.13	0%	Eating & Drinking Establishments (74)
0.12	0.39	0.21	0%	Amusements (76)
0.13	0.38	0.27	2%	Wholesale & Retail Trade (69)
0.21	0.38	0.28	4%	Stone and Clay Mining & Quarrying (9)
0.03	0.34	0.11	0%	Medical & Educational Services (77)
0.21	0.31	0.20	2%	Printing & Publishing (26)
0.07	0.26	0.06	0%	Agriculture, Forestry & Fishery Services (4)
Effective Producer Coordination Much Lower than Concentration Implies:				
0.85	-0.21	0.49	5%	Tobacco (15)
0.64	-0.28	0.17	2%	Ordnance & Accessories (13)
0.42	-0.30	0.12	13%	Electrical Industrial Equipment (53)
0.90	-0.31	0.34	1 %	Electric, Gas, Water & Sanitary Services (68)
0.41	-0.32	0.10	10%	Transportation Equipment (not cars/planes/trucks, 61)
0.51	-0.38	0.07	23%	Miscellaneous Electrical Machinery & Supplies (58)
0.35	-0.41	0.06	6%	Screw Machine Products & Stampings (41)
0.67	-0.44	0.14	16%	Household Appliances (54)
0.38	-0.49	0.05	15%	Iron & Steel (37)
0.59	-0.50	0.12	15%	Engines & Turbines (43)
0.64	-0.51	0.10	11 %	Aircraft & Parts (60)
0.73	-0.67	0.07	23%	Nonferrous Metal Ores Mining (6)
0.43	-0.67	0.03	13%	Nonferrous Metals (38)
0.82	-0.71	0.06	33%	Iron & Ferroalloy Ores Mining (5)
0.80	-0.76	0.07	28%	Motor Vehicles & Equipment (cars and trucks, 59)

Note: Columns are concentration (O), residual coordination (ϕ , computed from Eq. 4, $\mu = 0.31$), price-cost margin (P), market share of imports (footnote 5), and market name (input-output sector in parentheses). Markets are listed in descending order of residual coordination, the fifteen highest and the fifteen lowest.

other census measures of market structure such as concentration. *Effective competition is noteworthy here because it is more reliable (Table 4) and accurate (Table 5) than familiar measures of observed market structure, such as producer concentration, in revealing the contingency function integrating case and comparative research.*

More, the contrast between effective competition and observed market structure indicates residual coordination unseen in a market, which is a further guide to siting case analyses. Effective coordination is in some part explicit, observable from market concentration, and in some other part implicit, seen only in the ability of producers to obtain higher profit margins than one would expect from the observed structure of their market. Residual coordination defined in Eq. (4) measures the extent to which producers in a market are more effectively coordinated than they appear to be. Table 6 contains average scores across 1982, 1987, and 1992, ranking markets by residual coordination (detailed transaction data are available on diskette for 1982, 1987 and 1992, so they are the most likely to be analyzed for structure within the aggregate markets).

The 15 markets with the most positive residual-coordination scores are listed at the top of Table 6. These markets are strategic research sites for studying forms of organization in which producers with small shares of their market are organized so as to lower the effective level of competition in their market. Margins tend to be above average (e.g. 77¢ on the dollar in real estate and rentals, 39¢ in forestry and fish), but there are also markets in which margins are below average (e.g. 12¢ in livestock, and 6¢ in agriculture, forestry and fishery services). The common feature is that the margins, high or low, are higher than one would expect from the observed structure of these markets.

Real estate is an ideal-type at the top of the list. The largest real estate firms account for only a small proportion of all American real estate transactions. Concentration is close to zero. To obtain the high profits observed in real estate, producers must be coordinated in ways not apparent from concentration. In fact, real estate markets are organized locally by interpersonal referrals and dominant local brokerages, with city and state regulations an important factor in who gets to sell what (Case, 1965, pp. 2, 141; Fine, 1989, p. 10).

Business services are close to the top of the list. The business services market is a hodge-podge of services; 21% advertising, 13% architects and engineers, 12% lawyers, nine percent management consulting. These are services provided to firms in large part by internal suppliers (staff lawyers, engineers, managers) and local suppliers. Personal ties with clients are critical to success, and such ties are invisible to concentration data. Of course, social order in the market could have other origins (Mizruchi, 1992, Chap. 3). For example, status differentiation seems a likely source of residual coordination since socially

accepted distinctions between high- and low-status producers could help explain profits higher than expected from observed concentration and buyer-supplier constraint (Podolny, 1993; Podolny, Stuart & Hannan, 1996; Stuart, 1998; also note the markets at the top of Table 6 in which status distinctions more familiarly order the market; hotels, personal services, amusements, restaurants, medicine, education, and publishing).

Coal and stone mining are a different kind of example. Transportation costs are substantial and margins are thin. Producers rely on local customers (Rogers, 1986, p. 40, on coal mining; Ampian, 1989, p. 303; Tepordei, 1989, p. 1007, on clay and crushed stone). Government is another factor in residual coordination. Farming and fisheries are at the top of the list in Table 6. There are no dominant large firms in these markets (concentration is only two to five percent). But 26% of forestry and fish products are imported, and 20% of farming is exported. Both transactions are intimately linked with national policy; for example, fishing treaties on the one hand, grain sales on the other. There is also a complicated history of farm subsidies to consider when measuring coordination within American farming (see Browne, 1988; Cunningham et al., 1985, pp. 238–248, respectively on government's role in American farming and fishing).

The 15 markets with the most negative residual-coordination scores are listed at the bottom of Table 6. These markets are strategic research sites for studying how large organizations fail to coordinate. Again, margins tend to be below average (e.g. 7¢ on a dollar of motor-vehicle sales and 5¢ in steel), but there are also markets in which margins are above average (e.g. 34¢ in utilities and 49¢ in tobacco). The common feature is that the margins are lower than expected from observed market structure.

For example, tobacco producers enjoyed a high profit margin, but it is lower than one would expect from a market in which the largest producers so dominate the market. This is a rare instance of government and public opinion eroding producer coordination. Legal action against the tobacco industry weakened the informal arrangements through which competition was managed for so long in the industry (Miles, 1982).

Utilities are a different kind of example in that concentration is such an obviously poor measure of producer coordination. Utilities are usually local monopolies subject to government regulation, so in the absence of concentration data, we set concentration at a high, but not maximum, level of 0.9. The negative residual coordination for utilities in Table 6 shows that the effective level of coordination is lower, even in these years before energy deregulation.

Third, concentration is computed from the market shares of domestic producers, but imports hold a large share in many of these markets – making

the effective level of competition higher than would be expected from concentration data. Among the markets at the bottom of Table 6, auto imports are a familiar mass media story, as are stories about imported household appliances, electrical equipment, steel, and aluminum.

CONCLUSION AND DISCUSSION

It is difficult to overstate the influence of Lawrence and Lorsch's (1967) insight into the contingency of organization (see Scott, 1998, Chap. 4, for historical perspective). The work was a stark rejection of the idea that an optimum organizational form can be determined without understanding the market in which the organization is to operate. Focusing on coordination ties among research, production, and sales functions within their study firms, Lawrence and Lorsch showed that loose-coupling was a competitive advantage in plastics (a market of widely diverse customer demands so the firm that can quickly adapt to new customer needs has a competitive advantage) while tight-coupling was a competitive advantage in metal containers (a market dominated by strong customers able to insist on reliable quality at low price).

Beyond the question of *whether* different forms of organization are an advantage in different markets is the question of *how* their advantage varies across markets. A contingency hypothesis says that the association between two variables X and Y is contingent on the value of a context variable Z. Markets vary in many ways, any of which could be the context variable for a contingency hypothesis in organization theory, but market competition is a critical context variable for any study predicting the performance advantages of alternative organization forms.

Conclusion

This paper has been in three parts about competition as the market context factor in contingency theory. We began with an introduction to the dual structure of markets; the internal structure of producer coordination, vs. the external structure of producer buying and selling with other markets. We used a network model to describe the association between performance and the dual structure of American markets from 1963 through 1992. Summary results in Table 1 and Fig. 3 show how profit margins decrease with the internal market constraint of disorganized producers and the external market constraint of dependence on coordinated suppliers and customers.

Second, we reverse-engineered the model to infer the "effective" coordination of producers, and so the "effective" level of competition in a market. We

asked how coordinated producers must be in order to earn their observed level of profit from their observed pattern of dependence on specific other sectors of the economy. Instead of predicting producer performance from market structure, we used the network model with data on producer performance and external structure (the more reliable and detailed data) to infer internal structure. Producers have an *effective* level (as opposed to an *observed* level) of coordination, and so an effective level of competition, defined by the association between their observed performance and levels of coordination within the markets where they do business.

Third, we discussed the research value of effective competition as the market factor in contingency theory. We demonstrated its value as a reliable market factor (illustrated by its automatic adjustment for the exogenous shock of imports in 1982), its value as an accurate market factor (illustrated by it revealing the contingent value of a strong corporate culture in Kotter & Heskett's, 1992, study), and its value as a market factor integrating case with comparative research.

Our conclusion across the results presented is that the concept of effective competition opens an interesting and demonstrably productive perspective on the market factor in contingency theory.

Assumptions

The assumptions required to measure effective levels of competition seem to us less troubling than the assumptions required to measure competition with concentration data, but the cost of the new assumptions remains uncertain and so warrants brief discussion before we close.

Measurement Error

Market structure-performance models predict performance from market structure, allowing for measurement error in the performance variable. Effective competition assumes instead that producer coordination is measured with error. Of the three structure-performance variables – performance, producer coordination, buyer-supplier constraint – producer coordination is least accurately measured (by concentration ratios with or without ad hoc adjustments). This measurement error in the predictor can be expected to suppress market structure effects on performance and make effect estimates inconsistent. For reasons discussed in the text, it seems safe to assume that producer performance is better measured with input-output table price-cost margins, than producer coordination is measured by concentration data. Thus, effective competition puts the measurement error where measurement is most problematic. This is the point of

our demonstration that effective competition is a more reliable market factor for contingency theory. However, the assumption is more than performance being better measured; it is that performance is measured without any error at all. Error in the performance measure used to compute effective competition will be added to the effective level of coordination among producers. For example, if performance is measured erroneously high for a market, then the effective level of coordination within the market will be erroneously high – which can affect measures of effective coordination in other markets since the coordination reported for any one market depends on the effective level of coordination in its supplier and customer markets.

Complete System

Effective competition scores are computed under the assumption that all significant supplier and customer markets are in the analysis. Effective competition is inferred from producer dependence on buying and selling in specific other sectors of the economy, so it will be measured inaccurately if significant supplier or customer markets for any producers are excluded from the analysis. Concentration ratios do not require this assumption because they are computed from producer market shares without taking into account producer dependence on other sectors of the economy. The assumption of a complete system does not trouble us because the Department of Commerce benchmark input-output tables are a census of business establishments throughout the economy, but it is important to include all sectors in the analysis. Levels of aggregation can be an issue here as well, but selecting a proper level of aggregation is as much an issue for concentration measures of producer coordination (if not the same issue).

Producer Homogeneity

We compute effective competition scores for whole markets, without distinguishing individual producers for their relative exposure or contribution to the competition. The effective aggregate level of competition in a market mixes with aggregate buyer-supplier constraint to determine producer performance in the aggregate. How individual producers combine to define the aggregate is unknown. This is typical of market-level variables in organization research, but the next step in predicting organization performance is to add the market-level competition score to organization-level performance equations as in Table 5. That is where we see the contingent value of a strong corporate culture for individual firms.

NOTES

1. For example, photoengraving and electrotyping are combined in 1982 (detailed sectors 260804 and 260805 in 1977 are sector 260804 in 1982), all three of which are in the aggregate printing and publishing market (sector 26). There are two caveats to our statement about stable market boundaries. (a) The 1967 to 1972 transition involved a change to the computation of dollar flows between markets (secondary products completely excluded), and numerous SIC category revisions that moved commodities from one market into another, though the network pattern of buying and selling for even the most changed markets is similar between 1967 and 1972 (Burt, 1988, describes stability through the 1960s and 1970s). The one exception is that later tables contain a restaurant sector (sector 74) missing in the 1963 and 1967 tables. (b) The 1977 and 1982 tables follow 1972 with little change, but the transition to 1987 and 1992 involved a switch to the 1987 SIC categories, more efficient data processing that distinguished fewer detailed categories (528 detailed production markets in 1982 decrease to 469 in 1987, though most of the lost detail is within construction; 54 detailed construction categories in 1982, 5 in 1987), and more distinctions between aggregate markets. The 77 aggregate production markets in prior tables are 88 in the 1987 and 1992 tables (14 of the aggregate markets in prior tables are combined into 7, and 12 of the 77 are disaggregated into 30 – for an increase to 88, Lawson & Teske, 1994, p. 76). The goal was to combine small, declining markets and disaggregate large, growing markets (e.g. footwear and leather were combined into a single market and business services were divided into four markets each for a specific kind of service). To compare markets over time, we aggregated the 1987 and 1992 detailed data into the 77 aggregate market categories in preceding tables. Thus, we have six observations on each of 76 aggregate markets over time, plus observations on the restaurant market after 1967, for a total of 537 market observations.

2. The eight negative price-cost margins in 1982 (first graph in Fig. 4) are a problem because the log of a negative margin is undefined and effects are estimated by regressing $\ln(P)$ over $\ln(1-O)$, $\ln(C)$, and the controls. We tried truncating the P distribution and adding a constant to move the distribution above zero. Truncating better preserves relative performance over time for the markets with positive margins in 1982 (discussed under Table 4 below) so the eight negative margins in 1982 are re-coded to the 0.02 minimum in the preceding tables. Negative profit is deemed no profit for the purposes of estimating Models II and III.

3. Equation (3) also defines effective constraint at the level of transactions between markets (a predictor in resource dependence theories of organization, see Burt, 1992, pp. 236–252, for review). Re-write Eq. (3) with producer organization brought into the sum of constraint coefficients;

$$A_i = \alpha [\sum_j (k - \hat{O}_i)^r w_{ij} \hat{O}_j]^\gamma = \alpha [\sum_j \hat{c}_{ij}]^\gamma, i \neq j$$

where the expression being summed, $\hat{c}_{ij}$, is the effective constraint on producers in their transactions with supplier-customer market j . This is an attractive form. The effective constraint coefficient is producer dependence on market j (w_{ij}) times relative organization in the two markets (disorganized producers $k-i$ vs. organized

suppliers-customers j). Definition in terms of effective coordination means that the coefficients are adjusted for foreign competitors and the measurement issues that plague concentration data. Further, effective constraint coefficients have the practical advantage of not requiring concentration data.

4. There are seven variances and 21 covariances in Table 3 from which we estimated for the single-factor model seven error variances, the variance of the market factor, and six factor loadings (d_{j2} is set to 1.0 making 1972 structure the reference indicator). The 14 parameters estimated from 28 data in the variance-covariance matrix leave 14 degrees of freedom. Lack of fit generates chi-square statistics of 614.53, 275.03, and 205.27 for concentration, buyer-supplier constraint, and the w_{ij} dependence weights respectively, all of which reject the single-factor model beyond a 0.001 level of confidence. There are six panel-to-panel effects and seven error variances in the simplex model, which leaves 15 degrees of freedom. Lack of fit generates chi-square statistics of 731.43, 104.99, and 65.11 for concentration, buyer-supplier constraint, and the dependence weights respectively, all of which reject the simplex model beyond a 0.001 level of confidence.

5. Market share of imports, F , is dollars of imports divided by the sum of four variables; total commodity output (row sum in the input-output tables), plus sales from inventory, minus exports, plus imports. For example, domestic consumption of motor vehicles in 1982 was \$131,289 million, which was \$110,259 million of production, plus \$1,337 million of sales from inventory, minus \$12,305 million sold in foreign markets, plus \$31,998 million of buses, cars, trucks, and parts imported from foreign markets. Imports held a 24% share of the motor vehicle market in 1982 ($F = 0.244 = 31.998/131.289 = 31.998/[110.259 + 1.337 - 12.305 + 31.998]$). We do not know whether an American firm or a foreign firm produced an imported commodity. An automobile that Honda manufactures in the United States is a domestically produced commodity. A car that Ford manufactures in Europe and sells in the U.S. is an import. Our imports variable is the market share of foreign-made goods, not the market share of foreign firms. Regardless, the results in Table 4 show that imports erode the association between producer profit and concentration.

6. We are reassured by the fact that the 44% performance variance predicted by the 19 market distinctions is similar to the 45% estimated by others with return to assets over time for more detailed market categories (McGahan and Porter, 1997, p. 23, report the 45% in three components; 18.7% associated with four-digit SIC categories plus 31.7% associated with business segments within the categories, minus a 5.5% firm-industry covariance adjustment). However, there is little meaning to the 44% except as motivation for our decision to hold market performance differences constant. Claims regarding the exact portion of performance variance associated with market distinctions have little meaning since markets have no agreed-upon boundaries so the performance variance associated with them can be anything from none and all depending on research design (none – assign all sample firms to one market; all – define markets so narrowly that each sample firm operates in its own market). *Ceteris paribus*, more narrowly defined market categories mean a higher portion of corporate performance variance associated with market distinctions.

7. To compute effective competition scores for Fig. 1, we aggregated detailed input-output categories for 1982 to match the Kotter and Heskett market categories. The result was an aggregate input-output table distinguishing 82 rather than the Department of Commerce's 77 sectors: the food sector was divided into beverages vs. food

processing, transportation was divided into airlines vs. other transport, trade was divided into wholesale vs. retail, and finance was divided into three subsectors (banking, credit agencies, brokers and insurance).

8. We adjusted the position of publishing in Fig. 1, as described in Burt et al. (1994, Appendix), for McGraw-Hill's outlier effect on the association between performance and strong culture. The 0.33 correlation for publishing on the vertical axis in Fig. 1 is -0.04 before the adjustment. We have only done this in Fig. 1 to simplify the illustration. All results we report apart from Fig. 1 are based on the raw data, and the same conclusions about Figure 1 would be reached in this and the next paragraph with the data adjusted for McGraw-Hill or the raw data. The 0.85 correlation in the graph at the top of Fig. 1 is 0.81 for the raw data. The t -tests of 5.0, 0.4, and 1.5 in the next paragraph for effective competition, producer concentration, and buyer-seller constraint respectively are 4.2, 0.0, and 0.8 when based on the raw data.

ACKNOWLEDGMENTS

Portions of this paper were presented in James Coleman's "Mathematical Sociology" workshop at the University of Chicago (1993), an INSEAD conference on "Organizations in Markets" in Fontainebleu, France (1996), and a "Sociology of Strategy" workshop at the University of Chicago (1997). While graduate students at Columbia University, Martin Garguilo assembled the 1982 transaction data and the concentration data on manufacturing, and Shin-Kap Han assembled the 1982 concentration data on nonmanufacturing. While a graduate student at the University of Chicago, Ezra Zuckerman assembled the 1987 concentration data. We are grateful to colleagues for comments improving the text; William Barnett, James Coleman, Michael Hannan, Donald Palmer, Joel Podolny, Jesper Sørensen, Toby Stuart, and Brian Uzzi.

REFERENCES

- Ampain, S. G. (1989). Clays. In: *Minerals Yearbook 1989* (pp. 271–303). Bureau of Mines, U.S. Department of the Interior. Washington, D.C.: Government Printing Office.
- Baker, W. E. (1984). The social structure of a national securities market. *American Journal of Sociology*, 89, 775–811.
- Barker, J. R. (1993). Tightening the iron cage: concertive control in self-managing teams. *Administrative Science Quarterly*, 38, 408–437.
- Barley, S. R., Meyer, G. W., & Gash, D. C. (1988). Cultures of culture: academics, practitioners, and the pragmatics of normative control. *Administrative Science Quarterly*, 33, 24–60.
- Baum, J. A. C., & Mezias, S. J. (1992). Localized competition and organizational failure in the Manhattan hotel industry, 1898–1990. *Administrative Science Quarterly*, 37, 580–604.
- Bigelow, L. S., Carroll, G. R., Seidel, M.-D. L., & Tsai, L. (1997). Legitimation, geographical scale, and organizational density: regional patterns of foundings of American automobile producers, 1885–1981. *Social Science Research*, 26, 377–398.

- Bonacich, P. (1972). Factoring and weighting approaches to status scores and clique identification. *Journal of Mathematical Sociology*, 2, 113–120.
- Bonacich, P. (1987). Power and centrality: a family of measures. *American Journal of Sociology*, 92, 1170–1182.
- Browne, W. P. (1988). *Private Interests, Public Policy, and American Agriculture*. Lawrence, KA: University Press of Kansas.
- Burt, R. S. (1982). *Toward a Structural Theory of Action*. New York: Academic Press.
- Burt, R. S. (1983). *Corporate Profits and Cooptation*. New York: Academic Press.
- Burt, R. S. (1988). The stability of American markets. *American Journal of Sociology*, 94, 356–395.
- Burt, R. S. (1992). *Structural Holes: The Social Structure of Competition*. Cambridge: Harvard University Press.
- Burt, R. S., & Freeman, J. H. (1994). *Market structure constraint in Germany*. Graduate School of Business, University of Chicago.
- Burt, R. S., Gabbay, S. M., Holt, G., & Moran, P. (1994). Contingent organization as a network theory: the culture-performance contingency function. *Acta Sociologica*, 37, 345–370.
- Byron, J. E. (1983). Leather and leather products. In: 1983 U.S. *Industrial Outlook* (pp. 41.1–11). Bureau of Industrial Economics, U.S. Department of Commerce. Washington, D.C.: Government Printing Office.
- Case, F. E. (1965). *Real Estate Brokerage*. Englewood Cliffs, NJ: Prentice-Hall.
- Caves, R. (1992). *American Industry: Structure, Conduct, Performance*. Englewood Cliffs, NJ: Prentice-Hall.
- Coleman, J. S. (1972). Systems of social exchange. *Journal of Mathematical Sociology*, 2, 145–163.
- Coleman, J. S. (1990). *Foundations of Social Theory*. Cambridge: Harvard University Press.
- Collins, N. R., & Preston, L. E. (1969). Price-cost margins and industry structure. *Review of Economics and Statistics*, 51, 271–286.
- Cunningham, S., Dunn, M. R., & Whitmarsh, D. (1985). *Fisheries Economics*. New York: St. Martin's Press.
- DiMaggio, P. J., & Powell, W. W. (1983). The iron cage revisited: institutional isomorphism and collective rationality in organizational fields. *American Sociological Review*, 48, 147–160.
- DiMaggio, P. J., & Powell, W. W. (1991). Introduction. In: W. W. Powell & P. J. DiMaggio (Eds), *The New Institutionalism in Organizational Analysis* (pp. 1–38). Chicago: Chicago University Press.
- Durkheim, E. ([1897] 1951). *Suicide*, John A. Spaulding and George Simpson (trans.). New York: Free Press.
- Fine, T. K. (1989). Innovation, Concentration and the Residential Real Estate Brokerage Industry. Unpublished Ph.D. Dissertation, Department of Economics, Columbia University.
- Finkelstein, S. (1997). Inter-industry merger patterns and resource dependence: a replication and extension of Pfeffer (1972). *Strategic Management Journal*, 18, 787–810.
- Foss, N. J. (1997). *Resources, Firms, and Strategies*. New York: Oxford University Press.
- Galbraith, J. K. (1952). *American Capitalism*. New York: Houghton Mifflin.
- Gersen, J. E. (1999). Markets and corporate conflict: a substitution-cost approach to business litigation. *Law & Social Inquiry*, 24, 589–609.
- Han, S.-K. (1994). Mimetic isomorphism and its effect on the audit services market. *Social Forces*, 73, 637–664.
- Hannan, M. T., & Freeman, J. H. (1989). *Organizational Ecology*. Cambridge: Harvard University Press.

- Hirsch, P. M., & Levin, D. Z. (1999). Umbrella advocates vs. validity police: a life-cycle model. *Organization Science*, 10, 199–212.
- Hubbell, C. H. (1965). An input-output approach to clique identification. *Sociometry*, 28, 377–399.
- Ingram, P., & Roberts, P. W. (2000). Friendships among competitors in the Sydney hotel industry. *American Journal of Sociology*, 106, 387–423.
- Jang, H. (1997). Market Structure, Performance, and Putting-Out in the Korean Economy. Unpublished Ph.D. Dissertation, Department of Sociology, University of Chicago.
- Kalleberg, A. L., Marsden, P. V., Aldrich, H. E., & Cassell, J. W. (1990). Comparing organization sampling frames. *Administrative Science Quarterly*, 35, 658–688.
- Kalleberg, A. L., Knoke, D., Marsden, P. V., & Spaeth, J. L. (Eds) (1996). *Organizations in America*. Thousand Oaks, CA: Sage Publications.
- Klinger, F. L. (1983). Iron ore. In: *Minerals Yearbook 1983* (pp. 447–465). Bureau of Mines, U.S. Department of the Interior. Washington, D.C.: Government Printing Office.
- Kotter, J. P., & Heskett, J. L. (1992). *Corporate Culture and Performance*. New York: Free Press.
- Kwoka, J. E., Jr. (1979). The effect of market share distribution on industry performance. *Review of Economics and Statistics*, 61, 101–109.
- Lawrence, P. R., & Lorsch, J. W. ([1967] 1986). *Organization and Environment*. Boston: Harvard Business School Press.
- Lawson, A. M., & Teske, D. A. (1994). Benchmark input-output accounts for the U. S. economy, 1987. *Survey of Current Business*, 74, 73–115.
- Lie, J. (1997). Sociology of markets. *Annual Review of Sociology*, 23, 341–360.
- Lustgarten, S. H. (1975). The impact of buyer concentration in manufacturing industries. *Review of Economics and Statistics*, 57, 125–132.
- Marsden, P. V. (1981). Introducing influence processes into a system of collective decisions. *American Journal of Sociology*, 86, 1203–1235.
- Marsden, P. V. (1983). Restricted access in networks and models of power. *American Journal of Sociology*, 88, 686–717.
- Martin, J. (1992). *Cultures in Organizations: Three Perspectives*. New York: Oxford University Press.
- McGahan, A. M., & Porter, M. E. (1997). How much does industry matter, really? *Strategic Management Journal*, 18, 15–30.
- Merton, R. K. ([1957] 1968). Continuities in the theory of reference group behavior. In: R. K. Merton (Ed.), *Social Theory and Social Structure* (pp. 335–440). New York: Free Press.
- Meyer, J. W., & Rowan, B. (1977). Institutionalized organizations: formal structures as myth and ceremony. *American Journal of Sociology*, 83, 340–363.
- Miles, R. H. (1982). *Coffin Nails and Corporate Strategies*. Englewood Cliffs, NJ: Prentice-Hall.
- Mizruchi, M. S. (1992). *The Structure of Corporate Political Action*. Cambridge, MA: Harvard University Press.
- Mizruchi, M. S., Mariolis, P., Schwartz, M., & Mintz, B. (1986). Techniques for disaggregating centrality scores in social networks. In: Nancy Tuma (Ed.), *Sociological Methodology* (pp. 26–48). San Francisco, CA: Jossey Bass.
- O'Reilly, C. A., III (1989). Corporations, culture, and commitment: motivation and social control in organizations. *California Management Review*, 31, 9–25.
- Ott, J. S. (1989). *The Organizational Culture Perspective*. Chicago, IL: Dorsey.
- Peltzman, S. (1977) The gains and losses from industrial concentration. *Journal of Law and Economic Review*, 2, 105–140.
- Pfeffer, J. (1997). *New Directions for Organization Theory*. New York: Oxford University Press.
- Pfeffer, J., & Salancik, G. R. (1978). *The External Control of Organizations*. New York: Harper & Row.

- Podolny, J. M. (1993). A status-based model of market competition. *American Journal of Sociology*, 98, 829–872.
- Podolny, J. M., Stuart, T. E., & Hannan, M. T. (1996). Networks, knowledge, and niches: competition in the worldwide semiconductor industry, 1984–1991. *American Journal of Sociology*, 102, 659–689.
- Porter, M. E. (1980). *Competitive Strategy*. New York: Free Press.
- Powell, W. W., Koput, K. W., & Smith-Doer, L. (1996). Interorganizational collaboration and the locus of innovation: networks of learning in biotechnology. *Administrative Science Quarterly*, 41, 116–145.
- Press, W. H., Teukolsky, S. A., Vetterling, W. T., & Flannery, B. P. (1992). *Numerical Recipes in C*. Cambridge: Cambridge University Press.
- Raider, H. J. (1998). Market structure and innovation. *Social Science Research*, 27, 1–21.
- Ravenscraft, D. (1983). Structure-profit relationships at the line of business and industry level. *Review of Economics and Statistics*, 65, 22–31.
- Richards, W., & Seary, A. (2000). Eigen analysis of networks. *Journal of Social Structure*, 1, 1–15.
- Rogers, K. S. (1986). *U.S. Coal Goes Abroad*. New York: Praeger.
- Romo, F. P., & Schwartz, M. (1993). The coming of post-industrial society revisited: manufacturing and the prospects for a service-based economy. In: R. Swedberg (Ed.), *Explorations in Economic Sociology* (pp. 335–373). New York: Russell Sage Foundation.
- Romo, F. P., & Schwartz, M. (1995). The structural embeddedness of business decisions: the migration of manufacturing plants in New York State, 1960–1985. *American Sociological Review*, 60, 874–907.
- Schein, E. H. (1996). Culture: the missing concept in organization studies. *Administrative Science Quarterly*, 41, 229–240.
- Schmalensee, R. (1989). Inter-industry studies of structure and performance. In: R. Schmalensee & R. Willig (Eds), *Handbook of Industrial Organization* (pp. 951–1009). New York: North-Holland.
- Scott, W. R. (1998). *Organizations: Rational, Natural, and Open Systems*. Englewood Cliffs, NJ: Prentice-Hall.
- Simmel, G. ([1922] 1955). *Conflict and the Web of Group Affiliations*. K. H. Wolff & R. Bendix (trans.). New York: Free Press.
- Sorenson, O., & Stuart, T. E. (2001). Syndication networks and the spatial distribution of venture capital investments. *American Journal of Sociology*, 106, 1546–1588.

APPENDIX

We use a Newton-Raphson algorithm to compute effective competition. For the first iteration, we set effective coordination equal to the concentration scores in Table 1 to predict performance. It is convenient to work with natural logarithms of the variables in Eq. (3) with the r ratio taken from Table 4: 1.45 in 1963, 1.35 in 1967, 1.17 in 1972, 1.07 in 1977, 1.00 in 1982 (we use the ratio obtained with imports held constant since the baseline model is so obviously misspecified in 1982), 1.78 in 1987, and 1.02 in 1992. Given values of r , P , O , and the input-output network of market dependencies w_{ij} , the regression intercept α and market structure effect γ can be estimated to define vector A .

Compute adjustments in producer coordination to improve the match between expected and observed performance. Where performance is higher than expected, increase the effective coordination of producers and decrease effective coordination within key supplier and customer markets. The match is perfect for the final scores: $\mathbf{P} = \ln \mathbf{A} - \Delta \ln \mathbf{A}$, where $\mathbf{P}$ is a vector of N price-cost margins. Vector $\Delta \ln \mathbf{A}$ of adjustments to A comes from adjustments to the effective coordination of producers: $\Delta \ln \mathbf{A} = (\partial \ln \mathbf{A} / \partial \hat{\mathbf{O}}) \Delta \hat{\mathbf{O}} = \mathbf{J}(\Delta \hat{\mathbf{O}})$, so:

$$\Delta \hat{\mathbf{O}} = \mathbf{J}^{-1} (\ln \mathbf{A} - \mathbf{P}), \quad (5)$$

where $\mathbf{J}$ is the N by N Jacobian matrix of partial derivatives. Diagonal elements in the Jacobian are ratios of the internal market constraint effect over the effective level of producer competition $(k - \hat{\mathbf{O}}_i)$. In diagonal element (i, i) , the partial derivative of $\ln A_i$ in Eq. (3) with respect to producer coordination $\hat{\mathbf{O}}_i$ is: $r \gamma [\partial \ln (k - \hat{\mathbf{O}}_i) / \partial \hat{\mathbf{O}}_i]$, which equals:

$$r \gamma \left[\frac{1}{(k - \hat{\mathbf{O}}_i)} \left(\frac{\partial (k - \hat{\mathbf{O}}_i)}{\partial \hat{\mathbf{O}}_i} \right) \right],$$

which is:

$$(-r \gamma) / (k - \hat{\mathbf{O}}_i) \quad (6)$$

where the product $r\gamma$ equals β , the effect of producer competition. Eq. (6) is positive reflecting the fact that producer performance increases as producers become better coordinated. Off-diagonals in the Jacobian are ratios of the

external market constraint effect over the current level of supplier-customer constraint on producers. The partial derivative of $\ln A_i$ in Eq. (3) with respect to coordination $\hat{O}_j$ in supplier-customer market j is: $\gamma [\partial \ln (\sum_j w_{ij} \hat{O}_j) / \partial \hat{O}_j]$, which equals:

$$\gamma \left[\frac{1}{(\sum_j w_{ij} \hat{O}_j)} \left(\frac{\partial (\sum_j w_{ij} \hat{O}_j)}{\partial \hat{O}_j} \right) \right],$$

which is:

$$(\gamma w_{ij}) / (\sum_j w_{ij} \hat{O}_j), \quad (7)$$

where $i \neq j$. Equation (7) is negative reflecting the fact that producer performance decreases as key supplier or customer markets become better organized.

Third, adjust and evaluate. At each iteration, there is a new 77 by 77 matrix **J** to be computed, inverted, and inserted in Eq. (5) to define adjustments. Continue until the adjustment $\Delta \hat{O}$ in every market is less than a criterion. We use a criterion of 0.001 to secure three decimal places in each score. If any adjustment is larger than the criterion, go back to step one with adjusted score $\hat{O}_{i+1}$, equal to $\hat{O}_i$ minus $\Delta \hat{O}_i$. Adjusted scores are positive fractions relative to the maximum score ($\hat{O} = \hat{O} / \hat{O}_{\max}$). If any adjusted score is less than zero, the distribution is shifted up so scores remain positive fractions ($\hat{O}_i = (\hat{O} - \hat{O}_{\min}) / (\hat{O}_{\max} - \hat{O}_{\min})$). Adjusted scores are inserted in Eq. (3) to compute an expected price-cost margin for the next iteration. The iterations involve extensive computation, but few are required to reach convergence (38 to 94 iterations, depending on the year and start values).

ISSUES

Iterations were run under various conditions to check solution stability. Four issues are noted. (1) The first issue is convergence. The iterations are Newton-Raphson. As Press et al. (1992, p. 380) state, "This method gives you a very efficient means of converging to a root, if you have a sufficiently good initial guess. It can also spectacularly fail to converge." The problem is that $\Delta \hat{O}$ correctly indicates the direction in which coordination should be adjusted to improve the match between expected and observed performance, but it can overstate the magnitude of adjustment such that scores adjusted from one iteration to the next can weaken the match between expected and observed

performance. The algorithm works well for our initial iterations, but over-adjusts as iterations converge on the final solution. Since the direction of Newton adjustment is correct, Press et al. (1992, pp. 383–385) recommend fractional adjustments. We scale adjustments to the multiple correlation at each iteration. Correlation R between $\mathbf{P}$ and $\mathbf{A}$ is the multiple correlation between observed price-cost margins and the margins expected from market structure. The correlation increases across iterations to a value of one at convergence. At iteration i , compute R_i and adjustments for the next iteration $\Delta \hat{\mathbf{O}}_i$. If R_{i+1} computed with the adjusted scores $\hat{\mathbf{O}}_i - \Delta \hat{\mathbf{O}}_i$ is less than R_i , make a fraction λ of each adjustment and re-compute $\hat{\mathbf{O}}_i - \lambda \Delta \hat{\mathbf{O}}_i$. We use successive fractions of λ equal to $1(1-R)$, $(1-R)^2$, $(1-R)^3$, and so on as needed until R_{i+1} is larger than R_i , or a 0.01 minimum for λ is reached (which guarantees some change in each iteration). The multiple correlation is close to one within a dozen iterations.

(2) The second issue also involves convergence. The algorithm is sensitive to low levels of producer coordination. Figure 5 displays the association between $\ln(A)$ and producer coordination in 1972. The lines are evaluated at the mean level of external market constraint (C). The lines are higher (lower) for lower (higher) levels of external constraint. The dashed line describes the observed market structure effect – producer coordination is measured by the concentration data. Ordinary least squares estimates predicting performance are 1.020 for α and -0.037 for γ , with a 0.313 multiple correlation. The solid line in Fig. 5 describes the effective market structure effect – producer coordination is measured by effective coordination scores at convergence. Ordinary least squares estimates predicting performance are 0.860 for α and -0.066 for γ , with a 1.00 multiple correlation.

The slopes of these lines are the diagonal elements of the Jacobian matrix $\mathbf{J}$ (given in Eq. 6). The dashed line in Fig. 5 describes Eq. (6) in the first iteration. The solid line describes Eq. (6) in the final iteration. The points at which change in $\ln(A)$ is linear with change in producer coordination are marked by the 45 degree slope lines in the graph. Notice that the slope is near zero to the left of the graph, at low levels of producer coordination. At low levels of producer coordination, Eq. (6) is a small fraction ($-\gamma$ is 1.51 times 0.037 for the initial iteration in 1972) divided by a fraction $k - \hat{\mathbf{O}}$ that is close to one (at the left of the graph in Fig. 5). Near-zero diagonal elements in the Jacobian often make the matrix singular (zero determinant) so we can't compute the inverse for Eq. (5). Our solution was to impose a threshold of 1.0 on the diagonal elements in the Jacobian, which means that rates of change are as great or greater than the two 45-degree slope lines in Fig. 5. The exact value of the threshold is arbitrary. We tried lower values down to 0.75 before

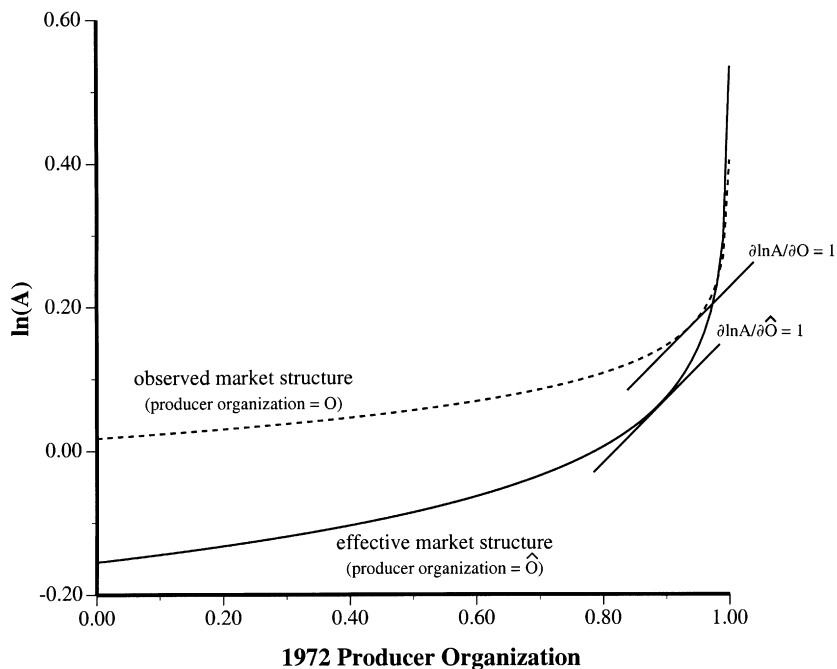


Fig. 5. Diagonal Elements in the Jacobian.

encountering problems, but obtained the same final scores. In different populations, other values could be appropriate. We also tried imposing the threshold until the iterations were in the neighborhood of the solution (multiple correlation greater than 0.9, or 0.99), then allowing the partial to vary over the whole range of its values. This cured the singularity problem, but the algorithm wouldn't converge. The near-zero diagonal elements in the Jacobian create large elements in the inverse of the Jacobian (on the order of 80 to 100), which, *ceteris paribus*, define large adjustments in Eq. (5) for the next iteration. Iterations cycle across the low coordination markets with a large adjustment in one market triggering a large adjustment in another, then back again. So, we imposed the threshold on the diagonal elements in every iteration. The effect is that the final scores are not exact. The 1977, 1982, and 1987 scores are not affected to three decimal places, but the 1972 data are. The multiple correlation should be 1.000, but it is instead 0.9976, and there are small differences between scores obtained with alternative start values (0.003

maximum, 0.0001 mean, discussed below). This is adequate precision for analyzing the 77 aggregate markets, but the potential effect of imposing a threshold should be noted for applications in other study populations. For example, we get less precise results if the diagonals are all forced to equal one (0.993 multiple correlation). The point is that the algorithm is sensitive to low levels of producer concentration in the sense that the Jacobian can be singular or the iterations can fail to converge. Our solution was to impose a threshold on the diagonal elements of the Jacobian, but results should be obtained with alternative thresholds to be sure that the threshold is not interfering with a desired level of precision.

(3) Given convergence, outliers are a third issue. Our effective coordination scores are sensitive to outliers because we give them a fixed range. Performance is especially poor for two markets in 1982: wood containers and nonferrous mining (-0.21 and -0.22 pricecost margins in Fig. 4). The two markets are outliers. The next lowest margin is -0.09 , and there are several markets at that level. The two outliers are at the bottom of the effective coordination distribution, but their scores are so low that all other markets are compressed into the interval between 0.5 and one. Differences between markets are obscured. We ran the iterations for 1982 with negative profit treated as no profit (negative margins set to zero). Effective coordination in nonferrous mining and wood containers remains at the bottom of the distribution, but the other markets are free to vary more widely over the full range of values (0.83 mean and 0.15 standard deviation in effective coordination scores before; 0.68 mean and 0.22 standard deviation after).

(4) The fourth issue is where to start. We use concentration data as initial estimates of effective coordination, but such data are not always available. We tried four alternative start values corresponding to different guesses about producer coordination within markets. We put random error into the observed concentration ratios (random increase or decrease of up to fifty percent of concentration, where errors are from a normal distribution centered on the observed concentration ratio), and we homogenized concentration ratios into a high-low dichotomy (markets above average concentration have a 0.8 start value, others have a 0.2 start value). We also tried random fractions drawn from a uniform distribution (to exaggerate the tails of the concentration distribution), and random fractions drawn from a normal distribution. The alternative start values affect the number of iterations required to reach to convergence, but have little effect on the final scores. Scores obtained with the four alternative start values are correlated 1.0000, with a maximum difference of 0.003 and average difference of 0.001 between scores.

Connection with Eigenvector Network Model

Model I in Table 1 can be written as: $A = a + bO - gC$. Replace C with its definition to get: $A_i = a + bO_i - g \sum_{ij} w_{ij} O_j$ where $i \neq j$. With three markets, there would be three equations:

$$A_1 = a + bO_1 - gw_{12}O_2 - gw_{13}O_3,$$

$$A_2 = a - gw_{21}O_1 + bO_2 - gw_{23}O_3,$$

$$A_3 = a - gw_{31}O_1 - gw_{32}O_2 + bO_3,$$

which is matrix equation:

$$\begin{bmatrix} A_1 - a \\ A_2 - a \\ A_3 - a \end{bmatrix} = \begin{bmatrix} b & -gw_{12} & -gw_{13} \\ -gw_{21} & b & -gw_{23} \\ -gw_{31} & -gw_{32} & b \end{bmatrix} \begin{bmatrix} O_1 \\ O_2 \\ O_3 \end{bmatrix},$$

or for N markets more generally: $\mathbf{A} = \mathbf{W}\mathbf{O}$, where $\mathbf{A}$ is a vector of N structural autonomy scores measuring the relative performance expected in each market ($\mathbf{A} = \{A - a\}$), $\mathbf{O}$ is a vector of N concentration ratios measuring coordination within markets, and $\mathbf{W}$ is an N by N non-symmetric matrix of market dependencies. Expected performance is defined by the coordination of producers within markets filtered through the network of market dependencies.

With the $\mathbf{W}\mathbf{O}$ form of the model explicit, take the market constraint effect out of $\mathbf{W}$ so the matrix equation for the three markets is:

$$\mathbf{A} = g\mathbf{W}\mathbf{O} = g \begin{bmatrix} b/g & -w_{12} & -w_{13} \\ -w_{21} & b/g & -w_{23} \\ -w_{31} & -w_{32} & b/g \end{bmatrix} \begin{bmatrix} O_1 \\ O_2 \\ O_3 \end{bmatrix}.$$

Assume that A and O converge at an equilibrium in which producers are autonomous to the extent that they are not dependent on autonomous suppliers and customers. The equation becomes; $\mathbf{A} = g\mathbf{W}\mathbf{A}$, which is the characteristic equation of the weight matrix $\mathbf{W}$; $0 = \mathbf{W}\mathbf{A} - (1/g)\mathbf{A}$, which can be written in a more familiar form; $0 = (\mathbf{W} - \mu\mathbf{I})\mathbf{A}$, where $\mathbf{I}$ is an identity matrix. Diagonal elements in the $\mathbf{W}$ matrix are the r ratio in Eq. (3). Assume equal effects of internal and external market constraint so ratio r equals 1. Matrix $\mathbf{W}$ is now

defined ($w_{ii} = b/g = r = 1$), μ is the dominant eigenvalue, and autonomy $\mathbf{A}$ is the corresponding right eigenvector.

Eigenvalue μ and eigenvector $\mathbf{A}$ are available with canned computer routines for solving the characteristic equation of square, nonsymmetric matrices. Market i has autonomy at equilibrium to the extent that it doesn't depend on autonomous supplier and customer markets; $a_i = g(a_i - \sum_j w_{ij} a_j)$, where $i \neq j$. The eigenvalue is an adjustment for the level of dependence between markets. Higher w_{ij} generate higher μ . For the network fragment in Fig. 2 and Table 2, the eigenvalue is 1.32 (so g equals 0.76), and the eigenvector elements scaled with respect to the most positive element are; $-1.1, 0.2, -3.1, -1.6, -0.1, -0.1$, and 1.0 . Market C is expected to reach the lowest level of autonomy at equilibrium (it is exclusively and completely dependent on the gray-dot market, row three in Table 2), and the gray-circle market is expected to reach the highest level (it has the lowest dependence on other markets, row seven in Table 2).

This model is a variation on the familiar eigenvector models of network centrality (also discussed as power, prestige or status; Hubbell, 1965; Coleman, 1972, 1990; Bonacich, 1972, 1987; Marsden, 1981, 1983; Burt, 1982, pp. 35–37; Mizuchi et al., 1986; Podolny, 1993; see Richards & Seary, 2000, for review). The variation is two-fold: dependence has a negative value (negative w_{ij} in the W matrix), and equilibrium autonomy is the right rather than the left eigenvector (row vector rather than column). Analogy to the familiar eigenvector model helps link effective coordination to the familiar concept of network centrality, but substantive study of the market networks is better served by the effective coordination model. A nonlinear network model better describes performance differences between markets (Table 1), and reliable performance data are available so there is no need to assume that relative performance evolves to equal relative producer coordination to get the characteristic equation, $\mathbf{A} = g\mathbf{W}\mathbf{A}$.

This Page Intentionally Left Blank

INSTITUTIONAL CHANGE IN REAL-TIME: THE DEVELOPMENT OF EMPLOYEE STOCK OPTIONS FOR GERMAN VENTURE CAPITAL

Jonathan Jaffee and John Freeman

ABSTRACT

We propose studying institutional change and the role of organizations behind it in real-time as the process unfolds and midstream without knowing the success or failure of the project. Our approach is in contrast to most analyses of institutional change that rely on retrospective accounts of successful institutionalization projects. This past methodology runs the risk of 'sampling on the dependent variable,' limiting knowledge of the institutional change process to a narrow slice of successful cases. The context for this new approach to institutional change is the development of 'American-style' employee stock options (ESOPs) in German venture capital contracts from 1997 to 2000. We examine the attempts of 'institutional entrepreneurs' (German law firms) to alter the existing institutional environment to implement American-style ESOPs for their clients (venture capital firms and entrepreneurs). In contrast to past research on institutional change, our analysis reveals a more complex

picture of the process of competition and collective action in leading to change. Our approach highlights the conflicting motives of organizational actors as they battle for and against institutional change.

INTRODUCTION

In the late-1990s, spurred on by the success of the United States's venture capital industry in generating economic growth, Germany attempted to jump on the venture capital bandwagon, using venture capital-backed start-ups to resurrect its flagging economy and enable the country to effectively compete in the *new* global economy. Although the German economy had long ranked among world leaders (e.g. in terms of Gross Domestic Product), German investment in venture capital had traditionally lagged behind most of Europe (in absolute terms, not just per capita), including smaller economic powers, such as Great Britain, Italy, and the Netherlands (Cowie, 1999). Rather than utilizing venture capital and venture capital-backed start-up companies, the German economy had traditionally relied on bank lending, large corporations, and the *mittelstand* (small to mid-sized privately-held businesses) to fuel economic growth (Black & Gilson, 1998; Freeman, Cramer & Jaffee, 1998).

In this period (1997 to 2000), Germany attempted to jump-start its venture capital industry both through highly attractive government subsidies of venture capital-backed investments and the creation of the *Neuer Markt*, a capital market for emerging growth companies, modeled after the United States's NASDAQ equity market. Based in part on government subsidies of venture capital and the creation of the *Neuer Markt*, both the amount and the number of deals invested by German venture capital firms in German-based start-ups companies more than tripled from 1994 to 1998. Moreover, the number of German venture capital firms exploded from 1994 to 1998, going from just one or two venture capital firms to over 20 firms doing 'American-style' venture capital and many companies with corporate venture capital arms (Freeman, Cramer & Jaffee, 1998).

In many ways, Germany had (and continues to have) great potential for a viable venture capital industry, given the country's huge stockpile of both human and financial capital (Pfirrmann, Wupperfeld & Lerner, 1997), essential ingredients for successful venture capital activity (Bygrave & Timmons, 1992). However, prior to the late-1990s, even with the attractive government subsidies of venture capital and the *Neuer Markt*, Germany still lacked many of the necessary institutions and organizations that underlay a successful venture capital industry, including key legal instruments of venture capital (e.g. employee stock options and limited liability partnerships for venture capital

funds) and organizations familiar with venture capital (e.g. venture capital firms and law firms with specialized legal knowledge) (Freeman, Cramer & Jaffee, 1998).

The Development of Employee Stock Options in Germany

Based on the ubiquity of employee stock options (ESOPs) in the United States, there was widespread interest in 'American-style' ESOPs in Germany by the late-1990s (Cowie, 1999; European Venture Capital Association, 1999).¹ From an economic standpoint, employee equity ownership through stock options is considered an excellent device to properly align the employee's and the company's (and its investors's) interests. ESOPs are also often used to lure senior managerial and technical talent from larger, more successful companies, which pay higher salaries, to start-ups, which usually cannot offer these competitive salaries. Because ESOPs are usually taxed at the sale of the stock (after the exercise of the option) and as capital gains, which often represents a lower rate of taxation than an individual's personal income rate, holders of ESOPs often accrue tax benefits (Scholes & Wolfson, 1992). For these reasons, American venture capitalists and entrepreneurs view ESOPs as an essential tool for compensating employees in start-up companies (Bygrave & Timmons, 1992).

Prior to the late-1990s, German venture capitalists faced two major obstacles in attempting to use American-style ESOPs in Germany: the administrative difficulty of creating the stock options and, even once created, unfavorable taxation of the options. Administratively, prior to 1997, Germany's existing corporate stock law did not easily allow companies to create reserves of shares of stock for options, thereby making stock options prohibitively costly (Roschmann, 1999). On the taxation-side, in 1971, the *Bundesfinanzhof* (BFH), Germany's highest court on taxation issues, ruled that certain employee stock options should be taxed at the *execution* of the option and as personal income, and not at the sale of converted stock and as capital gains, as in the United States (Killius, 1997).

In practical terms, the 1971 BFH decision meant that certain employee stock options were taxed at the relatively high personal income rate of approximately 54%. This was in sharp contrast to Germany's virtually non-existent capital gains rates: an individual who owns less than 25% of a capital asset for more than six months pays no capital gains tax in Germany (Killius, 1997). Due to both the administrative difficulties and high taxation of employee stock options, prior to the late-1990s, employee stock options were practically non-existent in Germany (Killius, 1997; Roschmann, 1999). In fact, many German corporate

executives *mistakenly* viewed stock options as ‘illegal’ in Germany (Killius, 1998; Pajunk, 2000; Roschmann, 1999).

By the late-1990s, however, with a bullish environment for German venture capital in place, a host of different organizational actors in Germany, including government officials (at many different levels), members of the judicial system, venture capital trade associations and firms, and private law firms began examining the potential for using American-style stock options in venture capital contracts. This paper examines how these different organizational actors in Germany attempted to create and, at times, forestall the development of American-style ESOPs in Germany.²

In examining the development of American-style ESOPs in Germany, we benefited from being contemporaneous witnesses to this institutional change effort, as it unfolded in real-time without knowing its success or failure. Moreover, by witnessing the development of ESOPs in Germany in real-time, we had a particularly good vista to assess the strengths and weaknesses of past theorizing on institutional change, which has been based on retrospective analyses of successful cases of institutional change.

Institutional Change in Real-Time

Past theories of institutional change have almost exclusively been based on historical case studies of successful institutionalization projects (see, e.g. Powell & DiMaggio, 1991). We believe that this past methodology may substantially limit institutional theory’s ability to offer a robust theory of institutional change because retrospective case studies, especially of successful institutional change, are constrained in their ability to assess the *intent* and *motives* of organizational actors. In the hindsight of successful institutional change, organizational action that looks rational and utility-maximizing might actually have been due to organizational power dynamics, conflict, or random decision making. The converse is also possible: retrospectively, organizational conduct chalked up to power, conflict, or random decisions may turn out to represent rather rational, utility-maximizing organizational conduct in the face of an ambiguous or uncertain institutional environment constraining organizational action. The reliance on successful institutionalization projects as the basis for a theory of institutional change is tantamount to ‘sampling on the dependent variable,’ and potentially creates sample selection biases. As Ingram (1998, p. 273) suggests for future research, “there must be many instances where new organizational forms fail to change institutions, and these failures should be compared to the successes.”

As a counter-balance to this past methodology, we advocate examining important institutionalization projects in real-time, as the process of institutional change unfolds, and midstream, without knowing the success or failure of these projects. To illustrate what a real-time approach to institutional change would look like, we analyze how different actors in Germany attempted to alter the existing institutional environment to implement American-style stock options. Since Germany's development of institutions of venture capital was in midstream during our observation period (1997 to 2000), we were able to avoid the problems of studying successful institutionalization projects retrospectively. A real-time approach allows an unusually good vista for assessing the *intent* and *motives* of organizational actors, which have tended to be understudied in institutional theory (DiMaggio & Powell, 1991). By using a real-time approach to examining institutional change, we can also assess how well past research in the new institutionalism – based on successful institutionalization projects – explains institutional developments that are still incomplete and uncertain.

On a substantive level, this paper also represents one of the few (if only) empirical studies of institutional change involving legal institutions underlying venture capital. Although there is a healthy literature on the effects of the institutional environment on venture capital (e.g. Black & Gilson, 1998; Bygrave & Timmons, 1992; Gompers & Lerner, 1999; Milhaupt, 1997), this literature neglects the crucial process of the creation of such institutions. In fact, this paper represents one of only a handful of empirical studies on the creation of private institutions. As Ingram and Clay (2000, p. 539) point out:

The challenge for institutionalists is to create a richer theory of the origin and change of institutions. One of the biggest barriers to creating a richer theory is a lack of empirical work, particularly on private institutions . . . Empirical studies of change are equally scarce.

As we subsequently point out, the development of legal institutions underlying venture capital is seldom easy and often highly contested.

Structure of the Paper

In the next section of this paper, we describe the research methodology used in our field study of the development of American-style ESOPs in Germany from 1997 to 2000. We then present the main facts of the development of American-style ESOPs in Germany, including the creation of the *Munich Model* of ESOP taxation by a handful of wily Munich law firms. In the following section, the heart of this paper, we interpret the development of American-style ESOPs in German venture capital contracts (e.g. the *Munich Model*) through the lens of past research on institutional change in the new institutionalism. In

order to do this, we iterate between predictions of past theory in the new institutionalism and what transpired in our particular case. The objective here is not only to assess how well the past literature on institutional change explains our case of institutional change, but also to add insight to supplement the past literature. We conclude this paper with an update on the development of American-style ESOPs in Germany, including the *Munich Model*, since we stopped our observation in September 2000.

RESEARCH METHODOLOGY

Our real-time analysis of the development of ESOPs in German venture capital contracts (1997–2000) relies on several different sources of information. Given the desire for detailed information on the motives and intent of organizational actors in the institutional change process, our primary source of data is extensive field notes taken from interviews with the main players in the development of American-style stock options in Germany. These include interviews with many of the major venture capitalists in the key German venture capital firms, lawyers who are involved in the legal structuring of venture capital, lawyers who represent both major and minor firms in the Munich legal scene, the officials at the three main public agencies that oversee government subsidies for seed and start-up investments, and, lastly, key government officials involved in the taxation of stock options, including local Munich and regional Bavarian tax officials.

Our interviews focused on Munich in the German state (*laender*) of Bavaria (*Bayern*), which is the center of venture capital in Germany, akin to California's Silicon Valley and Boston's Route 128 rolled into one. We conducted interviews with key organizational actors on two separate trips. Our first round of interviews occurred in the Summer of 1998 when there was excitement for and growing use of American-style stock options in Germany. We then returned to Munich in the Summer of 1999 for a second round of interviews and learned that the use of American-style ESOPs in Germany had been seriously restricted by government tax officials. We continued to correspond periodically with our informants throughout 2000 and 2001 to ascertain any changes in the development of American-style ESOPs in Germany.

Although our main source of information on the development of American-style stock options comes from interviews with key organizational actors, we supplement our analysis with an extensive literature review on German venture capital and stock options in Germany. We also rely on published information from both the European Venture Capital Association (EVCA) and the German Venture Capital Association (BVK), as well as extensive accounts in

international and German newspapers, magazines, and trade publications concerning the rise of venture capital in Germany. Lastly, we rely on publicly available quantitative data on German venture capital, including detailed information on German venture capital investments, the *Neuer Markt*, and prospectuses of initial public offerings (IPOs) on the *Neuer Markt*. We now turn to what transpired in terms of the development of American-style ESOPs in Germany.

THE DEVELOPMENT OF ESOPs IN GERMANY (1997–2000)

According to our informants, starting in approximately the middle of 1997, with the creation of the *Neuer Markt* and attractive government subsidies of German venture capital in place, German venture capitalists and their lawyers began examining methods to create American-style ESOPs in their venture capital contracts. A major aid to German venture capitalists and their lawyers were the concurrent, sweeping reforms by the German parliament, known as “KontraG,” which modernized Germany’s Stock Corporation Law (Roschmann, 1999). One of the many reforms in *KontraG* was the easing of the administration of creating employee stock options in Germany. As one legal scholar has pointed out, “KontraG . . . has created improved opportunities for establishing [employee] stock option plans . . . The way is now free for the ‘flexible’ use of a conditional capital increase for stock option plans in Germany” (Roschmann, 1999, p. 31).

By early-1998, without the administrative hassles of creating ESOPs, a small but influential group of Munich-based venture capital firms and their lawyers began focusing on the taxation issues of ESOPs. These organizational actors noted the differential tax treatment of personal income (a tax rate of about 54%) and capital gains (often no tax liability) in Germany. Since ESOPs were treated as capital gains in the United States, the German venture capital firms and their lawyers began examining ways to get around the German tax code and 1971 BFH decision, which had classified stock options as generating personal income.

The “Munich Model” of Stock Option Taxation

In early-1998, several prominent Munich law firms, separately, began discussing with local Munich and Bavarian tax authorities the timing and taxation of American-style stock options in venture capital contracts. These Munich lawyers contacted tax officials at two different branches of the Bavarian tax system: the Munich *Finanzamt*, which is the city of Munich’s local tax office, and the *Oberfinanzdirektion (OFD)*, which oversees the *Finanzamt* and includes

Munich and several others parts of Bavaria. When individuals or companies domiciled in Munich have tax questions that need formal interpretation, the *Finanzamt* is the administrative body that issues formal, binding rulings on such questions. The *OFD* offers research and policy direction for the *Finanzamt*. The *Bayern Staats Ministerium der Finanzen*, Bavaria's tax ministry, oversees both the *Finanzamt* and the *OFD*.

Due to Germany's general tax-exempt status of capital gains, these Munich law firms first discussed with the Munich tax authorities the possibility of having their ESOPs taxed at the sale of the stock and as capital gains. Informally, the Munich tax officials rejected such taxation treatment for the stock options, relying on Germany's federal tax code and the 1971 BFH decision, holding that all forms of employee compensation, including stock options, should be treated as personal income and taxed at the execution of the option.

The Munich law firms then decided to push the Munich tax officials to reconsider the issue of the timing of the taxation by claiming that ESOPs in venture capital contracts were assets and thus could be valued *at grant*, thereby triggering a taxable event. In order to value the option at grant, the Munich lawyers reasoned that tax officials could use the well-known *Black-Scholes Formula* for stock option valuation (Black & Scholes, 1973; Hull, 1997). By early-1998, two Munich law firms had convinced Munich tax officials to accept a model of "upfront" (i.e. at grant) taxation of ESOPs, which became known as the *Munich Model* (Haas & Potschann, 1998).

According to Munich tax officials, the *Finanzamt* issued approximately 30 binding tax rulings in favor of the *Munich Model* of upfront ESOP taxation in German venture capital contracts (Potschann, 1999). These favorable rulings by the *Finanzamt* were issued to just a "handful" of law firms on behalf of several venture capital firms (Potschann, 1999).

In practice, the *Munich Model* reduced the taxation of the ESOP to a negligible amount. Specifically, the Munich tax authorities accepted a modified *Black-Scholes Formula* that did not include a term for the option volatility (Haas & Potschann, 1998). The Munich tax authorities accepted the lawyers' argument that it was inappropriate to include the volatility term given the difficulty of assessing the option's volatility based on the lack of a track-record for the success of emerging growth companies in Germany, due to the recent creation of the *Neuer Markt*. However, by not including a term for the option's volatility, an option is worth virtually nothing. The main value of an option involves its volatility, or, more accurately, its upward potential (options that go down in value are simply not executed). By not including a measure for volatility, a favorable ruling from Munich tax officials on the *Munich Model* constituted little or no taxation on the option. In this regard, based on the

Munich Model, the option-holder had to pay only a nominal sum to the Munich tax authorities at the *grant* of the option, which then extinguished all subsequent tax liability.

Use of the Munich Model

Based on our first-round of interviews in the Summer of 1998, there was great variation in the knowledge and use of the *Munich Model*. Most venture capital firms outside of Munich had never heard of the *Munich Model*. In Munich, all of the venture capital firms we interviewed had some knowledge of the *Munich Model*. Moreover, in interviews with the *lawyers* of these venture capitalists, most of them were familiar with the general provisions of the Model and its taxation posture. However, notwithstanding the ubiquity of knowledge of the *Munich Model*, only two law firms were actively filing for binding rulings from the Munich tax authorities on the Model. The Munich tax authorities corroborated this account, stating that they had received phone calls from over 15 law firms concerning the Model, but had received briefs for rulings from only a handful of law firms (Potschann, 1999).

Outside of the few law firms that were championing the *Munich Model* in the Summer of 1998, the majority of lawyers in Munich, including top tax lawyers at several of the city's oldest and most prestigious firms, expressed general disdain and apprehension about the Model. These lawyers objected to the Model on many grounds. These lawyers pointed to the *Munich Model's* supposed lack of legal authority. Many of these lawyers felt that the *Munich Model* violated the 1971 BFH holding: they argued that the option in the *Munich Model* could not be sold on a regulated options markets, violating the BFH's requirement of transferability and tradability of the option. These lawyers also argued that the Model lacked political support because it reduced tax expenditure and did not have the support of a majority of other German *laender*. Lastly, these Munich lawyers argued that there would never be widespread acceptance of the *Munich Model* because of public outcry about allowing entrepreneurs the potential for creating enormous personal wealth, tax-free. The consensus of most lawyers was that a handful of smooth Munich lawyers had tricked the Munich tax authorities through legerdemain into granting favorable tax rulings that only benefited their clients and not Munich or Germany.

Although these law firm-detractors of the *Munich Model* couched their disdain for the Model on legal, political, and social grounds, economically, these law firms had much to lose by the Model's acceptance. According to certain informants (venture capital firms Where is here?), it was not coincidence that the two main law firms pushing for and receiving favorable rulings on the

Munich Model had nearly a duopoly of the venture capital legal work in Munich. In contrast, although many of the lawyers against the *Munich Model* practiced in prestigious and respected Munich firms, their primary clients were large, established German companies with little need for American-style stock options. In fact, several of the law firms against the *Munich Model* had been trying to muscle into the Munich venture capital legal scene, without much success. According to our informants, acceptance of the *Munich Model* would reinforce the control of the two *Munich Model*-proponent law firms as the major players on the venture capital legal scene.

Abeyance of the Munich Model

Based on a second round of interviews in the Summer of 1999, we learned that in December 1998, four months after our first set of interviews (Summer 1998), the Bavarian state tax ministry had temporarily prohibited the *Finanzamt* and the *OFD* from granting favorable rulings on the *Munich Model*. The Bavarian tax ministry was waiting for guidance from the *Lohnsteuerreferentzen der Laender* (Lohnsteuer), the ruling-body of finance ministers for all the German *laender*, at their next quarterly meeting in September 1999. This was the third time that the permissibility of the *Munich Model* had been on the *Lohnsteuer*'s agenda. Each previous time, there had been no vote on the *Munich Model* by the *Lohnsteuer* due to a lack of consensus on how to decide the Model's fate and the Model's relative lack of importance for most of the *laender*.

According to interviews with Munich tax officials, the temporary suspension of binding rulings on the *Munich Model* was based on the same difficulties with the Model that the detractor-law firms had mentioned. In particular, the Bavarian state ministry had received pressure from other German *laender* that disapproved of the Model and did not want Munich and Bavaria offering differential tax treatment on the ESOPs. For example, although Baden-Wittenberg, the German state that includes Stuttgart, had earlier considered accepting the *Munich Model*, Baden-Wittenberg decided against the use of the Model, due in part to pressure from other German states against the Model. Moreover, in most of Germany, where there was little venture capital activity, there was little interest in ESOPs for venture capital. Most German states also viewed the *Munich Model* as leading to little tax revenue.

By September 2000, the status of the *Munich Model* remained ambiguous and uncertain. The *Lohnsteuer* had several times passed on deciding the permissibility of the *Munich Model* (Pajunk, 2000). Similarly, the BFH had several opportunities to clarify the tax treatment of American-style stock options in Germany and had abstained from doing so. Munich lawyers were left with

little guidance on how to structure their American-style stock options for tax purposes. In addition, by this time, one of the two main proponent-law firms of the *Munich Model* had decided to stop using the *Munich Model* in its venture capital contracts until further guidance from German tax officials. In contrast, the other law firm, arguably the chief progenitor of the *Munich Model*, continued to use the Model in many of its venture capital contracts. This law firm did not believe that it put its clients (both entrepreneurs and venture capitalists) at any additional tax risk because the taxation at grant was so trivial. All told, most German venture capital firms and law firms were moving away from the use of any stock option plans in their venture capital contracts.

INSTITUTIONAL CHANGE IN REAL-TIME

In this section, we interpret the development of American-style ESOPs in German venture capital contracts through the lens of past research on institutional change in the new institutionalism. The objective here is to determine how well the past literature, based on retrospective analyses of successful institutionalization projects, explains our case of institutional change, which we witnessed as it evolved without knowing its subsequent success or failure. This approach hopes to uncover some fresh insights to supplement the past literature.

We examine three factors on which the past literature on institutional change has focused:

- (1) Where in a national-system will institutional change occur, including ‘top-down’ efforts, such as federal or state-level legislation or rulings, or ‘bottom-up’ efforts, including the action of grass-roots movements or local actors (Suchman, 1994)?
- (2) Which organizational actors will advocate for institutional change and which ones will resist change (Ingram, 1998)?
- (3) How (if at all) will the contestation over institutional change be resolved, including the resolution of collective action problems among different organizational actors (Holm, 1995; Ingram, 1998; Ingram & Inman, 1996)?

In order to address each of these three questions of institutional change, we iterate between predictions of institutional change based on past theory in the new institutionalism and what transpired in our particular case (Bates, Greif, Levi, Rosenthal & Weingast, 1998).

Where Will Institutional Change Occur?

We start by exploring at what level of Germany's national-system will the development of American-style ESOPs emanate. For our particular case of institutional change, we focus on the role of four different players in Germany's national-system: Germany's federal government, the individual German *laender*, Germany's federal judiciary (the BFH), the venture capital industry, including the venture capital trade associations and private venture capital firms, and, lastly, private law firms.

Theory: The Role of the German Federal Government, Laender, and BFH

Douglass North's (1990, 1998) theory of institutional change would suggest some role for Germany's federal government, *laender*, and federal judiciary (BFH) in the development of American-style ESOPs. According to North, competition between different nation-states and German *laender* should initiate institutional change to make Germany and specific *laender* attractive environments for venture capital. As described earlier, to some degree, at the federal and *laender*-level, Germany did exactly this in terms of creating highly attractive subsidies for venture capital-backed start-up companies and *KontraG*'s sweeping reforms of Germany's old corporate stock law, including substantially easing the administrative difficulties of creating ESOPs.

Although *KontraG* eased the administrative difficulties of creating American-style ESOPs, the question of the taxation of the ESOPs remained. Past research in the new institutionalism, especially in sociology, posit the reluctance of federal and state governments, and even the judiciary, to interpret arcane and ambiguous legal issues, such as the taxation of ESOPs in Germany. As this recent line of research point outs, federal and state governments defer this duty to the judicial system, but the judicial system itself often defers the interpretation of ambiguous legal mandates down to business organizations themselves (Dobbin & Sutton, 1998; Edelman, 1990; Edelman, Uggen & Erlanger, 1999).

Case: The Role of the German Federal Government, Laender, and BFH

In terms of the facts of our case, past research in the neo-institutionalism in sociology correctly predicted that Germany's federal government, state governments, and federal judiciary (the BFH, Germany's highest court on taxation) would all refuse to directly intervene in deciding the proper taxation of ESOPs in Germany. Several times, Germany's parliament was apprised of the uncertainty over the taxation of German ESOPs, and, each time, deferred

on such a decision, pushing it down to lower levels, including the *Lohnsteuer*, the ruling-body of state finance ministers, and the BFH.

The *Lohnsteuer* itself deferred any rulings on the taxation of ESOPs in Germany, even though the question of such taxation was officially on its agenda from 1998 to 2000. At each meeting, the *Lohnsteuer* stated that it would only offer an authoritative ruling if there was consensus on the taxation of American-style ESOPs among the different *laender*. However, since there was little consensus on the taxation of American-style ESOPs in Germany, the *Lohnsteuer* never officially determined the proper taxation posture for American-style ESOPs in Germany. The BFH also deferred making any general pronouncement on the taxation of American-style ESOPs in Germany. One possibility, consistent with the recent research in the neo-institutionalism in sociology (Dobbin and Sutton, 1998; Edelman, 1990; Edelman, Uggen, and Erlanger, 1999), is that the BFH was waiting to see how German venture capital firms and law firms would implement ESOPs in Germany.

Theory: The Role of Munich Venture Capital Firms and Munich Law Firms

Past theory in the new institutionalism, especially in sociology, would suggest a strong role for law firms in the development of legal institutions of venture capital. Suchman's work (1994) on the venture capital industry in Silicon Valley points out that private Silicon Valley law firms, and not the venture capital firms themselves, were the primary developers and diffusers of innovations in venture capital contracts. In this manner, in Silicon Valley, it was *not* the main actors directly affected by venture capital contracts – the entrepreneurs and venture capital firms – but the representatives of these actors – the law firms themselves – that instigated institutional change. There are several reasons why law firms acted as the main developers of institutions of venture capital in Silicon Valley, including the arcane nature of the (legal) issues involved, the special outsider role of law firms, which could minimize transaction costs between entrepreneurs and venture capital firms, and the law firms' desire to socially construct norms of venture capital activity (Suchman, 1994).

Case: The Role of Munich Venture Capital Firms and Munich Law Firms

The role of Munich venture capital firms and law firms in the development of American-style ESOPs in Germany had a remarkable similarity to what occurred in Silicon Valley over twenty years ago. Although informants tell us that the initial interest in American-style ESOPs originated with the German venture capital firms, the active role of German venture capital firms in creating American-style ESOPs soon vanished. Although the European Venture Capital Association (EVCA) and the German Venture Capital Trade Association (BVK)

issued reports endorsing the creation of American-style ESOPs in Germany (Cowie, 1999; European Venture Capital Association, 1999), the German venture capital firms took a back-seat role in the development of American-style ESOPs, ceding its development to their Munich-based law firms.

Instigators and Resistors of Institutional Change

Theory: North's Learning-based Theory of Institutional Change

North's (1990, 1998) learning-based theory of institutional change points to economic competition among organizations as the ignition for institutional change. According to North's theory, competition among organizations first induces the acquisition of new knowledge and skills by these organizations in order to effectively meet organizational goals. However, the existing institutional matrix often constrains organizational action in terms of accomplishing organizational goals that incorporate this new knowledge. In such cases, organizations are compelled to either alter the existing institutional framework to meet these goals or risk being out-competed by organizations that achieve this.

Case: North's Learning-based Theory of Institutional Change

The case of German venture capital firms generally fits well with North's (1990, 1998) learning-based model of institutional change and the role of economic competition in it. At a global and national level, the desire of Munich venture capital firms to use American-style ESOPs represented an important strategy to compete against other nations and within Germany for the best start-up deals and the best employees in terms of start-up companies. Consistent with North's model of economic competition, the proponent-law firms of the *Munich Model* were simply trying to maximize their organizational objectives and accomplish the tasks of their venture capital firm clients (attract the top deals and properly align start-up and venture capital investor interests) and start-up clients (minimize taxes and increase the potential for personal gain). As North's model would predict, the need to accomplish these tasks for their clients led some Munich law firms to attempt institutional change (e.g. the creation of the *Munich Model*).

Organizational Form Conflict over Institutional Change

North's (1990, 1998) model explains some, but not all, of the behavior of organizations involved in the development of stock options in Germany. In particular, although North offers an elegant theory for institutional change, his

theory tends to downplay the role of power and conflict in the institutional change process. North overlooks the paradox that while competition often initiates institutional change, it may also create opposition from other organizations who fear being out-competed by new innovators and thus resist institutional change efforts. Since institutional change may lead to a redistribution of resources within an industry (Knight, 1992), organizations with the most to lose will strongly resist change. Moreover, resisting organizations often represent the existing holders of resources and power, and can offer substantial resistance to institutional change.

Ingram (1998) argues for a modification of North's theory: not all organizations can successfully change to incorporate new knowledge and learning in pursuit of organizational goals. According to Ingram (1998, p. 259), "[r]ather than existing organizations [exploiting new knowledge] . . . it is new organizations representing new organizational forms that exploit new knowledge and become the agents of institutional change." Moreover, according to Ingram (1998, p. 262), existing organizations, which cannot incorporate this new knowledge, resist change:

Organizations that predate the new process or product may not be able to change to incorporate it due to inertia and will resist institutional changes that facilitate its use because such changes help new competitors and probably make it more difficult for existing organizations to use the processes and market the products they rely on.

Past case studies in the new institutionalism offer empirical support for Ingram's (1998) new vs. incumbent organizational form dichotomy for understanding institutional change (see, e.g. Brint & Karabel, 1991; DiMaggio, 1991). Ingram's framework should be especially applicable to our case since our micro-analytic perspective of the institutional change process highlights issues of conflict between and within different organizational forms involved with German venture capital.

Case: American-style Venture Capital Firms as Instigators of Institutional Change

There is much support for Ingram's (1998) new vs. incumbent organizational form dichotomy for understanding the development of American-style ESOPs in Germany. Consistent with Ingram's framework, it was the recently-founded American-style German venture capital firms that were employing the law firms that were the main proponents of the *Munich Model*. These American-style German venture capital firms were also endorsing the use of American-style ESOPs through the publication of "white-papers" in favor of these ESOPs by the European and German Venture Capital Trade Associations (Cowie, 1999;

European Venture Capital Association, 1999). These venture capital firms tended to be either of international origin (having connections with top venture capital firms in the United States and the United Kingdom) or very new entrants to the venture capital scene.

In contrast, the German corporate venture capital firms were mostly uninvolved spectators to institutional change. These corporate venture capital firms were wary of these innovative venture capital practices (e.g. American-style ESOPs), and, for the most part, did not partake or benefit from these institutional innovations. The corporate venture capital firms seldom (if ever) used American-style ESOPs in their venture capital contracts due to the lack of interest of their entrepreneurs (who were or had been employees of their corporate entity). Both the interest of American-style venture capital firms and the lack of interest of corporate venture capital firms in American-style ESOPs in Germany support Ingram's (1998) new vs. old form distinction for institutional change.

Case: New Law Firms as the Main Instigators of Change

In Munich's legal community, those law firms who were among the oldest and most prestigious members of the community were mostly critics of the *Munich Model*. In contrast, the main proponents of the *Munich Model* were recent entrants to the Munich legal scene. For these new entrants, acceptance of the *Munich Model* as the prevailing practice for ESOPs would cement their position as the top legal provider for start-up companies. The critics of the *Munich Model*, those older and more prestigious law firms, risked losing existing clients and not attracting new clients if the *Munich Model* became the dominant method for structuring American-style ESOPs in German venture capital contracts. The success of the *Munich Model* had important consequences for the present and future distribution of resources for all Munich law firms.

In the Appendix of this paper, we list the names (and number of lawyers and ages) of the main corporate law firms in Munich in 1998, the first year of observation in our study, based on the Martindale-Hubbell legal directory (Martindale-Hubbell, 1999).³ Although this list includes the main proponents and detractors of the *Munich Model*, due to issues of confidentiality, we cannot reveal their names. However, of the 81 Munich corporate law firms listed in the Appendix, four of these 81 firms were known proponents of the *Munich Model*, six of these firms were known detractors of the Model, and the remaining 71 firms had no known allegiance for or against the *Munich Model*.

Table 1 reveals several interesting features of Munich's legal population in terms of the size and age of the main proponents and detractors of the *Munich Model*. The average size of Munich law firms without known allegiance to the

Table 1. The Age and Size of Munich Law Firms in 1998.

Munich Law Firms	Average Size	Average Age	Number of Firms
No Allegiance with the Munich Model	8	21.1	71
Detractor of the Munich Model	32.7	32.3	6
Proponent of the Munich Model	14.75	6	4

Source: Martindale-Hubbell Legal Directory (International Edition), 1999.

Munich Model was eight lawyers and the average age was 21 years. For the main proponents of the *Munich Model*, the average size was 14.8 and the average age was six years. For the main detractors of the *Munich Model*, the average size was 32.7 lawyers and the average age was 32.3 years. Table 1 shows that the proponents of the *Munich Model* were substantially smaller and younger than the detractors of the Model. Moreover, the proponents of the *Munich Model* were among the youngest of all Munich corporate law firms, being less than one-third as old as the average firm without any allegiance to the *Munich Model* and less than one-fifth as old as the main detractors of the Model. In contrast, the detractors of the *Munich Model* were among the oldest and largest of all corporate law firms in Munich. These statistics bear out the new vs. old form hypothesis of institutional change.

Resolving Collective Action Problems

We conclude our discussion of institutional change by focusing on the importance of coalition-building and the resolution of collective action problems for successful institutional change (Holm, 1995; Ingram, 1998; Ingram & Inman, 1996). Arguably, the creation of American-style stock options in Germany, such as the development of the *Munich Model*, can be viewed as a problem of collective action.

In the short-run, not all German venture capital firms and their law firms will benefit equally from the improved institutional framework for venture capital activity, based on developments like the *Munich Model*. At least initially, if American-style ESOPs win widespread acceptance due to innovations like the *Munich Model*, those venture capital firms and law firms that are most knowledgeable about these innovations will become known as the key providers of venture capital and the main lawyers that structure venture capital deals. Those venture capital firms and law firms that do not use or approve of such innovations like the *Munich Model* will suffer a proportional loss of business as they lose out on existing and new business to those firms with experience with the new innovative practices.

In the long-run, however, if these venture capital innovations, such as the *Munich Model*, become accepted and help foment overall interest in venture capital in Germany, the *average* German venture capital firm and law firm will benefit from the overall increased level of venture capital work. Since the average venture capital firm and law firm needs to overlook its short-run harm for its eventual long-term benefit, we view the development of American-style ESOPs as a collective action problem.

Theorists in the institutional change literature have posited several different mechanisms to resolve collective action problems. For example, some scholars have emphasized the role of trade associations and other unifying bodies or mechanisms for creating collective action among organizations with disparate interests (see, e.g. Ingram, 1998). In Ingram and Inman (1996)'s study of the development of the U.S. and Canadian parks along-side Niagara Falls, they posit that inter-group rivalry between American and Canadian hoteliers motivated collective action among highly competitive hoteliers on each side of the Falls.

The *substance* of the institutional change effort may also affect the potential for collective action:

Building coalitions is also facilitated by the ability of actors to present interests as altruistic and by the consistency of proposed institutions with a society's basic values For instance, if an institutional entrepreneur were organizing a coalition to establish an institution that defended a nationalistic sentiment (an example of an interest that is both altruistic and consistent with basic values in society), it should be easier for that person to find sympathetic others than if the institution promised only to make the institutional entrepreneur wealthier (Ingram, 1998, p. 270).

Competing firms or strategic groups within an industry often use rhetoric invoking issues of *social welfare* to advocate for their side on an issue involving institutional change (see Shaffer, 1995, for an overview of this literature). For example, Dowell, Swaminathan and Wade (in this volume) argue that the debate over high-definition television was framed in terms of issues of social welfare.

Case: German Venture Capitalists as a Collective Entity

German venture capital firms generally operated as a consistent, collective body in favor of improving the institutional environment for venture capital, including innovations such as the *Munich Model*. This unity was true even though there was potential for conflict between the American-style and corporate venture capital firms. The corporate venture capital firms, based on their affiliation with old, main-line German businesses, had the potential to actively resist institutional change involving the legal structuring of venture capital contracts.

This did not happen for several reasons. First, most of the venture capital firms, both American-style and corporate ones, were members of the European and German venture capital trade associations. These trade associations were and are champions of reforming the German institutional environment for more American-style venture capital practices, including American-style ESOPs (Cowie, 1999; European Venture Capital Association, 1999). These trade associations helped maintain a collective and unified view of venture capital practices within Germany. A second explanation for the lack of resistance by corporate venture capital firms is that because they were less prestigious and influential than the American-style ones, the former tended to quietly follow the trade associations' directives rather than resist them.

A third explanation for why the corporate venture capital firms did not resist institutional change involves the general resource munificence for venture capital deals in Germany over this time. In Germany, due to extraordinary government subsidies of venture capital investment, the corporate venture capital firms were not having trouble finding money to invest in deals. The American-style venture capital firms often complained that the corporate venture capital firms were encroaching on their territory and driving up the price of venture capital deals. Moreover, because the bread-and-butter deals for the corporate venture capital firms were spin-offs of their corporate entities, their competition with American-style venture capital firms was limited.

Case: Conflict among German Law Firms

In contrast to the collective nature of German venture capital firms, there was severe in-fighting among Munich law firms over the propriety of the *Munich Model*. Unlike the German venture capital firms, German law firms did not belong to such a unified and cohesive trade association as the venture capital firms – Germany's legal trade (bar) associations were fractured and contentious (Killius, 1998; Pajunk, 2000). Moreover, the detractors of the *Munich Model* framed their disdain of the Model in lofty terms involving the Model's threat to Germany's social welfare. Although the proponents of the *Munich Model* countered that the Model would make Germany economically competitive with other nations for venture capital activity, the proponent-law firms were hard-pressed to overcome the criticism that the Model mainly led to individual wealth-creation. Given the real threat to the future livelihood of the old, prestigious Munich law firms, and the lack of an altruistic basis for the *Munich Model* in the eyes of most Germans, the chance of any collective vision among Munich law firms involving the Model was remote.

CONCLUSION

Past analyses of institutional change have relied on retrospective accounts of successful institutionalization projects to understand the institutional change process. In this paper, we focused on an institutionalization project – the development of American-style ESOPs in German venture capital contracts from 1997 to 2000 – that we contemporaneously witnessed, without knowing its success or failure. We examined three factors on which the past literature on institutional change has focused: where in a national-system will institutional change occur; which organizations will advocate for institutional change and which ones will resist change; and, how (if at all) will the contestation over institutional change be resolved, including the resolution of collective action problems among different organizational actors. We used our real-time approach to determine how well the past literature explains our case of institutional change and to uncover new insights to supplement the past literature.

The past literature would predict many aspects of the development of American-style stock options in Germany. For example, the past literature would correctly identify the Munich law firms as the center of institutional change efforts. Moreover, the past literature would also predict the reluctance of other organizational actors, including venture capital firms, Germany's federal and state governments, and even the German judiciary, in playing a major role in the development of ESOPs in Germany.

In comparison to past theory in the new institutionalism, our approach reveals a more complex and nuanced assessment of competition, firm conflict, and the resolution of issues of collective action. For example, the past literature would underestimate the double-edged sword of competition for institutional change. Competition may ignite the need for institutional change, but competition may also prevent institutional change from happening when change threatens the existing distribution of resources among organizations. In the German venture capital context, although competition among Munich law firms initiated the development of American-style employee stock options (e.g. the *Munich Model*), this institutional change effort also threatened the position of important, prestigious Munich law firms, which then strongly resisted the use of these innovative venture capital practices.

The past literature would also understate the complicated organizational dynamics underlying institutional change. In some cases, as with German venture capital firms, there was little conflict between the new entrants (American-style venture capital firms) and more incumbent firms (corporate venture capital firms). In other cases, as with the Munich law firms, there was substantial conflict between more recent entrants (the young law firms in favor

of the *Munich Model*) and more established firms (the older law firms that derided the use of the *Munich Model*). As our real-time analysis of institutional change points out, conflict between new and incumbent firms (and organizational forms) relates to the level of environmental resource munificence for these organizations, the directness of competition between firms (forms), and the strength of a unifying collective body, such as a trade association, in overcoming collective action problems. Researchers of institutional change should be vigilant in not glossing over intra-form, or inter-form, conflict.

The role of organizational actors, especially at lower-levels of analysis, in creating institutional change may be further heightened when the change effort involves arcane legal issues, which are susceptible to multiple interpretations. Edelman and colleagues (1999) have recently argued for the “endogenous” interpretation of legal mandates: in interpreting ambiguous legal mandates, courts may look at current business practice for the proper legal interpretation. In our case, the lack of an accepted business practice for implementing American-style stock options in German venture capital contracts, due to severe infighting among Munich law firms over the propriety of the *Munich Model*, may well have discouraged German and Munich tax authorities and courts from taking a strong stand on this issue.

Similarly, institution-building may be eased when the institutional change effort can be framed as benefiting the public good (Dowell, Swaminathan & Wade, in this volume; Ingram, 1998). In our case, since most Germans viewed American-style ESOPs as either ‘illegal’ or, for those in the know, as a vehicle for substantial individual wealth-creation, the *Munich Model* faced an uphill battle for its widespread acceptance and usage.

What Happened to the Munich Model, 2000 to 2002?

By September 2000, only one Munich law firm continued to use the *Munich Model*. By this time, there was practically no support for the Model from most of the German venture capital firms, the other Munich law firms, officials at the German federal and state governments, as well as the German judiciary (BFH). Given the *Munich Model’s* ambiguous legal mandate, the Model’s fate seemed perilously tied to its lack of acceptance by a majority of law firms practicing in Munich. Outside of some exogenous change in the German federal tax code or corporate stock law, or a substantial change in attitude toward individual wealth-creation in Germany, the *Munich Model* seemed “kaput,” as one of our informants told us. Since early-2001, we are told that the *Munich Model* has vanished from the German venture capital scene.

In the end, we are left with a beguiling and inherently unanswerable question about the *Munich Model*: is there any reason to believe that the *Munich Model* could have survived and even diffused if the old-line Munich law firms had united behind it? In some ways, the answer would seem to be no, as the notion of substantial individual wealth-creation based on the *Munich Model* would seem repugnant to most Germans, severely limiting the potential diffusion of the Model. However, it is important to remember that for a certain period of time in the late-1990s German tax officials formally issued rulings permitting the use of the *Munich Model* in German venture capital contracts. In fact, the prospectuses of some of the hottest IPOs on the *Neuer Markt* during this period (1997–2000), including Intershop Communications and Ixos Software, refer to the *Munich Model* of upfront ESOP taxation (Intershop Communications, 1998; Ixos Software, 1998).⁴

In this period, Germany and its individual *laender* were scrambling to jump on the venture capital bandwagon and not be left for dead in the new economic order. Arguably, elements of the *Munich Model* could have been retained. For example, German tax officials could have tweaked the *Munich Model* by including a term for option volatility (thereby increasing the value of the option and its taxation) or taxing the option at the sale of the shares, like in the United States, but at a special capital gains rate, which could have created some tax liability (and tax revenues).

The purpose of our real-time methodology has been to elicit these types of questions about how institutional change is created or forestalled. By focusing on successful institutionalization projects retrospectively, institutional theorists capture only a small slice of what actually occurs during the creation of institutions. For example, as Clay and Strauss's treatment of the taxation of internet commerce (in this volume) and Dowell, Swaminathan, and Wade's discussion of the development of high-definition television (in this volume) illustrates, institutional change is an inherently on-going social dynamic. Institutional theorists would do well to not only pay attention to past successful institutionalization projects, but also examine the fascinating ones that are presently evolving before our very eyes.

NOTE

1. ESOP is the well-known acronym for an employee stock option plan. An ESOP grants an employee the right or 'option' to purchase equity shares in his or her company. This option to purchase shares is either granted to the employee at the time of initial hiring or it accrues and vests over the course of the employee's tenure in the company. There are many different types of ESOPs. In this paper, we focus on Germany's attempt to imitate those ESOPs found in U.S. venture capital contracts, which we refer to as

'American-style' ESOPs. These ESOPs tend to be classified as employee-incentive stock options, which often offer a favorable tax rate for option-holders (see Scholes & Wolfson, 1992).

2. Although we focus on the development of American-style ESOPs in Germany, we also discuss other related changes affecting the German economy and the German venture capital industry during this period. For example, in the late-1990s, the German parliament passed major reforms of the country's rules for corporate governance, the well-known "KontraG" reforms (Roschmann, 1999). *KontraG* eased considerably the administrative difficulties of creating ESOPs in Germany. Moreover, during this period, the German government considered changing its federal tax laws (though it never did) to improve the institutional environment for venture capital activity (i.e. lower tax rates) (see Freeman, Cramer & Jaffee, 1998). We discuss these other reforms when relevant to the development of German ESOPs.

3. The Martindale-Hubbell legal directory is considered the 'yellow-pages' of lawyers. Most academic research on law firms uses the Martindale-Hubbell legal directory because of its ubiquitous use by lawyers and high quality and comprehensive nature. Due to the publishing process, the Martindale-Hubbell of 1999 records the names of law firms in the previous year of observation, 1998.

4. The prospectus for Intershop's initial public offering refers to their stock option plan as "[t]he European Stock Option Plan, which corresponds to the 'Munich Model' (*Muenchner Modell*) of Stock Option Plans for young growth enterprises" (Intershop, p. 23) (emphasis in original). The Intershop prospectus then details that the ESOP is taxed at grant, and this taxation status has been endorsed by German tax officials. The prospectus for Ixos's initial public offering does not explicitly refer to the *Munich Model*, but was written by one of the proponent-law firms of the Model and has terms that create upfront ESOP taxation. Moreover, the Ixos prospectus (Ixos, p. 49) refers to special taxation status based on Munich tax rulings: "Pursuant to a ruling by the Munich tax authorities in October 1997 (*Lohnsteueranrufungsauskunft*) . . . optionees must pay income tax at the time a German Option is granted."

ACKNOWLEDGMENTS

We would like to thank the editors of this volume, Paul Ingram and Brian Silverman, for helpful comments on an earlier draft. We also appreciate the comments of the participants at the "New Institutionalism in Strategic Management" conference held at the Columbia Business School. Paul Ingram, Sigurd Strack, and Nick Ziegler also provided wise counsel on the design and development of this project. This research was supported by the Lester Center for Entrepreneurship & Innovation, Haas School of Business, University of California, Berkeley and Dr. Hasso Plattner of SAP, AG.

REFERENCES

- Bates, R. H., Greif, A., Levi, M., Rosenthal, J.-L., & Weingast, B. R. (1998). *Analytic narratives*. Princeton, NJ: Princeton University Press.

- Black, B. S., & Gilson, R. J. (1998). Venture capital and the structure of capital markets: Banks vs. stock market. *Journal of Financial Economics*, 47, 243–277.
- Black, F., & Scholes, M. (1973). The pricing of options and corporate liabilities. *Journal of Political Economy*, 81, 637–654.
- Brint, S., & Karabel, J. (1991). Institutional origins and transformations: The case of American community college. In: W. W. Powell & P. J. DiMaggio (Eds), *The New Institutionalism in Organizational Analysis* (pp. 337–360). Chicago, IL: The University of Chicago Press.
- Bygrave, W. D., & Timmons, J. A. (1992). *Venture capital at the crossroads*. Boston, MA: Harvard Business School Press.
- Clay, K., & Strauss, R. (2002). Institutional barriers to electronic commerce: An historical perspective. In: P. Ingram & B. S. Silverman (Eds), *Advances in Strategic Management* (Vol. 19).
- Cowie, H. (1999). *Venture capital in Europe*. London: Federal Trust Press.
- DiMaggio, P. J. (1991). Constructing an organizational field as a professional project: U.S. art museums, 1920–1940. In: W. W. Powell & P. J. DiMaggio (Eds), *The New Institutionalism in Organizational Analysis* (pp. 267–292). Chicago, IL: The University of Chicago Press.
- DiMaggio, P. J., & Powell, W. W. (1991). Introduction. In: W. W. Powell & P. J. DiMaggio (Eds), *The New Institutionalism in Organizational Analysis* (pp. 1–40). Chicago, IL: The University of Chicago Press.
- Dobbin, F., & Sutton, J. R. (1998). The strength of a weak state: The rights revolution and the rise of human resources management divisions. *American Journal of Sociology*, 104, 441–476.
- Dowell, G., Swaminathan, A., & Wade, J. (2002). Pretty pictures and ugly scenes: Political and technological maneuvers in high definition television. In: P. Ingram & B. S. Silverman (Eds), *Advances in Strategic Management* (Vol. 19, pp. 97–133). Elsevier Science.
- Edelman, L. B. (1990). Legal ambiguity and symbolic structures: Organizational mediation of civil rights law. *American Journal of Sociology*, 97, 1531–1576.
- Edelman, L. B., Uggen, C., & Erlanger, H. S. (1999). The endogeneity of legal regulation: Grievance procedures as rational myth. *American Journal of Sociology*, 105, 406–454.
- European Venture Capital Association. (1999). *Special report: Stock options*. Brussels: S. J. Berwin & Co.
- Freeman, J., Cramer, C., & Jaffee, J. (1998). *Venture capital and growth businesses in Germany*. Working paper, Hass School of Business, University of California, Berkeley.
- Gompers, P. A., & Lerner, J. (1999). *What drives venture capital fundraising?* Working paper, Number W6906, National Bureau of Economic Research.
- Haas, W., & Potschann, H. (1998). Aussage von aktienoptionen an arbeitnehmer und deren lohnsteuerliche behandlung. *Steuerrecht*, 43, 2138–2141.
- Holm, P. (1995). The Dynamics of institutionalization: Transformation processes in Norwegian fisheries. *Administrative Science Quarterly*, 40, 398–422.
- Hull, J. (1997). *Introduction to futures and option markets* (3rd ed). Englewood Cliffs, NJ: Prentice Hall.
- Ingram, P. (1998). Changing the rules: Interests, organizations, and institutional change in the U.S. hospitality industry. In: M. C. Brinton & V. Nee (Eds), *The New Institutionalism in Sociology* (pp. 258–276). New York: The Russell Sage Foundation.
- Ingram, P., & Clay, K. (2000). The choice-within-constraints new institutionalism and implications for sociology. In: *Annual Review of Sociology* (pp. 525–546).
- Ingram P., & Inman, C. (1996). Institutions, intergroup competition, and the evolution of hotel populations around Niagara Falls. *Administrative Science Quarterly*, 41, 629–958.

- Intershop Communications, AG. (1998). *Prospectus of initial public offering* (July 15).
- Ixos Software, AG. (1998). *Prospectus of initial public offering* (October 6).
- Killius, J. (1997). *Business operations in Germany*. Washington D.C.: Tax Management, Inc.
- Killius, J. (1998). Personal correspondence.
- Knight, J. (1992). *Institutions and social conflict*. New York: Cambridge University Press.
- Martindale-Hubbell. (1999). *Legal directory (international edition)*. Summit, NJ: Martindale-Hubbell.
- Milhaupt, C. J. (1997). The market for innovation in the United States and Japan: Venture capital and the comparative governance debate. *Northwestern University Law Review*, 91, 865–898.
- North, D. C. (1990). *Institutions, institutional change and economic performance*. Cambridge: Cambridge University Press.
- North, D. C. (1998). Economic performance through time. In: M. C. Brinton & V. Nee (Eds), *The New Institutionalism in Sociology* (pp. 247–257). New York: The Russell Sage Foundation.
- Pajunk, A. (2000). Personal correspondence.
- Pfirmsmann, O., Wupperfeld, U., & Lerner, J. (1997). *Venture capital and the new technology based firms*. Heidelberg: Physica-Verlag.
- Potschmann, H. (1999). Personal correspondence.
- Powell, W. W., & DiMaggio, P. J. (1991). *The new institutionalism in organizational analysis*. Chicago, IL: The University of Chicago Press.
- Roschmann, H. (1999). Stock option plans in Germany. *International Law Practicum*, 12, 29–36.
- Scholes, M. S., & Wolfson, M. A. (1992). *Taxes and business strategy: A planning approach*. Englewood Cliffs, NJ: Prentice Hall.
- Shaffer, B. (1995). Firm-level responses to government regulation: Theoretical and research approaches. *Journal of Management*, 21, 493–512.
- Suchman, M. C. (1994). On advice of counsel: Law firms and venture capital funds as information intermediaries in the structuration of Silicon Valley. Ph.D. dissertation, Stanford University.

APPENDIX

Name of Law Firm	No. of Lawyers	Age of Law Firm (years)
Beiten, Burkhardt	59	32
Norr, Stiefenhofer	56	48
Schmidt-Sibeth, Heisse	38	24
Grunecker, Kinkeldey	34	.
Bardehle, Pagenberg	27	20
Haarman, Hemmelrath	26	11
Oppenhoff & Radler	26	26
von Boetticher, Hasse	23	.
Punder, Volhard	23	3
Schwarz, Kurtze	23	.
Wessing & Berenberg-Gossler	21	.
Fiedler & Forster	19	.
Heuking, Kuhn	17	.
Boesebeck & Droste	17	.
Vossius & Partner	15	37
Gaedertz	15	.
Frohwitter	13	1
Seufert	13	.
Raupach & Wollert-Elmendorff	13	.
Schlawien, Naab	13	.
Hoffmann, Eitle	13	106
Weinberger, Sottung	13	.
Hasche, Eschenlohr	12	.
Spitzweg & Partner	12	27
Stock, Strohm	12	18
Zirngibl, Langwieser	11	.
Heiss & Partners	11	.
Burger, Bohl	10	.
Weitnauer	9	.
Seelig, Preu	9	.
Gassner, Stockmann	9	.
Klaka Firm	8	.
Andersen, Freihalter	7	3
Buchholtz, Kisling	7	.
Abrell, Wendler	7	23
Fassbender, Schmitt-Walter	7	28

Rossner Firm	7	.
Doser, Amereller	7	6
Werner, Luger	7	.
Krauss, Schoepe	7	.
Lorenz, Seidler	7	36
Weiss, Walter	7	.
Viering, Jentschura	7	.
Scheele & Partner	6	.
Knorr Firm	6	7
Wirsing, Hass	6	.
Kador & Partners	6	24
Schmid, von Buttlar	6	.
Kloyer, Barthmes	6	2
Dissmann, Orth	5	.
Peters, Fleschutz	5	.
Graf, von Westphalen	5	.
Lenz & Kollegen	4	.
Veltins Firm	4	.
Maiwald	4	5
de Witt Oppler	4	23
Pollarth & Partners	3	4
Bernet & Böhner	3	.
Groll, Gross	3	.
Eisenfuhr, Speiser	3	.
Pracht, Rieg	13	12
Ehlers, Ehlers	3	.
Pollizien, Martens	3	.
Sydow, Rudolph	3	.
Roth, Roth	3	.
Fahr-Becker, Jackubowicz	3	8
Engelhard, Busch	3	.
Dr. Koppe & Partner	28	3
Menold, Herrlinger	2	.
Haseltine Lake	2	.
Thorn & Lunder	2	.
Johannes Fiala	2	.
Peter Czirnich	2	.
Dr. Grill	2	27
Nerz & Todt-Hang	2	.
Bureau D. A. Casalonga-Josse	1	20

Walter, Conston	1	.
Ladas & Parry	1	.
Peter Muller	1	.
Johannes Borries	1	.
Gleiss & Grosse	1	11

Source: Martindale-Hubbell Legal Directory (International Edition), 1999 Missing ages are denoted by ‘.’

INSTITUTIONAL BARRIERS TO ELECTRONIC COMMERCE: AN HISTORICAL PERSPECTIVE

Karen Clay and Robert P. Strauss

ABSTRACT

Although electronic commerce is currently a relatively small fraction of overall sales, the dollar amounts are significant and growing rapidly. Future growth, however, is likely to be limited by two factors – technical barriers and issues of trust and risk. Technical barriers such as delivery, bandwidth, and standardization are already beginning to erode. Problems of trust and risk require as yet undeveloped institutional solutions. The paper explores the possible form of these institutions by drawing lessons from institutions that emerged historically to address opportunism in remote commerce. Once such institutions emerge, remote commerce will begin to have real tax implications for states. The paper describes the institutional changes that will have to occur to address the tax shortfall once it becomes fiscally and therefore politically noticeable.

1. INTRODUCTION

Remarkably little has been written on institutional emergence, institutional change, and the relationship between private and public institutions (Ingram & Clay, 2000). The reasons lie in part with the stability of Western economies,

especially post-World War II. Institutions tend to emerge or undergo significant change during times of economic or political change. Stability has limited the need for, and thus observed change.

During the 1990s, the rise of the Internet initiated a period of significant economic change. The most dramatic changes occurred in capital markets with the rise of the Internet bubble. (See Jaffee & Freeman in this issue on German venture capital markets.) Less dramatic, but more economically enduring changes occurred in the markets for goods and services as consumers and businesses began to transact using the Internet. The dollar amounts involved are significant and growing rapidly (Census, 2001; Emarketer, 2001).

Future growth, however, is likely to be limited by two factors – technical barriers and the relative absence of institutions to control opportunism. Technical barriers include bandwidth and delivery for business to consumer (B2C) electronic commerce and a lack of common standards across vendors and exchanges for business to business (B2B) electronic commerce. Opportunism is an issue on several fronts. Sellers may deliver goods and services of lower quality than were contracted for or not deliver the goods and services at all. They may also misuse or accidentally permit misuse of information related to the sale. Buyers may not pay by engaging in credit card fraud or by accepting credit and then not repaying it.

Problems of opportunism and the relative absence of institutions to address opportunism suggest that we are likely to see new institutions emerge. To understand the types of institutions that are likely to emerge, we draw on results from game theory and a case study of the early days of mail order. The analysis suggests that private, reputation-based institutions are likely to emerge. There may be a role for the State, however, in ensuring that the information necessary for such institutions to operate is available.

If, or more likely when, such private institutions emerge, the resulting increase in electronic commerce will have significant tax implications for the State, specifically for state and local government. The public institutions that govern budgets will have to change. The only question is how these changes will be manifest. We outline alternative institutional approaches.

In sum, we examine the emergence of private institutions, the resulting change in public institutions, and the relationship between private and public institutions. Unlike most studies, we are not doing an *ex post* analysis of these changes. Rather, we are making admittedly ambitious *ex ante* predictions based on a detailed understanding of the current environment and the types of institutions that emerged historically to address similar issues. The paper is organized as follows. Section 2 examines the underlying problems of opportunism in electronic commerce. Sections 3 and 4 examine the institutions

that addressed opportunism in remote commerce in the pre-Internet era. Section 5 examines current barriers to the expansion of electronic commerce and some institutional solutions. Section 6 discusses the tax problems that will arise from further expansion of electronic commerce and the institutional changes that will have to occur to address the problem. Section 7 concludes.

2. OPPORTUNISM

In most types of transactions, either the buyer or the seller can act opportunistically. Buyers can act opportunistically by not paying for goods or services that they receive. In some contexts, they may also be able to engage in other behavior that imposes costs on sellers such as canceling orders. Sellers can act opportunistically by sending the wrong quality or quantity of goods, not sending the goods in the specified time frame, or by charging a higher price than initially agreed upon. Sellers can also act opportunistically by misusing or permitting misuse of transaction-related data such as personal or credit information.¹

In the absence of institutions that mitigate these risks of opportunism, we would not expect buyers and sellers to engage in remote commerce at all. That is, a buyer and seller who only plan to interact once in an environment with no legal infrastructure or other institutions would have no incentive to transact. If the opportunity arose, each would attempt to cheat the other.

Two different types of institutions – private order institutions and public legal institutions – can mitigate buyers' and sellers' incentives to act opportunistically. Private order institutions create incentives for participants to behave honestly by linking current behavior to future payoffs. For instance, if a buyer and a seller plan to transact repeatedly and both players derive value from the transactions, the threat of discontinuing the relationship may be sufficient to guarantee that both parties behave honestly. Alternatively, if a group of buyers plan to transact with a group of sellers repeatedly and all players derive value from the transactions, the threat of group punishment of a cheater may be sufficient to guarantee that all parties behave honestly. Note that multilateral repeated interaction may be sufficient to guarantee honesty, even if individual buyers and sellers do not interact sufficiently often to guarantee honesty. In the limit, as long as players can observe one another's past behavior it may be possible to support cooperation even if a buyer and seller never plan to interact again. It is important to note that under private order institutions the participants in trade, not the state, impose sanctions on cheaters.

Under public legal institutions, the State establishes the rules governing behavior, determines whether violations of the rules have occurred through the courts, and if necessary enforces the courts' decisions using law enforcement personnel. Using public legal institutions has a cost, so buyers and sellers who have been cheated may not always bother to bring a case. Alternatively, the parties may choose to settle 'in the shadow of the law' before the case is filed, before the case reaches the court, or before the court reaches a decision. One important difference between public legal institutions and private order institutions is that public legal institutions can at least in theory support single transactions between buyers and sellers who have no knowledge of one another's past behavior.

In practice, studies indicate that both types of institutions are often important in the contemporary business environment. For instance, Macauley's landmark study (Macauley, 1963) indicated that businessmen were often reluctant to use the legal system. Their preference for private dispute resolution arose because of informal norms in the business community specifying that disputes should be resolved privately. The greater flexibility and lower cost of private dispute resolution led businessmen to prefer dealing with other businessmen who followed that norm, creating an incentive for individuals to adhere to the norm. More recently, Thomas Palay's study of railroad carriers and freight shippers, Robert Ellickson's study of cattle ranchers, and Lisa Bernstein's study of New York Diamond dealers all support Macauley's main themes. For small, fairly homogenous groups, private order institutions may play a much more significant role than public legal institutions in ensuring that parties deal honestly with one another (Palay, 1985; Ellickson, 1991; Bernstein, 1992).

3. BUSINESS TO CONSUMER TRANSACTIONS: 1880–1995

Sears

Richard Sears and his customers faced problems of trust and risk in the late 1880s when he first began to sell mail order watches.² Sears could demand payment in full before shipping the watch. Customers were, however, likely to be concerned that Sears would take their money and not send anything or send a watch of lower than expected quality. Conversely, Sears could send the watch on credit.³ He then faced the risk that the customer would not pay the full amount. To address these issues, Sears established a policy of sending goods in return for partial payment. If the customer was not happy, he could return

the good with no questions asked and receive his money back. If the customer was happy, he would repay the credit over time.

The decision to act opportunistically depends on the relative magnitude of the gains from cheating today and the gains from honest behavior today plus the stream of benefits from future interaction. Relative to receiving full payment, partial payment decreased Sears' gain from cheating today. It also increased his gain from honesty by increasing the benefits from future interaction with that customer and any new customers obtained through referral. Partial payment did not guarantee that Sears would behave honestly. The diminished incentive to cheat together with Sears' low prices was, however, enough to attract the first customers and allow Sears to establish a good reputation. Once established, tremendous growth ensured that the sum of present and future profits were greater than the gains to cheating today.

Relative to receiving the good on credit, partial payment also decreased a customer's gain from cheating today. The stream of future benefits from interacting with Sears depended on two things. Sears had to offer a compelling value proposition relative to the alternatives. Otherwise, the customer had an incentive to make a partial payment to get the good, refuse to pay the balance and go elsewhere for future purchases. Sears and his competitors offered rural customers greater selection, higher quality, and lower prices than were available from local merchants. Sears also had to prevent individuals from being able to make purchases if they had cheated Sears in the past. Otherwise, even if the value proposition was compelling, individuals could cheat and then make a subsequent purchase under a new name or at a new address.

Sears carefully monitored the amount of credit that customers received, and cut off customers who did not pay. In a rural community, establishing a new name or address would have been difficult. Packages had to be signed for at the post office, and post office personnel knew virtually everyone in town. Further, postmasters were obligated to help prevent postal fraud. Thus, for the vast majority of customers, the benefits to future interaction with Sears were greater than the gains to cheating today.

A number of features of this early environment would later change – the cost of becoming a remote merchant, the delay between placement and receipt of an order, the cost of data transmission and the incidence of credit fraud. After the initial phase in which stores like Sears and Montgomery Ward grew from nothing to be large merchants, the cost of becoming a remote merchant was high.⁴ Remote merchants had to print catalogs, maintain warehouses, provide credit and establish a reputation for honesty. Suppose that customers believed that dealing with small remote merchants was riskier, in the sense that those merchants were more likely to not send goods, to not send goods of

appropriate quality, or to not accept returns. Then customers were likely to deal with them only if they offered a specialty product or had significantly better terms than the major catalog companies. Further, customers may have been more likely to accept credit and not pay, because the gains to future interaction were lower. Thus, reputation together with economies of scale and scope acted as effective barriers to new remote sellers.

The delay between a customer sending an order and receiving the goods was measured in weeks, limiting the scope for remote commerce. The fact that data was maintained in handwritten ledgers and later using typewriters made data transmission quite costly. As a result, privacy was not much of a problem. Finally, the incidence of credit fraud was low, because the costs of getting information were sizeable, the probability of being caught was high, and the gains were typically small.

Credit Cards

A major change occurred in the second half of the twentieth century with the development of universal credit cards. The founders of Diners Club, the first universal card, observed that salesmen in New York ate out nearly every night. Their innovation was to recognize that rather than having each restaurant offer accounts to all of their regular customers and having each customer maintain accounts at all of the restaurants in which they regularly ate, it was efficient for both salesmen and restaurants to deal with a single credit entity. And so, in 1949 the first universal credit card was born.⁵

Although merchant to customer credit continued to be important, within a decade, the number of individuals and merchants who used or accepted universal cards increased enormously. Figure 1 shows the explosion of consumer use of revolving credit, most of which is credit card credit.

Expansion in consumer confidence in credit cards and therefore their use was enhanced by passage of the federal Truth in Lending Act in 1970. The act's objective was to "assure a meaningful disclosure of credit terms so that the consumer will be able to compare more readily the various credit terms available to him and avoid the uninformed use of credit, and to protect the consumer against inaccurate and unfair credit billing and credit card practices." Among other things, the act limited cardholder liability for charges on lost or stolen cards to \$50 under most conditions.⁶ Merchants who had accepted the lost or stolen cards bore the remainder of the losses together with credit card issuers. Congress believed that such an act was necessary because of two factors. First, individual consumers had limited bargaining power relative to the issuing

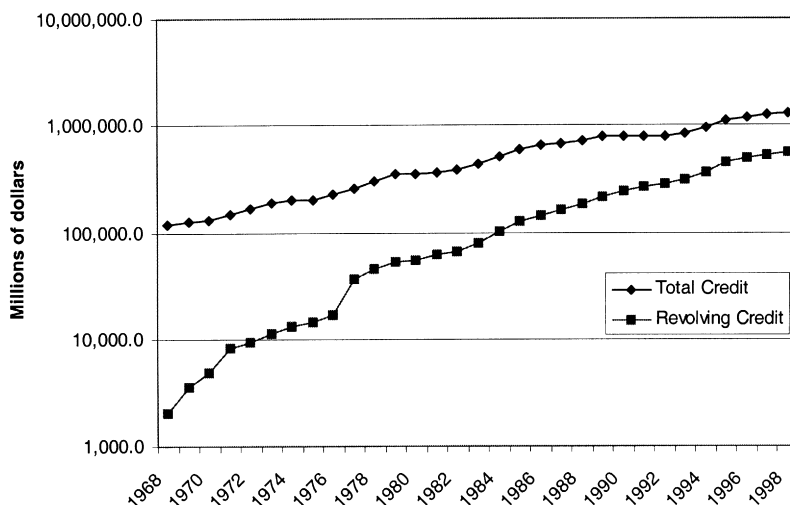


Fig. 1. Consumer Credit (Logarithmic Scale).

Notes: Table B-75.—Consumer credit outstanding, 1950–1999, Economic Report of the President, February, 2000, based on data from the Board of Governors of the Federal Reserve System. Data on revolving credit begin in 1968.

banks. Second, market forces were unlikely to be important because it was so difficult for consumers to compare offers.

With the rise of credit cards, the number of remote merchants increased, and the delay, cost of data transmission, and the cost of fraud fell. The risk and expense of offering credit had acted as a barrier to entry. With the reduction of this barrier, more individuals and firms found it profitable to establish themselves as remote merchants. The time from placement of an order to the arrival of the goods had shortened with the rise of telephones, trucks, airplanes and computers. These changes made remote commerce more attractive than it had been. Using emerging computer technology, remote merchants and credit cards began to collect personal data and to share it with major credit agencies and other merchants.

The change in the cost of fraud is attributable to three factors – the cost of obtaining the information necessary to commit fraud, the probability of being caught, and the gains to fraud. Waiters, clerks, and anyone else with access to credit cards could copy down numbers or falsify slips. And numbers obtained

from face to face or remote transactions were regularly bought and sold on the street. The emergence of markets in stolen credit cards was a reflection of a deeper change in the costs and benefits of credit card fraud (Mandell, 1990, pp. 64–69).

Remote merchants were larger, so clerks no longer had the personal knowledge of customers and their order patterns necessary to spot suspicious activity. At least initially, it was nearly impossible to determine whether a card was valid or stolen within a reasonable period of time. At the same time, urbanization made it less likely that the person behind the counter would know the person picking up the package or notice if a false change of address card had been submitted.

Universal credit cards made it possible for a thief to rack up charges much more quickly, because he could hit a large number of physical or remote stores within short period of time. The result was a high overall incidence of credit card fraud. In 1973, fraud represented 1.15% of sales. The highest incidence of fraud was in airline travel cards, gas cards, and mail order. The first two reflected the fact that airlines did not issue lists of stolen cards, and gas attendants rarely checked cards. The last, mail order, did not require a card at all, only a name and valid card number.

In response to consumer and merchant concerns, credit card companies worked to lower the incidence of fraud using a combination of education and technology. Education campaigns instructed merchants on how to identify fraudulent transactions. Similar campaigns aimed at consumers focused on the risks of giving out credit card information over the phone and what to do if a card was stolen. On the technology side, credit card companies began to offer telephone verification for large transactions and, when it became available, real time automated verification for most amounts. The net result was a fall in overall credit card fraud from 0.52% of sales in 1980 to 0.18% of sales in 1992 to 0.06% of sales in 1998.

In sum, merchants and consumers encountered problems of trust and risk in remote commerce before the advent of the Internet. Merchant opportunism was largely controlled through reputation and the extension of credit to consumers, which acted as a bond of good behavior. The need to establish a reputation and in the early days the need to offer credit created barriers to smaller merchants entering remote commerce. Merchant opportunism with respect to data received relatively little attention from consumers. Consumer opportunism in the form of credit fraud was initially controlled by limits on the amount of credit offered by a company, clerks' knowledge of customers, and tight-knit local communities. With the rise of universal credit cards and other societal changes, fraud became increasingly problematic for merchants. Through education and

technology, merchants, consumers, and credit card companies were able to drive credit card fraud to very low levels.

4. BUSINESS TO BUSINESS TRANSACTIONS: 1880–1995

In business to business remote transactions, the problems of opportunism are similar to the problems in business to consumer remote transactions. The fact that business to business relationships tend to be longer term than business to consumer relationships and in many cases offer greater value to both parties, however, lowers the incentives for opportunism.

Even with potentially more limited incentives for opportunism, the complexity of most business to business transactions, together with the incompleteness of the contracts that govern them, makes it likely that disputes will arise. Historically, two types of private institutions have evolved to address this issue. The first type of institution, decentralized institutions, is typically associated with small, usually ethnically homogeneous groups of traders who both buy and sell.⁷ Individuals governed by these institutions usually have a fairly clear, common understanding about what prevailing business norms state and what they mean in practice. Thus, dispute resolution can occur through informal investigation and settlement. Because these communities often punish cheaters through ostracism, it is critical that information about transgressions is transmitted. To maintain a common understanding of norms and transmit information efficiently, groups must remain small and fairly homogeneous.

The second type of private institution, centralized institutions, is usually associated with larger or more heterogeneous groups of traders for whom decentralized institutions would not be successful. Characterized by written rules, formal forums for dispute resolution, and efficient means for information dissemination, these institutions offer businesses or individual traders a framework for conducting trade in the absence of or as an alternative to public legal institutions.⁸ In the United States, many centralized, private institutions emerged at the end of the nineteenth century in the form of trade associations. Since there was often little agreement as to what the norms were or what they meant in practice, one of the earliest actions of these trade associations was to codify and to attempt to standardize merchant norms.

For disputes that could not be settled through private institutions, the courts took on the function of enforcing prevailing merchant norms. Local and regional differences, however, proved frustrating for business and the legal system, eventually leading the American Bar Association to make uniformity a high priority. One of the outcomes of this movement was the Uniform Commercial Code, which at least in theory codified existing United States business norms.⁹

The establishment of more uniform laws and the ability to bring disputes to the courts as a last resort mitigated risk and allowed the development of trust.

Paper contracts with physical signatures have played a key role in the resolution of disputes in all three types of institutions. Parties may or many not choose to rely on the contracts to reach a negotiated settlement, but the contracts do provide a means of tracking agreed upon price, quantity, and other dimensions of the transaction. EDI contracts have many of the same features as paper contracts. The fact that EDI is conducted over secure, private networks ensures that any transmission problems derive from the two parties and not third parties. At the time EDI is established, contingencies for dealing with transmission problems and other data security issues are spelled out in a contract between the two parties. So in practice, EDI contracts have the same status as paper contracts.

In sum, businesses also encountered problems of trust and risk in business to business remote commerce before the advent of the Internet. Since most buyers and sellers did not interact particularly frequently, private institutions arose in many segments of the United States economy. By providing rules and enforcement of rules tailored to a specific business community, these institutions allowed businesses to trust one another and trade to expand. Public legal institutions also supported trade, serving as a default forum for dispute resolution. With the rise of the Uniform Commercial Code, businesses faced a more certain legal environment. Prior to the Internet, the biggest recent change was the adoption of EDI over value added networks. Governed by private contracts and conducted over secure networks, EDI represented a fairly small change relative to the previous, paper-based paradigm.

5. CURRENT BARRIERS TO ELECTRONIC COMMERCE

Technical Barriers

As we mentioned in the introduction, bandwidth and transportation are key technical barriers to the expansion of B2C. Early predictions that catalog vendors would significantly reduce the number of paper catalogs they produce or cease publishing them altogether have been unfounded.¹⁰ The reason is fairly simple – looking at most online catalogs is tedious at low bandwidth. Currently only ten percent of households have broadband (always on, ISDN, DSL, or cable). This number is not expected to rise rapidly in near term. For instance by 2003, only 33–37% of households are predicted to have broadband (Cisco, 2001).

Within the transportation arena, there are a number of separate technological problems that have to be resolved before 24-hour delivery can become a reality. The first piece is reducing the time from order to truck. The next piece is delivery from truck to house. Route density is the primary issue here, because density is what will enable twice-daily or more frequent delivery for most households. Once the truck reaches the house, the other key – and as yet unresolved problem – is that of secure unattended delivery. Households in which all adults work are the households most likely to use remote commerce, yet they are precisely those for whom delivery is the most problematic. For some households, packages can be left on the front porch, or in urban areas, with a doorman. For remaining households, there are two options: creation of a secure delivery site, such as a drop box or garage with keypad, or extended delivery hours to ensure that someone will be home.¹¹

In the next decade, these technical barriers will be resolved. Using online catalogs will become more attractive, as households obtain faster connections, and online vendors create better catalogs. And 24-hour delivery will become a reality as more deliveries lead to greater density which in turn leads to more deliveries. Similarly, larger numbers of package deliveries will make it more attractive for either households or delivery companies to invest in secure package delivery or extended delivery hours.

The chief technical barrier for B2B is a lack of common standards across vendors and exchanges. The individuals who started many of the B2Bs drew on their own industry experience in a particular segment. During the initial rush to establish exchanges, these entrepreneurs had no standards to draw on for how these exchanges should operate. As a result, interfaces, catalogs, and document management systems were at least partially home grown. For managers interested in doing B2B procurement, this represented an unattractive outcome. Given the sunk costs in training and systems integration involved with joining each B2B exchange and the uncertainty about future standards, many managers adopted a wait and see attitude.

These technical barriers are already beginning to erode. Industry-wide exchanges such as Covisint either have or will establish standards that will reduce both the human and the systems integrations costs associated with B2B. Private networks are becoming increasingly popular, as large businesses begin B2B with existing suppliers. The large company sets the standards and then helps suppliers become compliant. Establishing a private exchange solves the standards problem for the parent company, although doing so pushes the standards problem down to the level of the suppliers. Companies that have significant experience with exchanges are increasingly acting as consultants to new exchanges. One result is that sets of client exchanges tend to have similar,

if not identical, interfaces, catalogs, and document management systems. Thus, within the next decade, we expect that a small number of standards or possibly even a single standard will emerge for B2B, making it more attractive for both buyers and suppliers to begin transacting online.

Trust and Risk

Four things have changed with the rise of the Internet to make remote commerce between businesses and individuals riskier. First, the fixed and marginal costs of becoming a remote merchant have fallen dramatically with the rise of the Internet. For many goods, economies of scale remain important on and off-line. Indeed, in commodity markets such as books, music, and computer equipment, we have seen a fairly rapid consolidation of web-based vendors, driven largely by economies of scale. For specialty sellers, however, the web has made it possible for them to reach large audiences much more cheaply than has been possible in the past. Expensive paper catalogs are not necessary, and problems of inventory management are diminished, because web-based catalogs can show actual holdings. Thus, large numbers of sellers have been able to use the web to initiate or expand remote sales. In the limit, the rise of marketplaces such as Ebay has made it feasible for individuals and small businesses to sell to one another quite profitably.

If the propensity for opportunistic behavior is correlated with size, then the rise of large numbers of small vendors has implications for the incidence of fraud. Fraud does seem to have risen; most of it associated with smaller vendors. Some partial solutions have arisen to address this problem. Examples include the rise of Bizrate, Gomez, Deja, and other organizations that pool customer experiences, making reputation more important for small firms than it otherwise might be. Ebay, the largest single marketplace for small vendors, explicitly incorporates reputation. Buyers can view feedback from previous buyers on their experience with a seller and vice versa. Even with these mechanisms, however, fraud remains persistent, in part because many buyers do not check sellers' reputations prior to purchase.¹²

Second, the fall in data transmission costs means that the risks associated with transmitting personal information through the Internet (and through conventional channels) are more significant than they were in the pre-Internet era. The problem is that intentional and unintentional security breaches can now expose individuals' personal and credit information to thousands of people. Most major merchants have implemented security to prevent third party interception of information in transit or from the site. Another significant threat is often internal. Without adequate controls, employees can access or sell

personal and credit card information stored on site (DiLonardo, 2000). Frequent news reports of security breeches have led consumers to infer that the risks of using the Internet are high.¹³

Related to this is the fact that customers are much more concerned about merchant opportunism with respect to data than previously (Buskin, 2000). Heightened awareness seems in part a result of the fact that the process is more transparent – entering personal data on a web page can almost immediately result in high levels of junk email. Fears are compounded by the possibility of meta-databases that can offer unprecedented levels of detail about customer behavior (Federal Trade Commission, 2000b, p. 21).¹⁴ These fears have had a real economic impact on the Internet channel.¹⁵ One study estimated foregone sales in the Internet channel of \$2.8 billion in 1999. Without intervention, this could rise to \$18 billion in 2002, nearly half of projected sales of \$40 billion (Federal Trade Commission, 2000b, p. 2).

Third, technology has made it more efficient to deliver some goods in digital rather than physical form. Examples include words, pictures, video, software and music. This change exposes a problem for merchants of these goods. As remote commerce with credit cards migrated from paper to telephone orders, merchants no longer had physical signatures. The signature on delivery, however, prevented customers from accepting delivery and then claiming that the goods were never delivered. If goods are delivered in digital form, there is no longer a signature. So, customers can and do regularly accept digital goods and then claim that they were never ordered or delivered.

Fourth, the probability of catching fraud on the Internet is low relative to the pre-Internet era. In telephone-based mail order, the customer representative may be able to make some determination, however imperfect, of the likely validity of the order. And written mail order has a much longer lag time and therefore a higher likelihood of discovery.

For business to business transactions, three things have changed with the rise of the Internet. First, the issues of opportunism are essentially the same as they were in the B2C context. Many smaller buyers and suppliers are either coming into existence or becoming known with the rise of the Internet. If propensity for opportunism is correlated with size, then fraud can be a significant risk.

Second, the risks associated with exchanging and storing sensitive information have increased. Businesses are particularly concerned about two issues – data security in public exchanges and data security of contracts entered into over the Internet. In public exchanges, businesses are concerned that competitors might be able to access data legally or illegally that would enable them to infer prices or other strategically sensitive information. Exchanges are mitigating this issue through additional security. For contracts entered into over the

Internet, there is the possibility – however remote – that the buyer's and seller's copies of the contract will not be identical. This could happen if the contract were altered in transit either accidentally or intentionally or if one of the parties changed their copy in the hopes of gaining some advantage. Both problems are being addressed today through a combination of encryption, trusted third party storage of contracts, and encrypted storage of important information on company sites. Although action has been taken to reduce these risks, data security issues are by no means resolved.

Third, if a Kansas firm purchases something from an Illinois firm through a server located in California, and there is a problem with the transaction, it is unclear what the default legal contract is. Viewed from a historical perspective, this is nothing new. Prior to the passage of the UCC and even subsequent to its passage, there have been differences in state law. Some of the uncertainty has traditionally been addressed through boilerplate purchase contracts. The remainder has been addressed through norms or in the case of remote contingencies, left unaddressed. With the advent of EDI, firms began to negotiate complicated contracts with partners to address potential contingencies associated with the new medium and clarify the default legal contract. Because they have neither the relative legal clarity of paper-based contracts nor the security of a comprehensive legal contract with their trading partner, however, businesses remain apprehensive about Internet-based transactions.

Institutional Solutions to Problems of Trust and Risk

Transactional opportunism is at heart an old problem. Sears solved the problem through a bond – sending the good with only partial payment – and reputation. Today the same kind of bond is not feasible, because retailers no longer extend credit directly to customers. One possibility is that consumers will quickly learn to deal only with established, brand name businesses on the Internet. Another possibility is that transactional opportunism will be addressed through enhanced reputation mechanisms together with a different type of bond.

The reputation mechanisms that underlie private institutions require three things, if they are to provide participants with incentives not to act opportunistically. First, it must be straightforward to observe other players' past behavior. In small groups, each member can often observe all other players' past behavior either directly or by asking another player. In larger groups, this is impossible. For reputation mechanisms to work, larger groups need a centralized authority that keeps records of past behavior. Some private sector firms such as Bizrate, Deja, Gomez and Ebay currently provide some information on some firms' past behavior. A trusted centralized site has,

however, not yet emerged. We believe that such a site, whether sponsored by a non-profit such as Consumers Union or a government agency, will emerge. To be effective, these ratings have to be readily available, ideally at the point of purchase. One possibility would be for merchants to implement this voluntarily. Another possibility is for ratings visibility to be a default browser-level option. Alternatively, the federal government could mandate the posting of ratings in much the same way that the Los Angeles department of public health mandates that restaurants post their health inspection letter and number grade in the front window, or states require the visible postings of gasoline prices by grade.

Second, it must be easy to determine whether violations of prevailing norms have taken place. For this to occur, the norms or rules must be clear, and the centralized authority must be able to verify the events surrounding a violation. A reading of comments available on commercial sites indicates that customers' expectations vary, suggesting that norms or rules have not emerged. Thus the trusted centralized site should establish rules, either by clearly posting existing laws (to the extent that they are applicable) or by writing new rules. If for some reason this proves infeasible, the federal government may need to act. Once the rules are established, the trusted centralized site must have a mechanism for investigating alleged violations of the rules.

Third, it must be feasible to link past behavior to future payoff. If enough consumers observe merchants' reputations and then act on that information when choosing a merchant, the link between the past and the future may be sufficiently strong to guarantee honest behavior. One way to strengthen this link is for merchants to post bonds with a trusted third party and for merchants to pay penalties if they violate the rules. An advantage of this approach is that new merchants can establish themselves by signing up with a trusted third party and paying the bond. Otherwise new merchants may be caught in a cycle of having no reputation and therefore no customers.

If a trusted third party emerges, reputations are readily observable, and the link between past and future is strong, all merchants will have incentives to deal fairly with consumers. Federal intervention may prove useful on two fronts in facilitating the operation of a reputation-based institution. The federal government may help overcome the informational problem by mandating disclosure of a merchant's reputation on its website. It may also act as a trusted third party, if one does not emerge from the non-profit sector. Thus, the institution may end up being a hybrid private-public institution, rather than a purely private one.

If the foregoing conditions are not met, not all merchants may have incentives to deal fairly with consumers. Given the uncertainty about the trustworthiness of most merchants and the high costs of pursuing complaints against merchants in

the courts, the vast majority of consumers will simply choose to deal with large trusted merchants such as Amazon.com or Walmart.com. It is worth noting that federal intervention will not be effective in resolving this market failure unless it addresses the underlying informational problem.

Data opportunism has only become a significant issue in the last decade or so.¹⁶ The term data opportunism includes a number of discrete issues including merchant tracking of customers within their site, third party tracking of customers within and across sites, the sale or exchange of customer information without adequate notice or permission, and data security. The first step in mitigating this issue is for sites to prominently post comprehensive privacy policies. A survey in late 1999 by the Federal Trade Commission found that only 20% of all websites and 42% of the top 100 websites comply with the four fair information practice principles of notice, choice, access, and security (FTC, 2000b, pp. 4, 12). Although these numbers represent an improvement over the previous year, compliance remains inadequate. If a merchant's privacy reputation could be observed along with its transactional reputation, potential customers could act on this information. Hence, a reputation-based institution could offer merchants incentives to safeguard customer data.

Creating an institutional solution to credit card fraud is inherently more difficult, because criminals exploit the fact that they can temporarily acquire an individual's reputation to make purchases. Criminals then disappear, usually without a trace. Thus, merchants must rely to a large degree on self-help. In Sears' day, making use of another customer's reputation was fairly difficult. Among other things, clerks' personal knowledge of customers served as a protection against fraud. Merchants today use sophisticated algorithms that flag suspicious transaction in much the same way to protect themselves.¹⁷ Even with active screening, charge backs for Internet retailers are typically around two percent, about ten times the rate in physical transactions.¹⁸ The problem is that these algorithms over screen, rejecting large numbers (20–40%) of valid transactions along with some fraudulent ones (Orr, 2000). Thus, merchants are actively limiting B2C, because of problems with fraud.

As a number of commentators have noted, digital goods are ideally suited for sale over the Internet, because they can be copied and delivered over the Internet at close to zero marginal cost (Shapiro & Varian, 1998; Bakos & Brynjolfsson, 1999, 2000; de Figueiredo, 2000). These properties, however, also make them the most vulnerable to credit card fraud. At the moment, rates of fraud are as high as 30% in these segments, particularly pornography and online gambling (Bicknell, 1999; *Card News*, 2000). The problem is so severe that American Express no longer serves merchants whose primary business is the sale of pornography, and Visa and Mastercard are imposing ever-stiffer

penalties on sites with high levels of charge backs, one indication of fraud (Hisey, 2000). In the absence of effective anti-fraud measures, transaction costs for this segment may prove to be prohibitive, leading to market failure. Alternatively, merchants may begin to require cash payments or delay transfer of the good for several days if a credit card is used, negating many of the benefits of the Internet for delivering these types of goods.

We expect many of the foregoing problems to be resolved. Opportunism can be dealt with through private or hybrid private-public institutions that use reputation to provide merchants with incentives to deal honestly with their customers. Although credit card fraud is an inherently difficult problem, merchant self-help measures appear to have lowered transactions costs for physical goods to reasonable levels. Transactions costs for digital goods are currently very high and appear to be restricting the sale of these goods. Data security may be addressed through more stringent measures at exchanges or through the current movement towards private exchanges. On the contract side, current movements towards more stringent security may be fruitful. And for high value contracts, paper or fax may remain the dominant mode of business. Finally, public legal institutions in the form of amendments to the UCC and expanded case law may individually or jointly diminish the legal uncertainty of transacting over the Internet to acceptable levels.

6. THE COMING EXPLOSION OF B2C AND B2B: IMPLICATIONS FOR THE STATES

Bella Hess and Quill

With the removal of barriers to B2C and B2B and the consequent expansion in sales through this channel, tax issues will become increasingly important. As a result of two landmark United States Supreme Court decisions, *National Bella Hess* in 1967 and *Quill* in 1992, remote sellers without a physical presence in a state are not currently obligated to collect and remit compensating use taxes. The Court's rationale in both cases rested on the complexity and therefore high costs of compliance by remote vendors given the very large number of taxing units in the United States. Destination individual and businesses are legally responsible for paying use taxes to their state of residence. State oversight and direct collection of use taxes from households have, however, relied largely on voluntary reporting.¹⁹ The current conflict over retail sales over the Internet is an extension of historical conflict between the state and local sector and remote mail order vendors who have been able to avoid use tax collection responsibilities. The stakes have become higher in the post-Internet era, however, as

the dollar amounts of remote commerce increased.

Whether or not sales and use tax collections are being significantly eroded by the diversion of transactions from traditional commercial channels to the Internet is a hotly debated issue. Cline and Neubig (1999) in a paper for AOL concluded on the basis of an examination of 1998 data that diversion and therefore revenue loss would be rather modest. Duncan (1999) subsequently complained that 1998 was hardly relevant to what might happen in 2003. Bruce and Fox (2001), utilizing Forrester Research, Inc. estimates of B2B and B2C activity growth, concluded that the loss of sales and use tax collections resulting from diversion could be substantial. The loss comes at a time when the sales and use tax base continues a secular decline as consumers substitute out of taxed commodities to untaxed commodities. As a result of this secular decline and the diversion of taxable sales to the Internet, Bruce and Fox estimate that states will have to raise their 2011 sales and use tax rates anywhere from 0.46 to 0.94 percentage points in order to maintain sales and use tax collections (Bruce & Fox, 2001, p. 14). Interestingly, at least 75% of the incremental revenue loss due to diversion results from B2B tax base loss rather than B2C tax base loss. This reflects the fact that business has more rapid adopted electronic commerce (Bruce & Fox, 2000, Table 2) than typical consumers. Table 1 displays the Bruce-Fox (2000) direct sales and use tax loss estimates, and shows the importance of B2B.

Table 1. Bruce-Fox (2000) Estimates of State Sales and Use Tax Losses Due to Diversion of B2C and B2B Sales.

	1999	2000	2001	2002	2003
Total Business-to-Business	106.59	244.87	486.63	821.80	1297.80
Less Exempt Sales	-47.54	-105.05	-208.76	-369.81	-616.45
Less B2B on which sales/use tax collected	-34.07	-80.96	-164.77	-281.59	-444.24
Equals B2B Base Loss	24.98	58.87	113.09	170.40	237.11
Less substitution for other remote sales	-12.49	-29.43	-56.55	-85.20	-118.55
Equals Incremental B2B Base Loss	12.49	29.43	56.55	85.20	118.55
Approximate Revenue Loss from B2B	0.80	1.88	3.61	5.44	7.57
Total Business-to-Consumer	19.75	37.79	62.59	98.62	140.19
Less Exempt B2C	-8.32	-15.34	-23.53	-32.74	-41.78
Less B2C on which sales/use tax collected	-1.14	-2.60	-5.51	10.54	-20.57
Equals B2C Base Loss	10.29	19.85	33.55	55.34	77.85
Less substitution for other remote sales	-3.60	-6.95	-11.74	-19.37	-27.25
Equals Incremental B2C Base Loss	6.69	12.90	21.81	35.97	50.60
Approximate Revenue Loss from B2C	0.43	0.82	1.39	2.30	3.23
Approximate Incremental Revenue Loss	1.23	2.70	5.00	7.74	10.80

Two Institutional Solutions

Some institutional change will have to occur, either to adjust to lower sales and use tax revenues or to address the concerns raised by the Court in *Bella Hess* and *Quill*. The high-tech industry, political commentators and elected federal, state and local officials have not overlooked the implications of an explosion of B2C and B2B for the American federal system. Most attention has focused on the fiscal implications to state and local governments of B2C replacing face to face or traditional commerce, and state and local sales and use tax collections decelerating or atrophying. The Internet Tax Freedom Act, enacted by Congress in October 1998 and renewed as the Internet Tax Non-Discrimination Act of 2001 on November 28, 2001, successfully froze existing state and local taxes on Internet access, and placed a moratorium on new taxes through October 2003.

While the no-tax movement was largely successful in forestalling the imposition of new or broader transactions taxes to the Internet, there arose, almost simultaneously, discussions among the state and local sector, various segments of the business community and some academics, including one of this paper's authors, about the dimensionality of some sort of grand political trade that might resolve at least some of the emerging tax issues. The idea of a grand political trade contained the following major elements:

- radical simplification of extant state and local sales and use taxes in return for
- political support from the business community to expand their duty to collect and remit use taxes with the effect of overturning *Bella Hess* and *Quill*.

This grand political trade was debated in two forums during 1997–2000: the National Tax Association's Project on the Taxation of Telecommunications and Electronic Commerce (1997–1999), and the Advisory Commission on Electronic Commerce (1998–2000). The range of issues which might effect the above grand political trade includes not only the adoption of uniform administrative rules in such areas as estimated payment, penalty interest rates, uniform registration forms and rules, uniform definitions of casual sales, uniform definitions of timeliness of payments, vendor's discount for turning over use taxes to the destination state, adoption of "ship-to address" or credit card mailing address for determination of state of destination, but also:

- the elimination of all local sales and use taxes and adoption of one tax rate per state; and

- acceptance of uniform definitions of commodities and commodity exemptions.

Given that there are better than 15 states whose local government budgets depend 20% or more on *local* sales and use taxes, the elimination of local authority to impose sales and use taxes will entail a major upheaval in state-local fiscal relations (Strauss, 1997, Table 16). Overall, better than 7,900 local governments utilize some form of sales and use tax to finance their services.

The NTA project issued its report on September 8, 1999. While there was not agreement among the participants on the final form of a simplified sales and use tax, or an agreement on the particular mechanism (see discussion below) that would implement this reformed sales and use tax, there was a thorough discussion of the options available in solving virtually all of problems arising from the development of a simplified sales and use tax.

The Advisory Commission had a tumultuous start with several government members suing the others in federal court about the propriety of AOL and Netscape each having representation on the Commission after AOL purchased Netscape. Then there was controversy over the naming of its executive director whose spouse was a lobbyist for technology interests, controversy over \$20,000 per person donations to be seated next to Commission Members at private dinners, and finally controversy over the issuance of a Final Report that was not agreed to by the supermajority required in the federal legislation which created the Commission.

Subsequent to the dissolution of the Advisory Commission, several government groups (the National Conference of State Legislatures, National Governor's Conference, and Federation of Tax Administrators) decided to meet to work out the details of a model statute which each state might adopt as a means for the state and local sector to demonstrate both its good faith and its resolve to move forward to collect use tax once the Moratorium expires. Currently 39 states are involved in the cooperative development of a uniform, standardized sales and use tax.²⁰ Importantly, Kansas, Michigan, North Carolina and Wisconsin are running a pilot to test current technology viz-à-viz real-time sales and use tax calculations, collection and reporting services. The industry participants in this project are esalestax.com, Inc. Pitney Bowes (Vertex, Inc is a subcontractor), Taxware International (Hewlett-Packard and Pitney Bowes are subcontractors to Taxware). The basic idea being tested is to construct a trusted third party, or Certified Service Provider (CSP), to contract with merchants and vendors to perform their sales tax functions. The CSP's functions include tax application, sales and use tax exemption administration, computation, filing, and remittance to the pilot states. The CSP is reimbursed by the pilot state for its services, not

by the merchant or vendor. Moreover, pilot activities cover not only remote sales over the Internet, but also phone, mail order, and face to face sales as well. As of March, 2001, 19 states have introduced model sales and use tax legislation based on the work product of the 39 states.

While there is much state and local legislative activity surrounding the development of the model sales and use tax statute and the presumption that widespread adoption would create a multi-state compact, there remains a variety of vexing compliance questions about the ultimate efficacy of such an approach. Hellerstein (2000), among others, points to the uneven state adoption of the Uniform Division of Income for Tax Purposes Act, some 40 years since first proposed, and the observed heterogeneity in state provisions in such important areas as the nature of the apportionment formula as grounds for pessimism about such a purely cooperative approach. It is worthwhile to remember that the Supreme Court asserted in *Moorman* in 1978 that Congress possesses the power to legislate uniform state tax rules among the states (Hellerstein, 2000, pp. 1309–1310).

Should Congress move forward to do so, the questions of the precise federal role and policy design arise. As McLure (1998) points out, a prominent role for the federal government inevitably runs afoul of state (and local) sovereignty concerns. On the other hand, without some sort of active federal role, it is difficult to envision how use tax compliance can be achieved. A new federal consumption tax with the proceeds shared back to replace extant sales and use taxes may appear straight-forward; however, it is likely that state sovereignty concerns would be vigorously expressed before Congress and make that option unattractive.

The problem of federal policy design involves whether the Congress should provide financial incentives to the states for state by state adoption of what the states agree on as a model statute, and/or provide financial incentives to private interests to achieve the desired goal of an enlarged private responsibility to collect and remit use taxes. Depending on how the enlarged federal role is expressed, there may or may not be new, important implications for other state taxes, especially business net income taxes. An ongoing concern of the business community involves the potential expansion of the concept of nexus for other business taxes, primarily business activity, income and franchise taxes should the Supreme Court through a new decision or the federal government statutorily overturn *Bella Hess* and *Quill*.

There are a variety of other approaches that the federal government might legislatively pursue that could effect an enlarged private duty to collect use taxes without risking an expansion or alteration of nexus concepts for income and franchise taxes. These approaches generally involve conditioning the

application of federal authority to require remote vendors to collect and remit use taxes on state adoption or participation in a reformed and simplified system. The question that then arises is the form of incentive the federal government might legislatively construct to encourage all states that currently impose a sales and use tax to adopt or participate in a reformed system. Given the extent of interdependencies originating between the federal government and the states, this does not seem like a difficult design problem. For example, all states benefit from the bilateral exchange of tax information between the IRS and their state revenue agencies. The federal government might condition further access to such information on the voluntary adoption by any state of a “qualified sales and use tax” which would allow the state to choose the one rate of tax to meet its fiscal needs, but in all other respects be identical among the states. Fiscal sovereignty would thus be maintained along with simplification, albeit the elimination of local sales and use taxes. This sort of approach, however, does not ensure an information base available to the federal government, presumably the IRS, to ensure that remote collection of use taxes occurs.

Conditioning continued eligibility for the Federal Unemployment Tax Act credit on the positive agreement by any employer to participate in the collection and remittance of use taxes under a harmonized system addresses that problem; however, this financial incentive may not be sufficient to ensure employer/remote vendor compliance. Federal adoption of a high penalty tax, say a 15% excise, on all interstate transactions into states which have adopted a reformed and simplified sales and use tax regime unless the vendor demonstrably collects and remits the use tax to the destination state, would appear to provide sufficient incentives for private compliance, and also ensure state adoption of the reformed and simplified sales and use tax. To provide clarity and longevity to such an arrangement, the federal mechanism would be well advised to place the details of the reformed and simplified sales and use tax in the Internal Revenue Code.²¹

7. PAST AS PROLOGUE?

Our historical review of the way nineteenth century retail commerce evolved reminds us that the states followed suit by modernizing their commercial law institutions to address merchant and customer concerns about risk.²² Within the past six months, most of the states have adopted some form of digital signature legislation whose purpose is to provide an improved level of assurance to remote vendors, primarily on the web. Outstanding in most states, however, is counterpart legislation that will protect customers on the web from various forms of vendor fraud. It is readily imaginable that state inaction to

reduce the risks of B2C commerce for vendors and consumers, through the enactment of either a digital Universal Commercial Code, or a uniform adoption of UCITA, will put off the supposed explosion of B2C commerce, and forestall the adverse revenue consequences that pessimists expect. On the other hand, state action to solve the commercial law problems facing B2C may be a precursor to achieving some sort of political solution to either convincing Congress that federal legislation is necessary to overcome *Bella Hess* and *Quill*, or finding a cooperative solution that is truly workable.

What is clear to us is that meaningful solutions to the collection and remittance of remote use tax problem will require not only vast simplifications and harmonization of extant state sales taxes, which are long overdue, but also a likely role for either the federal government through a federal agency to determine whether each state's version of a reformed sales and use tax adheres to an agreed upon template, or some other credible third party agency which could prove effective to induce both private and public sector compliance with the model statute.

To date, the states have not recognized that the value to private economic forces from state adoption of uniform commercial laws governing both digital commerce and digital privacy that could then make the explosion in electronic commerce a reality could readily be harnessed to move forward the states use tax agenda.

NOTES

1. This section reviews some basic results from game theory. These results are discussed in more detail in Clay and Ingram (2000).

2. This paragraph is based on Weil (1977).

3. Like other merchants of the period, the cost of the credit was included in the price of the good. This made accounting simpler and allowed merchants to avoid state banking and usury laws.

4. To the extent that Sears and Montgomery Ward legitimized remote selling in the minds of consumers, these firms may have lowered costs relative to what they would have been. Overall, however, costs appear to have been higher for these later entrants.

5. The discussion of the credit card industry draws on Mandell (1990).

6. "The Truth-in-Lending Act originated as part of former President Lyndon Johnson's Great Society plan providing "federal consumer protection." One commentator explained that the passage of TILA was an effort to "level the playing field" between consumers and "large corporations." Harrington (2001), p. 113. The other big innovation in 1970 was the introduction of standardized magnetic strips on credit cards.

7. For examples of this type of institution, see Greif (1989, 1994), Clay (1997a, 1997b), Macauley (1963), Ellickson (1991), and Reid (1980).

8. For examples of this type of institution, see Greif, Milgrom, and Weingast (1994) and Bernstein (1992, 1996, 2001).

9. Recent legal scholars have suggested that it did not actually codify existing business norms, since little empirical work was done on the content of the norms and the UCC contains some parts that ran counter to existing norms. See Bernstein (1996, 1999) and references cited therein.

10. "We saw some major consumer catalogers cut back their mailings and watched their sales dramatically decline" (Cyr, 2000).

11. We are assuming that most people will find it inconvenient to have large quantities of goods delivered to their workplace.

12. It is particularly problematic on Ebay, because many small vendors do not take credit cards. This forces customers to pay by check or some other cash equivalent payment system. Unless a customer explicitly chooses to use escrow, he has no recourse if the merchant acts opportunistically. For a list of Federal Trade Commission Internet Auction Fraud cases, see Federal Trade Commission (2000a). On international ecommerce readiness and the role of the state in fostering trust, see Oxley and Yeung (2001).

13. Weekly media reports of compromised data probably only represent a small fraction of the true incidence. See, for instance, Stoughton (2000) on recent breeches at Eve.com, Western Union, and AOL. Exactly what consumers are afraid of is unclear given that liability is limited to \$50. It may be that costly outcomes such as identity theft are more common in the Internet channel than in physical channels. Note that data on the last six months indicates that credit card fraud is a relatively minor problem relative to Internet scams of various types and auction fraud.

14. The FTC survey indicates that more than half of all sites permit third party (e.g. Doubleclick) placement of cookies and less than half of those tell consumers that third parties may be placing cookies.

15. The overall impact is less clear in the sense that those same purchases may be occurring in the physical channel.

16. Interestingly, the FTC survey showed that privacy seals have not been widely adopted overall, although 45% of the sites in the top 100 had them. Further, having a seal did not guarantee that the site complied with the four principles of fair information.

17. Major credit card companies are also working to address fraud. For instance, American Express will begin offering disposable credit card numbers (Sapsford, 2000). And customers may soon enter passwords that are routed directly to the bank to authenticate transactions or use digital signatures. Digital signatures (or equivalently fingerprints, retinal scans, or typing) still have unresolved data security issues.

18. The Internet rate is similar to physical rates in 1973. One problem in the current environment is the differential across channels. A second problem is that remote sellers are often smaller now, and so more vulnerable to fraud. A third problem is that the penalties are higher now and card issuers are more likely to cut off merchants with high chargebacks (Anguin, 2000; Guernsey, 2000).

19. Due and Mikesell (1994, p. 265) report that 17 state income tax forms contain a use tax reporting line.

20. For a description and timeline of the project, see <http://www.geocities.com/streamlined2000/>

21. See Strauss (2000) for a further discussion of these alternatives, Peha and Strauss (1997) for a discussion of technology-related problems, and Hellerstein (2000) for a positive, constitutional review of these mechanisms.

22. See Armstrong (1991) for a discussion of the role of the National Conference of Commissioners on Uniform State Laws.

ACKNOWLEDGEMENTS

We would like to thank the participants at the Conference on New Institutionalism in Strategic Management and the editors of this special issue, Brian Silverman and Paul Ingram, for helpful comments. We would also like to thank session participants at the National Tax Association 2000 meetings for comments on an earlier version of this paper.

REFERENCES

- Anguin, J. (2000). Credit-Card Scams Bedevil E-Stores – With No Signatures to Prove Who Placed Orders, Sites Are Left Footing the Bills. *Wall Street Journal* (September 19), B1.
- American Bar Association. (2000). *The American Bar Association's Internet Jurisdiction Project, A Report on Global Jurisdictional Issues Created by the Internet*. <http://www.abanet.org/buslaw/cyber/initiatives/proj-documentation.html>
- Armstrong, W. P. (1991). *A Century of Service: A Centennial History of the National Conference of Commissioners on Uniform State Laws*. West Publishing Company.
- Bakos, Y., & Brynjolfsson, E. (2000). Bundling and Competition on the Internet: Aggregation Strategies for Information Goods, *Marketing Science*, 19, 63–82.
- Bakos, Y., & Brynjolfsson, E. (1999). Bundling Information Goods: Pricing, Profits and Efficiency. *Management Science*, 45, 1613–1630.
- Bernstein, L. (1992). Opting Out of the Legal System: Extra-legal Contractual Relations in the Diamond Industry. *Journal of Legal Studies*, 21, 2583–2612.
- Bernstein, L. (1999) The Questionable Empirical Basis of Article 2's Incorporation Strategy: A Preliminary Study. *University of Chicago Law Review*, 66, 710–780.
- Bernstein, L. (1996). Merchant Law in a Merchant Court: Rethinking the Code's Search for Immanent Business Norms. *University of Pennsylvania Law Review*, 144, 1643–1696.
- Bernstein, L. (2001). Private Commercial Law in the Cotton Industry: Value Creation Through Rules Norms and Institutions. Forthcoming in: *Michigan Law Review*.
- Bicknell, C. (1999). Credit Card Fraud Bedevils Web. *Wired News* (April 2). <http://www.wired-news.com/news/business/0,1367,18904,00.html>
- Bruce, D., & Fox, W. (2000). E-Commerce in the Context of Declining State Sales Tax Bases. *National Tax Journal*, 53, 1373–1388.
- Bruce, D., & Fox, W. (2001). *State and Local Sales Tax Revenue Loses from E-Commerce: Updated Estimates* (September). Center for Business and Economic Research, University of Tennessee, Knoxville.
- Buskin, J. (2000). E-Commerce (A Special Report): Revamping the Model – Choice and Trust. *Wall Street Journal* (April 17), R34.
- Card News (2000). Merchants and Issuers Must Address Internet Credit Card Fraud. *Card News*, 15 (September 6), 1.
- Census. (2001). *E-Stats, Measuring the Electronic Economy*. <http://www.census.gov/estats>

- Cisco. (2001). Government Affairs, Facts and Stats, Broadband Deployment. <http://www.cisco.com/warp/public/779/govtaffs/factsNStats/broadbanddeployment.html>
- Clay, K. (1997a). Trade without Law: Private-Order Institutions in Mexican California. *Journal of Law, Economics, and Organization*, 13, 202–231.
- Clay, K. (1997b). Trade, Institutions, and Credit. *Explorations in Economic History*, 34, 495–521.
- Cline, R., & Neubig, T. (1999). The Sky is Not Falling: Why State and Local Revenues Were Not Significantly Impacted by the Internet in 1998. *State Tax Notes*, 17, 43–49.
- Cyr, D. (2000). Your New Customer. *Catalog Age*, 17, 125–132.
- DiLonardo, M. J. (2000). Little E-Shop of Horrors: Don't be Afraid to Buy Online, but Take Precautions – and Never Use a Debit Card, *The Atlanta Journal-Constitution* (July 9), 1.
- Due, J. F., & Mikesell, J. L. (1994). *Sales Taxation: State and Local Structure and Administration*. Urban Institute Press (2nd ed.).
- Duncan, H. (1999). State Revenue Losses from E-Commerce Underestimated. *State Tax Notes*, 17, 245–246.
- Ellickson, R. C. (1991). *Order Without Law*. Harvard University Press.
- Emarketer. (2001). *ECommerce: B2B*. http://www.emarketer.com/ereports/ecommerce_b2b/welcome.html
- Federal Trade Commission. (2000a). *Going, Going, Gone, Law Enforcement Efforts to Combat Internet Auction Fraud*. <http://www.ftc.gov/bcp/reports/int-auction.pdf>
- Federal Trade Commission. (2000b). *Privacy Online: Fair Information Practice in the Electronic Marketplace, A Report to Congress*.
- de Figueiredo, J. M. (2000). Finding Sustainable Profitability in Electronic Commerce, *Sloan Management Review*, 41, 41–52.
- Goolsbee, A., & Zitttrain, J. (1999). Evaluating the Costs and Benefits of Taxing Internet Commerce. *National Tax Journal*, 52, 413–428.
- Greif, A. (1989). Reputation and Coalitions in Medieval Trade: Evidence on the Maghribi Traders. *Journal of Economic History*, 49, 857–882.
- Greif, A. (1994). Cultural Beliefs and the Organization of Society. *Journal of Political Economy*, 102, 912–950.
- Greif, A., Milgrom, P., & Weingast, B. R. (1994). Coordination, Commitment, and Enforcement: The Case of the Merchant Guild. *Journal of Political Economy*, 102, 715–776.
- Guernsey, L. (2000). Putting Your Family Heirlooms Up for Bids May Make You a Target for Credit Card Fraud. *New York Times* (August 3), G1.
- Harrington, J. (2001). To Litigate or Arbitrate? No Matter – The Credit Card Industry is Deciding For You. *Journal of Dispute Resolution*, 101–118.
- Hellerstein, W. (2000). Federal Constitutional Limitations on Congressional Power to Legislate Regarding State Taxation of Electronic Commerce. *National Tax Journal*, 53, 1307–1325.
- Hisey, P. (2000). At War Over Merchant Risk. *Credit Card Management*, 13, 59–62.
- Ingram, P., & Clay, K. (2000). The Choice-Within-Constraints New Institutionalism and Implications for Sociology. *Annual Review of Sociology*, 26, 525–546.
- Jaffee, J., & Freeman, J. (2002). Institutional Change in 'Real-Time': The Development of Employee Stock Options in German Venture Capital Contracts, 1997 to 2000. *Advances in Strategic Management* (forthcoming).
- Johnson, D., & Post, D. (1996). Law and Borders-The Rise of Law in Cyberspace. *Stanford Law Review*, 48, 1367. <http://www.temple.edu/lawschool/dpost/Borders.html>
- Macaulay, S. (1963). Non-contractual relations in business: A preliminary study. *American Sociological Review*, 28, 55–67.
- Mandell, L. (1990). *The Credit Card Industry: A History*. Twayne Publishers.

- McLure, C. E., Jr. (1998). Electronic Commerce and the Tax Assignment Problem: Preserving State Sovereignty in a Digital World. *State Tax Notes*, 17, 1169–1173.
- Orr, B. (2000). Will e-commerce reverse card fraud trend? *ABA Banking Journal*, 92, 59–62.
- Oxley, J., & Yeung, B. (2001). E-Commerce Readiness: Institutions and International Competitiveness. *Journal of International Business Studies*, 32, 705–723.
- Palay, T. M. (1985). Avoiding Regulatory Constraints: Contracting Safeguards and the Role of Informal Agreements Source. *Journal of Law, Economics and Organization*, 1, 155–175.
- Peha, J., & Strauss, R. P. (1997). A Primer on Changing Information Technology and the Fisc. *National Tax Journal*, 50, 607–621.
- Reid, J. P. (1980). *Law for the Elephant*. The Huntington Library.
- Sapsford, J. (2000). American Express Credit Cards to Offer Disposable Numbers for Web Shopping. *Wall Street Journal* (Sept. 8), B6.
- Stoughton, S. (2000). Web Retailers' Growth Spurt May Be Ebbing. *Boston Globe* (Sept. 18), A1.
- Strauss, R. P. (1997). Administrative and Revenue Implications of Alternative Federal Consumption Taxes for the State and Local Sector. *The American Journal of Tax Policy*, 14, 361–452.
- Strauss, R. P. (2000). Federal Tax Mechanisms to Enable State Taxation of Final Consumption. *Tax Notes*, 87, 361–452.
- Weil, G. L. (1977). *Sears, Roebuck, U.S.A.* Stein and Day.

This Page Intentionally Left Blank

PART 3:
WHAT LIMITS THE SCOPE
OF THE FIRM?

This Page Intentionally Left Blank

INFORMAL AND FORMAL ORGANIZATION IN NEW INSTITUTIONAL ECONOMICS

Todd R. Zenger, Sergio G. Lazzarini and
Laura Poppo

ABSTRACT

Exchanges are governed by a set of formal institutions (contracts, incentives, authority) and informal institutions (norms, routines, political processes) that we argue are deeply intertwined. However, for the most part, informal institutions are treated as exogenous forces changing the benefits to using in an alternative formal structures, and formal institutions are treated as mere functional substitutes for informal elements governing exchanges. As a result, scholars have not sufficiently explored the interactions between formal and informal institutions. We contend that the failure to integrate these concepts into a common theory has led to faulty reasoning and incomplete theories of economic organizations. In this paper, we highlight three potential areas of research exploring the interplay between formal and informal institutions: first, whether formal institutions support (complement) or undermine (substitute for) the contributions of informal institutions; second, how vacillation in formal organizational modes allows managers to efficiently alter the trajectory of informal institutions; and third, how certain informal institutions can lead to hierarchical failure, thereby requiring managers to constrain the boundaries of the firm.

The New Institutionalism in Strategic Management, Volume 19, pages 277–305.
Copyright © 2002 by Elsevier Science Ltd.
All rights of reproduction in any form reserved.
ISBN: 0-7623-0903-2

INTRODUCTION

The core argument of new institutional economics (NIE) is that “institutions matter and are susceptible to analysis” (Matthews, 1986, p. 903; Williamson, 1996, p. 3). Institutions provide rules, constraints and incentives that are instrumental to the governance of exchanges, and can be formal or informal in nature¹ (North, 1990). We define formal institutions as rules that are readily observable through written documents or rules that are determined and executed through formal position, such as authority or ownership. Formal institutions, thus, include explicit incentives, contractual terms, and firm boundaries as defined by equity positions. We define informal institutions, in turn, as rules based on implicit understandings, being in most part socially derived and therefore not accessible through written documents or necessarily sanctioned through formal position. Thus, informal institutions include social norms, routines, and political processes.

Despite the recognition by some NIE scholars that informal institutions play a crucial role in defining societal rules (e.g. Denzau & North, 1994; Ensminger, 1997; Greif, 1997), the application of NIE to the study of micro-level issues relevant to business strategy – such as organizational design, firm boundaries, and interorganizational relations – has largely focused on formal institutions. Discussing social norms, Hart (2001, p. 15) contends that “it has been difficult to incorporate norms into the theory of organizations . . . although there has been some interesting recent work on this topic, this work has not to date greatly changed our views about the determinants of organizational forms.” Other authors, while acknowledging the role of informal institutions, treat them as exogenous forces that simply change the benefits to using alternative formal structures. For instance, Williamson (1991, p. 291) considers the presence of reputation from social networks as a “shift parameter” reducing the incidence of opportunistic behavior and thus favoring non-hierarchical forms of governance. This type of analysis sharply differs from the work of organizational theorists and economic sociologists, who have stressed the central role of informal mechanisms in governing exchanges both internal (Crozier, 1964; Roethlisberger & Dickson, 1939; Trist & Bamforth, 1951) and external to the firm (Granovetter, 1985; Powell, 1990; Uzzi, 1996).

Thus, for the most part, formal institutions have been analyzed and evaluated quite independently of informal institutions. The converse is also true: the study of informal institutions has largely abstracted from the importance of formal institutions, often viewing them as mere functional substitutes. As a result, scholars have not sufficiently explored the *interactions* between formal and informal institutions. We contend that the failure to integrate these concepts

into a common theory has led to faulty reasoning and significant weakness in theories of economic organization. In this essay, we explore how a careful treatment of both informal and formal institutions in the analysis of economic organization provides key insights into the most fundamental predictions of NIE. We submit that a more careful exploration of this relationship is vital to developing a theory of efficient governance.

In particular, we highlight three potential areas of research. First, an improved understanding of the relationship between formal and informal institutions allows us to assess whether formal institutions support (*complement*) or undermine (*substitute* for) the contributions of informal institutions. Currently, managers have surprisingly little basis for determining whether and when formal contracts enhance, damage, or replace trust in interorganizational exchanges or whether and when formal incentives enhance, damage, or replace cooperative behavior within the firm. In this paper, we attempt to provide the beginnings of a theory to untangle this relationship.

Second, a more careful examination of the relationship between formal and informal institutions suggests a need to rethink the fundamental proposition of static alignment pervasive in NIE (Williamson, 1991, p. 277) and other contingency theories in organization theory (Lawrence & Lorsch, 1967). We contend that scholars often mistakenly predict static alignment – matching formal mechanisms to exchange conditions – when the prediction that logically emerges from a more careful assessment of organization theory assumptions is *dynamic* alignment by vacillation in formal mechanisms. This is, informal institutions affecting the operation of an organizational form, such as political processes and routines, cannot be adjusted directly and thus require changes in formal structures to gradually alter their trajectory. By observing how formal governance causes changes in informal elements, a completely different theory of organizational choice emerges: under some circumstances, even static exchange conditions may demand dynamic responses in terms of formal structures (Nickerson & Zenger, 2002).

Third, a more careful assessment of the relationship between formal and informal governance mechanisms is critical to understanding the boundaries of the firm. NIE has successfully built upon Coase's (1937) seminal work to explain how market failure – namely, transaction costs – shifts the organization of exchanges from markets to hierarchies. However, the theory has not adequately explored the sources of *hierarchical failure* that pose limits to the size and vertical scope of firms. We submit that the analysis of informal institutions within firms is critical to understanding how firms determine their boundaries.

This remainder of this paper is organized into five sections. In the first section, we examine the relationship between informal and formal governance and

highlight four common assumptions used in the remainder of the essay. The next three sections use these assumptions to explore three ways in which a careful look at the relationship between formal and informal contributes to our development of NIE. Our final section concludes.

THE INTERPLAY BETWEEN FORMAL AND INFORMAL INSTITUTIONS

We argue that the managers' implicit task is to shape informal and formal institutions influencing the operation of an organizational form in such a way as to increase the functionality that they *collectively* deliver. By functionality, we mean a variety of dimensions such as capacity to coordinate tasks, to achieve levels of cooperation, or to respond to changing market conditions, which invariably determine governance costs. In this section, we examine how formal and informal institutions interact and jointly define the functionality of organizational forms.

The Role of Informal Institutions

Research in organization theory has stressed the central role of informal institutions in defining how work is performed and tasks are accomplished within firms (Barnard, 1938; Crozier, 1964; Roethlisberger & Dickson, 1939; Trist & Bamforth, 1951). While formal institutions define the "normative system designed by management" or the "blueprint for behavior" (Scott, 1981, p. 82), informal institutions define the actual behavior of players. Thus, Roethlisberger and Dickson (1939, p. 559) observe that:

Many of the actually existing patterns of human interaction have no representation in the formal organization at all, and these are inadequately represented by the formal organization . . . Too often it is assumed that the organization of a company corresponds to a blueprint plan or organization chart. Actually it never does (1939, p. 559).

Consistent with this view, many authors remark that the "informal organization," supported by informal institutions within firms, is not only distinct from formal rules, but also has a critical role in influencing the operation of firms. For example, decision making within firms is strongly influenced by political processes (Pfeffer, 1978); patterns of communication are largely a function of informal relationships and shared language (Zenger & Lawrence, 1989); tacit knowledge is rooted in organizational routines (Argote, 1999; Nelson & Winter, 1982); and perceived obligations between employer and employee transcend job descriptions and formal contracts (Rousseau & McLean Parks, 1993).

The importance of informal institutions is also recognized in market contexts as well. Granovetter (1985) insists that formal institutions have limited ability to support exchange and thus social networks embodying informal institutions such as norms and trust play a crucial role in shaping economic institutions of governance. Macaulay's (1963) famous study of the governance of business relationships is consistent with this view. He observes that "businessmen often prefer to rely on a 'a man's word' in a brief letter, a handshake, or 'common honesty and decency' – even when the transaction involves exposure to serious risks" (1963, p. 58). Thus, informal dealings have the advantage of promoting flexibility and responsiveness to changing conditions, avoiding costly renegotiation of contract clauses (Macneil, 1978). The advantages of such informal contracting mechanisms – commonly referred to as *relational governance* – are now extensively discussed. Relational governance supports cooperation through norms and reciprocal obligations that transcend initial contract clauses and economize on the costs to use the legal system (Dore, 1983). This discussion leads to:

Assumption 1: Informal institutions strongly influence the functionality of organizational forms.

Formal Institutions as Mechanisms of Change

The functional consequences of informal institutions call for managerial action seeking their optimization. For instance, Lincoln (1982, p. 11) observes that "informal networks are indispensable to organizational functioning, and managers must learn to manipulate them for organizational ends." However, the fact that informal institutions in general are difficult to manipulate engenders major managerial challenges. Fortunately, changes in formal institutions, which can be directly manipulated, appear to strongly influence changes in informal institutions within firms. Internal routines, norms, and networks of influence develop over time in response to an organization's formal structure (Shrader, Lincoln & Hoffman, 1989; Stevenson, 1990; Tichy, 1980).

Research shows, for instance, that the operation of informal networks is influenced by the positions of individuals in the formal hierarchy (Brass, 1984; Krackhardt, 1990). Thus, when managers centralize a previously decentralized activity, not only do formal reporting relationships change, but informal patterns of communication also shift. In addition, formal institutions appear to influence the trajectory of informal elements in interorganizational relationships. Thus, long-term contracts and joint equity stakes (ownership) may help transform weak informal ties into strong ties involving mutual trust (Doz, 1996; Parkhe, 1993).

Changes in formal institutions can also be used to disrupt strong, but dysfunctional, informal ties between firms. For instance, Humphrey and Ashforth (2000, p. 719) document that U.S. automakers expressed “concerns that interpersonal relationships could lead buyers to award contracts to suppliers who had higher unit costs, lower quality and slower delivery times.” The adoption of competitive bidding – such as in the case of “business-to-business” exchanges through the Internet – is seen as an opportunity to circumvent reciprocal deals between sales and buying representatives which may not be in the best interests of one employer or the other. Therefore, since “the formal largely orders the direction the informal takes” (Dalton, 1959, p. 237), formal institutions constitute a tool available to managers through which informal institutions can be shaped. This supports:

Assumption 2: Formal institutions influence the trajectory of informal institutions.

The Nature of Formal and Informal Changes

Several organizational perspectives share the assumption that formal institutions involve *discrete* modes comprised of “bundles” of mutually consistent, complementary features. In discussing TCE, Williamson (1991, p. 271) stresses that “... each viable form of governance ... is defined by a syndrome of attributes that bear a supporting relation to one another.” Organizational economists Milgrom and Roberts (1991, p. 84) also submit that organizations involve activities that are “mutually complementary and so tend to be adopted together with each making the others more attractive.” For instance, centralization is characterized by structural interdependence between units, lower-powered incentives, and centralized decision making; decentralization is characterized by structural autonomy, higher-powered incentives, and local decision making. Each element reinforces the other: for instance, the use of higher-powered incentives is expected to discipline autonomous units to act efficiently, while autonomy supports those incentives since performance is assessed on an individual basis. On the other hand, the use of one element in isolation or in conjunction with another incompatible element will yield sub-optimal results. Thus, the adoption of higher-powered incentives jointly with centralized decision making is likely to trigger dysfunctional attempts by local managers to influence the central manager’s decisions or alter performance standards.

The assumption that organizational forms are discrete is also pervasive in organization theory. The configuration literature considers that organizational structures are composed of clusters of consistent traits (Meyer, Tsui & Hinings,

1993; Mintzberg, 1979). Miller and Friesen's (1980, p. 593) empirical study evidences that changes in organizational variables "tend to occur together, or . . . follow one another after a very brief interval, in order to maintain an appropriate balance or 'configuration'." Punctuated equilibrium models describing organizational change share the same perspective. Gersick (1991) uses the term "deep structure" to describe systems with distinct and interdependent parts, which change in an abrupt, comprehensive fashion rather than gradually. Likewise, Tushman and Romanelli (1985) consider that change is followed by periods of convergence that align the diverse activities within a firm into a consistent portfolio. Change (or reorientation) causes consistent changes of these activities toward a new alignment or deep structure, which clearly suggests discrete choices.

In contrast to formal organizational structures, which correspond to menus of discrete choices, informal elements are *continuously* arrayed. Social attachments, for instance, differ in degree rather than in kind. Thus, Granovetter (1973, p. 1361) defines tie strength as a combination of the "amount of time, the emotional intensity, the intimacy (mutual confiding), and the reciprocal services which characterize the tie." These elements clearly have a continuous flavor. Talking about individual commitment to an organization, Salancik (1977, p. 4) points out that "there are degrees of commitment [which derive] from the extent to which a person's behaviors are binding." Krackhardt (1990) uses the continuous measure of individual centrality in a social network to indicate the degree of an individual's power within an organization, derived from his or her ability to control information flows.

Thus, while changes in formal institutions are expected to involve discrete, abrupt movements of consistent variables, changes in informal institutions have a more continuous and gradual character. This leads to:

Assumption 3: Formal institutions are discretely arrayed, while informal institutions operate comparatively on a continuum.

The Pace of Formal and Informal Changes

Although changes in formal institutions trigger changes in informal institutions, the latter do not respond instantaneously. The concept of *inertia*, pervasive in organization theory, implies that webs of interdependent relationships, political coalitions, patterns of communication, established routines impede organizational change (Hannan & Freeman, 1984; March & Simon, 1958; Nelson & Winter, 1982; Tushman & Romanelli, 1985). The existence of inertia causes the functionality of an organizational form to change slowly, thereby

creating a lag between the implementation of a new formal structure and the change in the overall functionality, which derives in large part from informal elements.

The concept of inertia is also pervasive in NIE. Building upon Arthur's (1989) and David's (1985) ideas, North (1990) discusses how institutions exhibit *path dependence* in that the trajectory of an economic system is largely a function of its past position. North attributes a great deal of such path dependence to informal institutions derived from "available mental constructs – ideas, theories, and ideologies" (1990, p. 96), which create resistance to change.² In the same vein, Greif (1997, p. 89) argues that "past behavior, cultural beliefs, social structures, and organizations impact the development of values and social enforcement mechanisms that inhibit flexibility in departing from past patterns of behavior." A common theme in the NIE literature discussing economic change is that although it is easy to implement changes in formal institutions (laws, decision rights, etc.), existing informal institutions are difficult to disrupt, responding gradually and slowly to formal changes. Thus:

Assumption 4: Formal and informal institutions differ in the pace with which they change. Informal institutions possess inertia that slows the pace of change.

In the sections that follow, we discuss how the assumptions presented above imply three general propositions based on the relationship between informal and formal institutions:

Proposition 1: Formal and informal institutions are interdependent governance mechanisms in that the use of one mechanism can either promote (complement) or undermine (substitute for) the use of the other.

Proposition 2: Even in static environments, achieving the optimal functionality of an organizational form may require dynamic changes in formal institutions. Thus, under some circumstances, a pattern of vacillation in formal institutions supporting distinct organizational forms (market vs. hierarchy, centralized control vs. decentralized control, etc.) is warranted (Nickerson & Zenger, 2002).

Proposition 3: Firm boundaries are determined in large part by the need to adjust informal institutions within hierarchies. In particular, managers must sever the boundary of the firm to suspend dysfunctional informal processes.

Each general proposition is explained in turn. Taken together, these propositions exemplify how a more careful examination of the relationship between formal

and informal institutions may provide important insight into our understanding of organizational choice.

INFORMAL AND FORMAL INSTITUTIONS: COMPLEMENTS OR SUBSTITUTES?

If informal institutions influence the functionality of an organizational form (Assumption 1) and their trajectory is determined in part by formal institutions (Assumption 2), then a key question centers on the nature of relationship between these two mechanisms. At the most basic level, one must ask whether the use of one type of institution increases or decreases the functionality of the other – i.e. whether formal and informal institutions function as complements or substitutes. As it turns out, the literature focusing on this issue diverges along several paths.

Formal and Informal Institutions as Substitutes

Substitution arguments cluster around two basic claims. One perspective, which we refer to as *weak* substitution, argues that formal constraints are unnecessary because informal relationships based on trust and social norms can support cooperation without the costs and complexity incurred with formal agreements (Ellickson, 1991; Gulati, 1995; Powell, 1990; Ring & Van de Ven, 1994; Uzzi, 1996). Granovetter (1985, p. 489) contends that formal institutions “do not produce trust but instead are a functional substitute for it.” According to this view, social norms support the emergence of trust and, in the presence of trust, formal governance institutions are unnecessary. One of the most widely discussed social norms is *reciprocity*, meaning that individuals tend to cooperatively respond to generous offers even if reciprocity violates their own self interest (Berg, Dickhaut & McCabe, 1995; Dore, 1983; Rabin, 1993). Thus, weak substitution implies that the presence of informal institutions such as norms and trust removes the need for formal institutions.

A starker substitution perspective, which we refer to as *strong* substitution, argues that formal institutions are not only unnecessary, but also damaging to the formation and operation of informal elements. Macaulay (1963, p. 64) contends that “not only are contract and contract law not needed in many situations, their use may have, or may be thought to have, undesirable consequences . . . Detailed negotiated contracts can get in the way of creating good exchange relationships between business units.” He argues that some firms discourage the use of an elaborate contract because it “indicates a lack of trust

and blunts the demands of friendship, turning a cooperative venture into an antagonistic horsetrade" (Macaulay, 1963, p. 64).

Similarly, Sitkin and Roth (1993, p. 376) posit that "legalistic remedies can erode the interpersonal foundations of a relationship they are intended to bolster, because they replace reliance on an individual's 'goodwill' with objective, formal requirements." Ghoshal and Moran (1996) also stress that the use of rational, formal control has a pernicious effect on cooperation.³ They contend that for those parties being controlled . . .

. . . the use of rational control signals that they are neither trusted nor trustworthy to behave appropriately without such controls . . . For the controller, negative feelings arise from what Strickland (1958) described as 'the dilemma of the supervisor' viz., the situation when the use of surveillance, monitoring, and authority led to management's distrust of employees and perceptions of an increased need for more surveillance and control . . . (1996, p. 24).

Social psychologists have provided an explanation for this effect: explicit incentives or punishments may reduce partner's intrinsic motivation to perform certain tasks (Deci & Ryan, 1985). This effect has received the name *motivation crowding out* in economics (Frey, 1997). According to proponents of motivation crowding out theory, reciprocity is a particular form of intrinsic motivation originated from social norms that is violated when formal incentives or punishments are present (Gächter & Falk, 2000). In other words, those mechanisms can signal that trust is absent and no reciprocity is expected, thereby framing the relationship in a strictly economic, rather than social, orientation (Lubell & Scholz, 2001; Tenbrunsel & Messick, 1999). One possible consequence is that the outcomes achievable through incentives and controls can be less efficient than those that could naturally flow from an individual's voluntary willingness to cooperate, manifested through social norms and trust.

Formal and Informal Institutions as Complements

An alternative argument that has received comparatively less attention is that formal institutions *complement* informal mechanisms. In settings where hazards are severe, the combination of formal and informal safeguards may deliver greater functionality than either institutional type in isolation. As North (1990, pp. 46–47) puts it, "formal rules can complement and increase the effectiveness of informal constraints. They may lower information, monitoring, and enforcement costs and hence make informal constraints possible solutions to more complex exchange." Poppo and Zenger (forthcoming) provide supportive empirical evidence. They find that customized contracts and relational governance, characterized by alignment of goals, trust and collaborative

orientation, appear to operate as complements in generating improvements in exchange performance.

Three distinct arguments support the complementarity proposition. First, formal contracts may both extend the expected duration of a relationship and restrict the gains from one-time deviations from cooperative behavior in an exchange relationship (Baker, Gibbons & Murphy, 2002). Contracts not only have this source of advantage because of their formal specification of a long-term commitment to exchange, but because they limit the domain and lessen the gain of potential opportunistic behavior through clearly articulated clauses that specify punishments. This reduction in short run gains heightens comparatively the gains from cooperating in the exchange relationship. By contrast, failing to contractually specify elements of the exchange that are easily specified merely heightens incentives for short-run cheating and lowers expectations of cooperation (Baker, Gibbons & Murphy, 1994; Klein, 1996; Milgrom, North & Weingast, 1990). Note that complementarity arguments assume that formal institutions are to some extent *incomplete*, since otherwise any outcome could be legally enforced without the need for informal institutions. Due to the costs to write clauses, limits of enforceability by courts, and individuals' cognitive limitations (bounded rationality), it is not possible for parties to pre-specify all future contingencies in a comprehensive contract.⁴

Lazzarini, Miller and Zenger's (2001) experiment provides support for the complementarity argument outlined above. Their experiment involves buyer-seller exchanges with moral hazard on the part of sellers. To operationalize contract incompleteness, the authors consider that the good being transacted has two distinct dimensions. One dimension is easy to specify and therefore is contractible in advance: buyers can structure contingent payments based on the supplied level of that dimension. The other dimension is difficult to measure and enforce by third parties (such as courts), and thus is non-contractible: no contingent payment can be applied. Lazzarini, Miller and Zenger (2001) find that contractual incentives (contingent payments) applied to the contractible exchange dimension facilitate the enforcement of the *non*-contractible dimension, precisely because they limit the gains that sellers could attain from short-term defection.

The second argument providing support for the complementarity view is that formal institutions can set the stage for the development of trust within a long-term interaction. This is a direct implication of Assumption 2: formal mechanisms can influence the trajectory of informal elements. Cooperative behavior in the present – as a result of supporting formal mechanisms – reinforces an expectation of cooperation in the future. Supportive of this logic, empirical work suggests that past success in contracting with a particular

exchange partner yields greater success in the present (Anderson & Weitz, 1989; Larson, 1992). Formal contracts help ensure that the early, more vulnerable stages of exchange are successful. Durkheim (1933) appears to have this idea in mind when he writes that:

... in order for [parties] to co-operate harmoniously, it is not enough that they enter into a relationship, nor even that they feel the state of mutual dependence in which they find themselves. It is still necessary that the conditions of this cooperation be fixed for the duration of their relations. The rights and duties of each must be defined, not only in view of the situation such as it presents itself at the moment when the contract is made, but with foresight for the circumstances which may arise to modify it. Otherwise, at every instant, there would be conflicts and endless difficulties (1933, pp. 212–213).

Thus, formal institutions may also be designed to create procedures to adapt to changing conditions. Unexpected disturbances may place considerable strain on an exchange relationship (Williamson, 1991, pp. 271–273). Formal contracts that shift from merely specifying deliverable outcomes to providing frameworks for bilateral adjustments may facilitate the evolution of highly cooperative exchange relations. Crocker and Masten (1991, p. 95) suggest that “it seems more appropriate to view contracts as means of establishing procedures for adapting exchange and resolving disputes rather than purely as incentive mechanisms.” In addition, the process of contracting may itself promote expectations of cooperation consistent with relational governance. The activity of creating complex contracts requires parties to *mutually* determine and commit to processes for dealing with unexpected changes, penalties for non-compliance, and other joint expectations of trade. Hence, the process of developing complex contracts in response to exchange hazards positively affects future exchange performance through the development of social relations resulting from the very act of bilaterally negotiating contract terms.

The third argument supporting the complementarity view is that informal elements may also promote the refinement (and hence increased complexity) of formal institutions. As discussed before, informal institutions increase the performance of formal institutions because explicit arrangements are inherently incomplete. Thus, not only do formal institutions promote the stability of informal institutions, but informal institutions also play a role in filling contractual gaps over time. As a close relationship is developed and sustained, lessons from the prior period are reflected in revisions of the contract. Exchange experience, patterns of information sharing, evolving performance measurement and monitoring may all enable greater specificity (and complexity) in contractual provisions and exchange conditions. As a consequence, relational exchanges may gradually develop more complex formal contracts, as mutually agreed upon processes become formalized.

An Integrated Assessment

We note that the substitution and complementarity effects described above are not mutually exclusive; the interaction between formal and informal institutions is more complex. For instance, even if explicit incentives or control mechanisms reduce individuals' intrinsic motivation to provide extra effort, they can *at the same time* discourage short-term defection. Thus, the effects described above can be best viewed as *partial* effects; the net outcome is dependent on particular exchange conditions. Although research on this topic is in its early stages, we tentatively outline some specific propositions. We submit that formal and informal institutions act more as complements than substitutes when: (1) individuals are not likely to or not committed to transacting repeatedly; and (2) the procedures involved in the operation of formal institutions are perceived to be "fair". When these conditions do not hold, the use of formal institutions may be unnecessary and even detrimental to the operation of certain informal institutions.

The reason for our first claim is that non-repeated interactions provide neither a "shadow of the future" increasing individuals' perceived benefits from co-operation (Axelrod, 1984) nor a "shadow of the past" promoting the gradual development of relational norms and trust (Macneil, 1978; Ring & Van de Ven, 1994). Hence, the benefit of formal institutions becomes relatively more important in new or non-recurring relationships, since informal enforcement will tend to be weak or absent (Lazzarini et al., 2001). One could argue that, instead of relying on formal structures to support non-repeated exchanges, parties might be better off encouraging repeated interaction as a way to create norms and trust (e.g. Dyer & Singh, 1998; Kollock, 1994; Krackhardt, 1992). However, parties must *credibly* commit to a repeated interaction since at any moment they can switch to alternative partners. Formal arrangements are thus a way to lock parties into relationships with sufficient duration (Baker et al., 2002). In addition, repeated interaction between the same agents limits the opportunities and information that they can attain with external relations (Blau, 1964). Non-recurring exchanges or "weak ties" are a fundamental way to transfer new information and knowledge between specialized agents (Granovetter, 1973). By contrast, "overembedded" systems are likely to involve low knowledge diversity and hence less propensity to innovate (Greif, 1997; Uzzi, 1996). Formal institutions are thus crucial for economic growth marked by specialization and fewer recurring exchanges (North, 1990; Zucker, 1986).

Our second claim derives from the growing recognition that individual attitudes are dependent on *procedural* issues, which alter their perceptions about the fairness of processes employed by parties to achieve certain outcomes (Bies

& Shapiro, 1988). Consider for instance the effect of contract “framing”: empirical studies provide evidence that people tend to prefer *equivalent* contracts that specify rewards or “bonuses” rather than punishments or “damages” (Luft, 1994; McLean, Parks & Coelho-Kamath, 1999). It is possible that equivalent contracts stipulating bonuses instead of punishments discourage defection without crowding out implicit norms. In addition, the procedural perceptions of formal institutions are dependent on the extent to which parties expect the use of these institutions for certain types of exchanges, i.e. whether their use is “taken for granted” or not. Thus, Lewicki, McAllister and Bies (1998, p. 454) argue that “quality control people do their work because it’s their job – not necessarily because they personally distrust others.” Also, two firms engaging in an alliance may employ a formal contract not because they do not trust one another in particular, but because it is a standard procedure in their industry. Thus, whether a formal institution complements or substitutes for informal institutions depends in part on how fair that formal institution is perceived.

In conclusion, formal and informal institutions are not merely alternative ways to govern exchanges. In most cases they are employed simultaneously and interact in complex ways. The complete assessment of complementarity and substitution effects, and the conditions under which one effect supplants the other, are an important research agenda both within and outside NIE.

THE DYNAMIC ALIGNMENT OF ORGANIZATIONAL FORMS

Theories within NIE argue that the choice of organizational forms involves matching formal structures to strategies, exchange conditions, and environments in some discriminating way (Chandler, 1962; Williamson, 1975). NIE is fundamentally a theory of static alignment and, as such, shares many of the elements of “contingent-fit” models. Namely, selection and external pressures (e.g. from owners and capital markets) will either prompt managers to choose organizational forms that are aligned to particular exchange conditions, or wash out misaligned forms. Thus, managers choose to govern exchanges through markets when outputs are easily measured (Barzel, 1982; Holmstrom & Milgrom, 1994; North, 1981), and the production of those outputs involves low levels of specialized assets (Klein, Crawford & Alchian, 1978; Williamson, 1985). By contrast, managers choose hierarchy when outputs are difficult to measure and specialized assets are substantial. Intermediate conditions may favor “hybrid” organizational forms such as long-term contracts and interfirm

alliances (Williamson, 1991). Similar discriminating alignment logic governs the choice of governance forms within the firm. Thus, managers may choose decentralized forms with their incumbent high-powered incentives when innovation is desired, but adopt centralization with its ready access to authority when coordination is desired (Williamson, 1985).

Static alignment implies that if the diverse contingencies that affect firms remain stable, then organizational forms should also remain unchanged. Changes in formal mechanisms are precipitated by changes in environment or exchange conditions, which are influenced in part by strategy decisions. The patterns of change in formal governance that we commonly observe, however, are often difficult to reconcile with this theory of static alignment. Many changes in formal governance appear to occur with little change in environment or exchange conditions. Indeed, we seem to often observe firms engaged in *vacillating* patterns of choice in formal organization (Carnall, 1990, p. 20; Cummings, 1995, p. 112; Eccles & Nohria, 1992, p. 127; Mintzberg, 1979, p. 294; Nickerson & Zenger, 2002). Thus, firms centralize, then decentralize, then centralize, etc. Over a sixteen-year period, Hewlett-Packard, for instance, made six fundamental shifts between centralizing and decentralizing core activities within the firm. Similar processes are observed in sourcing decisions. Firms outsource an activity, then internalize it, only to outsource it, again. In the choice of compensation, many firms seem to vacillate between aggressive incentive pay for sales personnel and flat salaried pay.

Firms may also cycle among more than two forms. Consider KPMG Peat Marwick. Prior to 1992, the consulting and accounting services firm was structured geographically around local managing partners, which helped to leverage close ties with regional clients into a broad range of services. In 1992, KPMG shifted to a functional structure where associates reported to nationwide practice managers rather than local managers. This structure promoted knowledge sharing and resource allocation within particular consulting and accounting practices. In 1994, KPMG shifted again its organizational configuration to an industry-focused structure based on categories such as healthcare, government services, retail, and manufacturing. This new structure promoted the development of in depth knowledge about client industries. In 1996, KPMG returned to its former geographic structure centered on local managing partners. However, by 1999 they shifted back to an industry-focused industry, although now globally centralized. Within seven years, KPMG had cycled among three (discrete) formal structures – geographic, functional, and industry-focused – with five events of structural change. How do we explain this pattern, since it is very unlikely that within this short period KPMG had faced so many changes in exchange and environmental conditions?

Arguably, such vacillation (or cycling) in formal organizational modes can result from managers' fickle behavior, thereby being simply a manifestation of noise in the decision-making process. However, as Nickerson and Zenger (2002) maintain, a richer understanding of the relationship between formal and informal institutions leads to the conclusion that such commonly observed vacillation can indeed be efficient. The explanation for functional vacillation follows logically from the assumptions described in Section 2.

As indicated by Assumption 1, the functionality of an organizational form is determined as much by informal mechanisms, as by formal mechanisms. In addition, informal institutions vary systematically in response to changes in formal institutions (Assumption 2). But although managers can influence changes in informal institutions, they only have discrete formal choices or "levers" – e.g. centralization vs. decentralization, make vs. buy, etc. – to promote such changes (Assumption 3). If the desired level of functionality – which is largely dependent on informal institutions – lies in between the functionality delivered by these formal levers and change is not too costly, then managers will have an incentive to modulate between two or more discrete formal choices to achieve *temporarily* the desired intermediate level. Given that informal institutions display inertia (Assumption 4), each switch between formal choices triggers a gradual change in the trajectory of informal elements. Thus, by vacillating between distinct formal choices, managers can influence the trajectory of informal institutions towards a desired position that is unavailable if the organization remains fixed with a particular formal structure.

The choice between centralization and decentralization is a useful illustration of the virtues of vacillation. Centralization and decentralization are discrete organizational modes characterized by distinct sets of formal institutions. While centralization involves structural interdependence between units, lower-powered incentives, and centralized authority, decentralization involves structural autonomy, higher-powered incentives, and local authority. These organizational modes also exhibit distinct and conflicting patterns of functionality. Centralization facilitates coordination, but at the cost of low-powered incentives and reduced innovation potential. Decentralization yields high-powered incentives and increased innovation, but at the cost of coordination. Managers would like to maintain a level of functionality such that improved coordination *and* higher-powered incentives/innovation coexist; by choosing permanently either centralization or decentralization, managers will necessarily sacrifice one of these dimensions of functionality. They can, however, dynamically modulate between these two formal structures to achieve temporarily a level of functionality that lies in between centralization and decentralization.

Thus, a decision to change from centralization to decentralization triggers changes in informal elements that alter the functionality of the organizational form. As managers initiate decentralization and the firm begins to reap benefits from higher-powered incentives and innovation, it still enjoys, albeit temporarily, the dense communication channels and social attachments supported by the informal institutions (e.g. routines, norms, etc.) that accompanied the formerly centralized structure. However, these patterns will tend to diminish as time elapses, since the formal change to decentralization will sever social attachments and communication flows within the firm, which in turn will dampen coordination. As a result, the overall organizational performance will migrate towards the steady-state functionality delivered by decentralization. When this occurs, managers can do the reverse: they can change the formal structure from decentralization to centralization to restore coordination while keeping some innovation and incentives reminiscent from the decentralized structure. However, after some time, sticky routines and excessive politicking promoted by centralization will again cause a reduction in functionality. Managers will then need to initiate a new cycle by decentralizing the organization in order to alter these informal elements.

Nickerson and Zenger (2002) show that this theory provides a counterintuitive result that *inertia can be performance enhancing*. The reason is that inertia reduces the frequency with which managers must change the formal structure. When inertia is high, informal institutions change slowly in response to changes in formal institutions. Since informal elements are critical determinants of organizational functionality, the latter will remain close to the optimal level – which will combine elements of both discrete forms – for a long period of time before a new switch is necessary. Were changes in informal institutions instantaneous (i.e. no inertia), it would not be possible to keep informal institutions at the desired, intermediate level for a time period sufficient to warrant vacillation: they would quickly converge to the steady-state position of the chosen formal structure

Thus, the manager's task is not simply to observe changes in environment or exchange conditions, but rather to monitor the trajectory of informal institutions and manipulate them indirectly through changes in formal structures. Vacillation between organizational forms comprised of discrete formal institutions is efficient when the costs of change are moderate and the desired functionality lies intermediate to that delivered by either formal form in steady state, since informal institutions will lag formal changes and therefore will stay temporarily at that intermediate position. Hence, taking informal institutions seriously has in this case led us to a conclusion that contradicts a fundamental proposition from NIE (and TCE in particular): rather than

statically align institutions to exchange or environmental conditions, under certain circumstances, managers must pursue a dynamic alignment by vacillating between discrete formal institutions. In this sense, changes in formal structures can occur even when exchange or environment contingencies remain unchanged.

INFORMAL INSTITUTIONS AND FIRM BOUNDARIES

The prototypical institutional choice in NIE is the choice between market governance and hierarchical governance. This is fundamentally viewed as a choice between two formal institutions. The standard story – derived from Coase’s (1937) insight – is that the institution of hierarchy is chosen when markets fail. Thus, we have well developed explanations for why markets fail and therefore why managers choose to replace markets with internal organization. We also have a large body of confirmatory empirical evidence: managers appear to choose hierarchy when exchange conditions present hazards in using markets and choose markets when they do not (e.g. Poppo & Zenger, 1998; Shelanski & Klein, 1995). Managers choose hierarchy because hierarchy possesses governance features to which markets have limited access. Namely, hierarchy’s low-powered incentives discourages expropriation of rents in the presence of specific assets (Williamson, 1985) and avoids dysfunctional responses to incentives when performance attributes are difficult to measure (Holmstrom & Milgrom, 1994).

However, as numerous scholars have noted, why managers choose markets is not as well understood. Parallel logic suggests that markets are chosen when hierarchies as institutions fail. Interestingly, Coase (1937) in his seminal paper asks:

... why, if by organizing one can eliminate certain costs and in fact reduce the cost of production, are there any market transactions at all? Why is not all production carried on by one big firm? (1937; reprinted in 1993, p. 23).

Coase tentatively answers this question by invoking, among other things, the nebulous concept of “diminishing returns to management,” prevalent in early industrial organization writings, which asserts that managers have limited ability to coordinate large flows of resources. This explanation is not satisfactory because, for instance, one could solve the problem of diminishing returns by splitting the firm into smaller independent units, and then using internal markets to allocate resources. At a more fundamental level, we must ask why hierarchies

cannot selectively use the high-powered incentives of markets and thereby do all that markets can and more (Williamson, 1985, p. 133). What are the limits to using market-like instruments to reduce the deficiencies of hierarchies?

Within this perspective, Williamson (1985) provides a more convincing explanation of the limits of firms: it is not possible to “selectively intervene” inside hierarchies by infusing market-like incentives without incurring additional costs or creating undesirable side effects. He notes, for instance, that performance indicators can be manipulated, and incentives (such as piece rates) can lead to the overutilization of a firm’s assets by employees. Williamson (1985, p. 142) also briefly mentions the role of fairness considerations in dictating the allocation of gains and losses inside firms, which are rooted in informal institutions. This, however, is only the tip of the iceberg. We submit that a thorough understanding of informal institutions within organizations is critical to understanding the limits of the firm and to developing a theory of firm boundaries.

The basic argument is as follows. A key reason why hierarchies reduce market failure is that they trigger the formation of informal institutions – norms, routines, organizational culture, etc. – which affect organizational functionality (Assumption 1) by facilitating communication, coordination and cooperation (Barnard, 1938; Kogut & Zander, 1996). However, such informal institutions also bring side effects. For instance, social attachments cause biased decision making, and firm-specific routines constrain the ability of those within the firm to externally communicate and acquire external knowledge. Thus, the problem that managers of hierarchy face is that the informal institutions which hierarchy triggers cannot be selectively shut down with any great success. Consequently, to suspend these informal processes, managers must shift or constrain the boundaries of the firm. Since these informal processes are influenced by formal decisions (Assumption 2), we contend that firms can adjust their formal boundaries to alter the dynamics of informal institutions. Thus, firms suspend hierarchy as an institution (i.e. sever the organizational boundary) to avoid the informal processes that run rampant within their boundaries. This shows that, in taking seriously the concept of informal institutions, we are able to develop a theory of *hierarchical failure* – a theory that allows us to explain why firms constrain their boundaries. While informal features of hierarchies reduce market failure, they create costs that need to be factored in boundary choices.

We identify here four critical informal processes, largely based on informal institutions, creating hierarchical failure: (1) influence activities; (2) social attachments; (3) social comparison processes; and (4) development of firm-specific routines. We discuss each in turn.

Influence Activities

Milgrom and Roberts (1990) argue that influence activities – attempts to influence the allocation of rents within firms in order to preferentially reward particular individuals or coalitions – are the primary costs of internal organization. Hierarchies not only create an environment where individuals can engage in lobbying to distort the allocation of resources, since exchanges are not disciplined by market forces, but also magnify the extent to which such activities are feasible. This is because the increased communication channels within hierarchies, although instrumental in facilitating coordination (Eccles & White, 1988; Pfeffer, 1978), represent ways in which individuals can reach and influence decision makers. In this sense, politicking becomes by itself a fundamental informal institution governing and affecting the functionality of hierarchies.

Poppo (1995) provides evidence on how influence activities have implications for boundary decisions. She finds that although hierarchy improves coordination by facilitating the exchange of information among internal units, it engenders difficulty in negotiating internal (transfer) price adjustments due to costly bargaining among divisional managers. Consequently, since the coordination-based benefits of hierarchy necessarily unleash influence activities, firms must constrain their boundaries to interrupt communication channels that facilitate such dysfunctional political behavior.

Social Attachments

Similar to influence activities, social attachments can distort the allocation of resources within firms, which in turn affects the costs of internal organization. Thus, poor decision and resource misallocation may occur even in the absence of lobbying efforts or other influence activities. Such social attachments are largely governed by a host of informal institutions. Decision makers may overfund projects or divisions with rather limited promise in an act of reciprocity to friends/managers, or underfund projects and divisions with more substantial promise but involving managers with whom they lack such social attachments. Evidence also suggests that reciprocity norms embedded in friendship ties can reduce individuals' willingness to negotiate freely and pursue better opportunities (Halpern, 1994). In addition, excessive socialization can induce a bias towards shared perspectives, values, and culture, thereby undermining the firm's ability to find innovative solutions (Gruenfeld, Mannix, Williams & Neale, 1996; Katz, 1982). Thus, to the extent that social attachments within

firms lead to biased decision making, managers may wish to limit their boundaries in order to reduce the reach and consequences of these distortions. Resource allocation decisions governed through the market are likely to be less contaminated by such distortions, precisely because they will be relatively less affected by (even though not completely insulated from) deep social attachments.

Social Comparison Processes

The social norm of equity in the allocation of rewards creates difficulty in the design of compensation schemes within firms (Adams, 1965; Deutsch, 1985). Employees directly, or indirectly through managers, compare their rewards to all other employees within the boundary of the firm. When they perceive rewards to be inequitably distributed – i.e. when they consider that their compensation to effort ratio is lower than other colleagues – they reduce effort, seek to alter the distribution, or simply depart the firm (Adams, 1965). All such outcomes are costly to the firm. The challenge that a manager faces in rewarding employees is that employees possess highly inflated perceptions of their own performance, exacerbated by the fact that the manager does not possess a fully accurate measure of performance. Hence, efforts to aggressively reward performance in the absence of accurate performance measures trigger social comparison processes that impose costs upon the firm. In response to these social comparison processes triggered by equity norms, managers simply adopt low-powered performance incentives. However, this brings two side effects (Zenger, 1994). First, low-powered incentives are likely to reduce employees' effort compared to the situation involving high-powered incentives. Second, other firms offering contracts with a closer match between pay and performance will lure more skilled people, since the latter will be able to reap higher rents from their superior talent. Thus, firms offering low-powered incentives are likely to face not only turnover costs, but also the departure of skilled people to other firms (Zenger, 1992).

Constraining the boundaries of the firm and more specifically reducing the size of the firm constrains the scope of social comparison processes. The costs of social comparison that are associated with a high-powered incentive scheme are strongly dependent on the size of the firm. In large firms there is a much larger group of individuals who will respond negatively when a colleague within the firm is granted a significant performance-based increase in pay.⁵ In small firms, the number of comparisons triggered by a single adjustment in pay is substantially less. Further, in small firms, information about individual performance levels is more easily disseminated. Consequently, a reduction in

firm size is likely to enhance consensus about individuals' perceived relative performance and hence reduce the costs created by social comparisons (Zenger, 1992). It follows that small firms will tend to offer higher-powered incentives than large firms, since they will face lower inequity perceptions resulting from such incentive policies. For this reason, managers may sever the boundaries of the firms to avoid, at least in part, social comparison processes created by equity norms within hierarchies.

Firm-specific Routines

Firm-specific routines develop as a by-product of repeated interaction within firms, and represent informal institutions in the form of tacit codes of interaction (Nelson & Winter, 1982). While some consider such common vocabulary and procedures as enhancing communication (Allen & Cohen, 1969; Lawrence & Lorsch, 1967), it does not come without cost. Tushman and Katz (1980) point out that:

The evolution of local languages and coding schemes helps the unit deal with its local informal processing requirements; yet, it also hinders the unit's acquisition and interpretation of information from external areas. External communication is vital, however, both in terms of feedback and for evaluating and acting on the unit's environment (1980, p. 1072).

Thus, idiosyncratic routines can lock partners into one particular field of knowledge and lock them out of external opportunities and sources of information (Cohen & Levinthal, 1990; Leonard-Barton, 1995; Poppo & Zenger, 1998). Employees sharing common vocabulary and interpretations are not only likely to have limited capacity to interpret external sources of information, but also to recognize the importance of those sources (Levinthal & March, 1993). As Katz (1982) explains,

One of the main principles of human communication...is the strong tendency for individuals to communicate with others who are most like themselves, or who are most likely to agree with them. Over time, project members learn to interact selectively or avoid messages and information that might conflict with their established practices and dispositions, thereby reducing their overall levels of outside contact (1982, pp. 84–85).

A manager determining firm boundaries must take this effect into account. Theoretical approaches emphasizing the firm as a repository of routines (Kogut & Zander, 1996; Nelson & Winter, 1982), must also consider that these informal elements put limits on an organization's ability to acquire external information and adapt to changing circumstances (Leonard-Barton, 1995). Faced with these two opposing effects, managers will need to monitor the development of

firm-specific routines and, if necessary, expose internal units to external sources of information by severing existing boundaries.

The discussion above shows that informal institutions within firms are fundamental in determining the limits of firms. While certain informal institutions solve market failure within hierarchies, they also bring their own costs. Thus, informal channels improve coordination but create avenues for influence activities; social norms create trust and facilitate cooperation but induce biased decision-making and trigger social comparison processes; firm-specific routines are instrumental in facilitating the internal exchange of information but lock the firm out of external sources of information. The paradox is that the same informal processes that help to mitigate market failure create hierarchical failure once they are internalized within firms. We stress that a careful examination of such processes is necessary to develop a more complete theory of the firm – a theory not only based on how firms expand their boundaries in the presence of market failure, but also how they sever their boundaries in the presence of hierarchical failure.

CONCLUDING REMARKS

NIE has contributed to our understanding of how formal institutions such as explicit incentives, authority, contracts, and ownership can be aligned to exchange conditions to increase governance efficiency. Organization theory and economic sociology, in turn, have provided a deep examination of the role of informal institutions such as social norms, trust, routines, and political processes. However, by focusing on each type of institution in isolation, scholars have not paid sufficient attention to the interplay between formal and informal institutions and its performance implications. It is our goal in this paper to demonstrate how the joint assessment of the role of formal and informal mechanisms governing exchanges within and between firms provides new insights and expands the explanatory power of existing theories of organization.

The insights generated by this line of research have important implications for business strategy. It is widely recognized that informal institutions have a critical role in the performance of organizations. Informal institutions can be either performance enhancing, such as relational governance, or performance damaging, such as influence activities. In some cases, the *same* informal institution can promote or undermine performance depending on the circumstances; for instance, firm-specific routines can improve coordination but at the same time reduce an organization's ability to respond to external changes. Thus, it is fundamental that scholars not only outline the benefits or drawbacks of informal mechanisms, but also inform managers about how to adjust them.

In most cases, managers have only formal mechanisms at their disposal to change the trajectory of informal processes. Thus far, theories of organization have provided insufficient guidance on how informal institutions can be shaped through changes in formal institutions or by other means.

The good news is that there is still much research needed in understanding the interaction between formal and informal institutions and exploring how this articulation can potentially deliver superior performance for those who manage institutions. A more extensive collaboration among disciplines that have traditionally focused on each type of institution will certainly be a necessary step in developing this stream of research.

NOTES

1. Davis and North (1971, pp. 6–7) distinguish between institutions related to the political, legal and social (broadly defined) environment of an economic system, and institutions related to arrangements “between economic units that govern the ways in which these units can cooperate and/or compete.” In this paper, we focus on the latter, which relates to the concept of governance (Williamson, 1991) defined as the “*institutional* matrix in which the integrity of a transaction is decided” (Williamson, 1996, p. 378, emphasis added). Governance mechanisms, which are comprised of formal and informal institutions, support organizational forms for the production and/or exchange of assets.

2. See also Denzau and North (1994).

3. Interestingly, Williamson (1996, p. 271) makes a similar argument, but restricts its application to purely social, non-economic relationships.

4. For a recent assessment of the incomplete contracting literature, see Tirole (1999).

5. Such logic may explain the tendency for increases in plant size to cause the pay of all employees to rise even those in secondary jobs which are rather easily filled in the labor at well established market wages. Concerns about the fallout from negative perceptions of equity result in rising pay (Rebitzer & Robinson, 1991).

ACKNOWLEDGMENT

We particularly thank Jackson Nickerson whose collaborative work with the first author forms the basis for the section on the dynamic alignment of organizational forms. We also acknowledge the helpful comments of participants at the Workshop on New Institutionalism in Strategy Research held at Columbia. Finally, we thank Brian Silverman and Paul Ingram for helpful editorial advice and the discipline they provided with a long stream of reminders.

REFERENCES

- Adams, J. S. (1965). Inequity in social exchange. In: L. Berkowitz (Ed.), *Advances in Experimental Social Psychology*. New York: Academic Press.

- Allen, T. J., & Cohen, S. (1969). Information flows in R&D labs. *Administrative Science Quarterly*, 14, 12–19.
- Anderson, E., & Weitz, B. (1989). Determinants of continuity in conventional industry channel dyads. *Marketing Science*, 8, 310–323.
- Argote, L. (1999). *Organizational learning: creating, retaining and transferring knowledge*. Boston: Kluwer Academic Publishers.
- Arthur, W. B. (1989). Competing technologies, increasing returns, and lock-in by historical events. *The Economic Journal*, 99, 116–131.
- Axelrod, R. (1984). *The evolution of cooperation*. New York: Basic Books.
- Baker, G., Gibbons, R., & Murphy, K. J. (1994). Subjective performance measures in optimal incentive contracts. *The Quarterly Journal of Economics*, 109, 1125–1156.
- Baker, G., Gibbons, R., & Murphy, K. J. (2002). Relational contracts and the theory of the firm. *The Quarterly Journal of Economics*, 117, 39–83.
- Barnard, C. (1938). *The functions of the executive*. Cambridge: Harvard University Press.
- Barzel, Y. (1982). Measurement costs and the organization of markets. *Journal of Law and Economics*, 25, 27–48.
- Berg, J., Dickhaut, J., & McCabe, K. A. (1995). Trust, reciprocity, and social history. *Games and Economic Behavior*, 10, 122–142.
- Bies, R. J., & Shapiro, D. L. (1988). Voice and justification: their influence on procedural fairness judgments. *Academy of Management Journal*, 31, 676–685.
- Blau, P. (1964). *Exchange and power in social life*. New York: Wiley.
- Brass, D. J. (1984). Being in the right place: a structural analysis of individual influence in an organization. *Administrative Science Quarterly*, 29, 518–539.
- Carnall, C. A. (1990). *Managing change in organizations*. London: Prentice Hall.
- Chandler, A. D. (1962). *Strategy and structure*. Cambridge: MIT Press.
- Coase, R. H. (1937). The nature of the firm. *Economica N.S.*, 4, 386–405. (Reprinted in: O. E. Williamson & S. G. Winter (Eds), *The Nature of the Firm: Origins, Evolution, Development* (pp 18–33). New York: Oxford University Press, 1993).
- Cohen, W. M., & Levinthal, D. A. (1990). Absorptive capacity: a new perspective on learning and innovation. *Administrative Science Quarterly*, 35, 128–152.
- Crocker, K. J., & Masten, S. E. (1991). Pretia ex machina? Prices and process in long-term contracts. *Journal of Law & Economics*, 34, 69–99.
- Crozier, M. (1964). *The bureaucratic phenomenon*. Chicago: University of Chicago Press.
- Cummings, S. (1995). Centralization and decentralization: the never-ending story of separation and betrayal. *Scandinavian Journal of Management*, 11, 103–117.
- Dalton, M. (1959). *Men who managed*. New York: John Wiley and Sons.
- David, P. (1985). Clio and the economics of QWERTY. *The American Economic Review*, 75, 332–337.
- Davis, L. E., & North, D. C. (1971). *Institutional change and American economic growth*. Cambridge: Cambridge University Press.
- Deci, E. L., & Ryan, R. M. (1985). *Intrinsic motivation and self-determination in human behavior*. New York: Plenum.
- Denzau, A. T., & North, D. C. (1994). Shared mental models: ideologies and institutions. *Kyklos*, 47, 3–30.
- Deutsch, M. (1985). *Distributive justice: a social-psychological perspective*. New Haven: Yale University Press.
- Dore, R. (1983). Goodwill and the spirit of market capitalism. *The British Journal of Sociology*, 34, 459–482.

- Doz, Y. L. (1996). The evolution of cooperation in strategic alliances: Initial conditions or learning processes? *Strategic Management Journal*, 17, 55–83.
- Durkheim, E. (1933). *The division of labor in society*. Glencoe: The Free Press.
- Dyer, J. H., & Singh, H. (1998). The relational view: cooperative strategy and sources of interorganizational competitive advantage. *Academy of Management Review*, 23, 660–679.
- Eccles, R. G., & Nohria, N. (1992). *Beyond the hype: rediscovering the essence of management*. Cambridge: Harvard Business School Press.
- Eccles, R. G., & White, H. (1988). Price and authority in inter-profit center transaction. *American Journal of Sociology*, 94, 17–51.
- Ellickson, R. C. (1991). *Order without law: how neighbors settle disputes*. Cambridge: Harvard University Press.
- Ensminger, J. (1997). Changing property rights: reconciling formal and informal rights to land in Africa. In: J. N. Drobak & J. V. C. Nye (Eds), *The Frontiers of the New Institutional Economics* (pp. 165–196). San Diego: Academic Press.
- Frey, B. S. (1997). A constitution for knaves crowds out civic virtues. *Economic Journal*, 107, 1043–1053.
- Gächter, S., & Falk, A. (2000). *Work motivation, institutions, and performance* (Working paper). Zurich: Institute for Empirical Research in Economics, University of Zurich.
- Gersick, C. J. G. (1991). Revolutionary change theories: a multilevel exploration of the punctuated equilibrium paradigm. *Academy of Management Review*, 16, 10–36.
- Ghoshal, S., & Moran, P. (1996). Bad for practice: a critique of the transaction cost theory. *Academy of Management Review*, 21, 13–47.
- Granovetter, M. (1973). The strength of weak ties. *American Journal of Sociology*, 78, 1360–1380.
- Granovetter, M. (1985). Economic action and social structure: the problem of embeddedness. *American Journal of Sociology*, 91, 481–510.
- Greif, A. (1997). On the interrelations and economic implications of economic, social, political, and normative factors: reflection from two late medieval societies. In: J. N. Drobak, & J. V. C. Nye (Eds), *The Frontiers of the New Institutional Economics* (pp. 57–94). San Diego: Academic Press.
- Gruenfeld, D. H., Mannix, E. A., Williams, K. Y., & Neale, M. A. (1996). Group composition and decision making: how member familiarity and information distribution affect process and performance. *Organizational Behavior and Human Decision Processes*, 67, 1–15.
- Gulati, R. (1995). Does familiarity breed trust? The implications of repeated ties for contractual choice in alliances. *Academy of Management Journal*, 38, 85–112.
- Halpern, J. J. (1994). The effect of friendship on personal business transactions. *Journal of Conflict Resolution*, 38, 647–664.
- Hannan, M. T., & Freeman, J. (1984). Structural inertia and organizational change. *American Sociological Review*, 49, 149–164.
- Hart, O. D. (2001). *Norms and the theory of the firm*. Harvard Institute of Economic Research Discussion Paper.
- Holmstrom, B., & Milgrom, P. (1994). The firm as an incentive system. *American Economic Review*, 84, 972–991.
- Humphrey, R. H., & Ashforth, B. E. (2000). Buyer-supplier alliances in the automobile industry: how exit-voice strategies influence interpersonal relationships. *Journal of Organizational Behavior*, 21, 713–730.
- Katz, R. (1982). The effects of group longevity on project communication and performance. *Administrative Science Quarterly*, 27, 81–104.

- Klein, B. (1996). Why hold-ups occur: the self-enforcing range of contractual relationships. *Economic Inquiry*, 34, 444–463.
- Klein, B., Crawford, R. A., & Alchian, A. (1978). Vertical integration, appropriable rents, and the competitive contracting process. *Journal of Law and Economics*, 21, 297–326.
- Kogut, B., & Zander, U. (1996). What firms do? Coordination, identity, and learning. *Organization Science*, 7, 502–518.
- Kollock, P. (1994). The emergence of exchange structures: an experimental study of uncertainty, commitment, and trust. *American Journal of Sociology*, 2, 313–345.
- Krackhardt, D. (1990). Assessing the political landscape: structure, cognition, and power in organizations. *Administrative Science Quarterly*, 35, 342–369.
- Krackhardt, D. (1992). The strength of strong ties: the importance of philons in organizations. In: N. Nohria & R. G. Eccles (Eds), *Networks and Organizations* (pp. 216–239). Boston: Harvard University Press.
- Larson, A. (1992). Network dyads in entrepreneurial settings: a study of the governance of exchange relationships. *Administrative Science Quarterly*, 37, 76–104.
- Lawrence, P. R., & Lorsch, J. W. (1967). *Organization and environment*. Boston: Harvard University Press.
- Lazzarini, S. G., Miller, G. J., & Zenger, T. R. (2001). *Order with some law: complementarity vs. substitution of formal and informal arrangements* (Working Paper). St. Louis: Olin School of Business, Washington University.
- Leonard-Barton, D. (1995). *Wellsprings of knowledge: building and sustaining the sources of innovation*. Boston: Harvard Business School Press.
- Levinthal, D. A., & March, J. G. (1993). The myopia of learning. *Strategic Management Journal*, 14, 95–112.
- Lewicki, R. J., McAllister, D. K., & Bies, R. J. (1998). Trust and distrust: new relationships and realities. *Academy of Management Review*, 23, 438–458.
- Lincoln, J. R. (1982). Intra- (and inter-) organizational networks. *Research in the Sociology of Organizations*, 1, 1–38.
- Lubell, M., & Scholz, J. T. (2001). Cooperation, reciprocity, and the collective-action heuristic. *American Journal of Political Science*, 45, 160–178.
- Luft, J. (1994). Bonus and penalty incentives: contract choices by employees. *Journal of Accounting and Economics*, 18, 181–206.
- Macaulay, S. (1963). Non-contractual relations in business: a preliminary study. *American Sociological Review*, 28, 55–70.
- Macneil, I. R. (1978). Contracts: adjustment of long-term economic relations under classical, neoclassical, and relational contract law. *Northwestern University Law Review*, 72, 854–905.
- March, J. G., & Simon, H. (1958). *Organizations*. New York: John Wiley & Sons.
- Matthews, R. C. O. (1986). The economics of institutions and the sources of economic growth. *Economic Journal*, 96, 903–918.
- McLean Parks, J., & Coelho-Kamath, C. (1999). *Give & take: framing physician compensation contracts* (Working Paper). St. Louis: Olin School of Business, Washington University.
- Meyer, A. D., Tsui, A. S., & Hinings, C. R. (1993). Configurational approaches to organizational analysis. *Academy of Management Journal*, 36, 1175–1195.
- Milgrom, P., North, D. C., & Weingast, B. W. (1990). The role of institutions in the revival of trade: the Law Merchant, private judges, and the Champagne Fairs. *Economics and Politics*, 2, 1–23.
- Milgrom, P., & Roberts, J. (1990). Bargaining costs, influence costs, and the organization of economic activity. In: J. Alt & K. Shepsle (Eds), *Perspectives on Positive Political Economy* (pp. 57–89). Cambridge: Cambridge University Press.

- Milgrom, P., & Roberts, J. (1991). Complementarities, momentum, and the evolution of modern manufacturing. *American Economic Review*, 81, 84–88.
- Miller, D., & Friesen, P. H. (1980). Momentum and revolution in organizational adaptation. *Academy of Management Journal*, 23, 591–614.
- Mintzberg, H. (1979). *The structuring of organizations*. Englewood Cliffs: Prentice Hall.
- Nelson, R. R., & Winter, S. G. (1982). *An evolutionary theory of economic change*. Cambridge: Harvard University Press.
- Nickerson, J. A., & Zenger, T. R. (2002). Being efficiently fickle: a dynamic theory of organizational choice. *Organization Science*, 13(5).
- North, D. C. (1981). *Structure and change in economic history*. New York: W. W. Norton.
- North, D. C. (1990). *Institutions, institutional change and economic performance*. Cambridge: Cambridge University Press.
- Parkhe, A. (1993). Strategic alliance structuring: a game theoretic and transaction costs examination of interfirm cooperation. *Academy of Management Journal*, 36, 794–829.
- Pfeffer, J. (1978). *Organizational design*. Arlington Heights: AHM Publishing.
- Poppo, L. (1995). Influence activities and strategic coordination: two distinctions of internal and external markets. *Management Science*, 41, 1845–1859.
- Poppo, L., & Zenger, T. R. (1998). Testing alternative theories of the firm: transaction cost, knowledge-based, and measurement explanations for make-or-buy decisions in information services. *Strategic Management Journal*, 19, 853–877.
- Poppo, L., & Zenger, T. R. (forthcoming). Substitutes or complements? Exploring the relationship between formal contracts and relational governance. *Strategic Management Journal*.
- Powell, W. W. (1990). Neither market nor hierarchy: network forms of organization. *Research in Organizational Behavior*, 12, 295–336.
- Rabin, M. (1993). Incorporating fairness into game theory and economics. *The American Economic Review*, 83, 1281–1302.
- Rebitzer, J. M. and M. D. Robinson. (1991). Employer size and dual labor markets. *The Review of Economics and Statistics*, 73, 710–715.
- Ring, P. S., & Van de Ven, A. H. (1994). Developmental processes of cooperative interorganizational relationships. *Academy of Management Review*, 19, 90–118.
- Roethlisberger, F. J., & Dickson, W. J. (1939). *Management and the worker*. Cambridge: Harvard University Press.
- Rousseau, D. M., & McLean Parks, J. (1993). The contracts of individuals and organizations. In: L. L. Cummings & B. M. Staw (Eds), *Research in Organizational Behavior* (pp. 1–43). Greenwich: JAI Press.
- Salancik, G. R. (1977). Commitment and the control of organizational behavior and belief. In: B. M. Staw & G. R. Salancik (Eds), *New Directions in Organizational Behavior* (pp. 1–54). Chicago: St. Clair Press.
- Scott, W. R. (1981). *Organizations: rational, natural, and open systems*. Englewood Cliffs: Prentice Hall.
- Shelanski, H. A., & Klein, P. G. (1995). Empirical research in transaction cost economics: a review and assessment. *Journal of Law, Economics and Organization*, 11, 337–361.
- Shrader, C. B., Lincoln, J. R., & Hoffman, A. N. (1989). The network structures of organizations: effects of task contingencies and distributional form. *Human Relations*, 42, 43–67.
- Sitkin, S. B., & Roth, N. L. (1993). Explaining the limited effectiveness of legalistic “remedies” for trust/distrust. *Organization Science*, 4, 367–394.
- Stevenson, W. (1990). Formal structure and networks of interaction within organizations. *Social Science Research*, 19, 113–131.

- Tenbrunsel, A. E., & Messick, D. M. (1999). Sanctioning systems, decision frames, and cooperation. *Administrative Science Quarterly*, 44, 684–707.
- Tichy, N. M. (1980). Networks in organizations. In: P. C. Nystrom, & W. H. Starbuck (Eds), *Handbook of Organizational Design*. London: Oxford University Press.
- Tirole, J. (1999). Incomplete contracts: where do we stand? *Econometrica*, 67, 741–781.
- Trist, E., & Bamforth, W. (1951). Some social and psychological consequences of the longwall method of coal-getting. *Human Relations*, 4, 3–38.
- Tushman, M. L., & Katz, R. (1980). External communication and project performance: an investigation into the role of gatekeepers. *Management Science*, 26, 1071–1085.
- Tushman, M. L., & Romanelli, E. (1985). Organizational evolution: a metamorphosis model of convergence and reorientation. *Research in Organizational Behavior*, 7, 171–222.
- Uzzi, B. (1996). The sources and consequences of embeddedness for the economic performance of organizations: the network effect. *American Sociological Review*, 61, 674–698.
- Williamson, O. E. (1975). *Markets and hierarchies: analysis and antitrust implications*. New York: The Free Press.
- Williamson, O. E. (1985). *The economic institutions of capitalism*. New York: The Free Press.
- Williamson, O. E. (1991). Comparative economic organization: the analysis of discrete structural alternatives. *Administrative Science Quarterly*, 36, 269–296.
- Williamson, O. E. (1996). *The mechanisms of governance*. New York: Oxford University Press.
- Zenger, T. R. (1992). Why do employers only reward extreme performance? Examining the relationships among performance, pay, and turnover. *Administrative Science Quarterly*, 37, 198–219.
- Zenger, T. R. (1994). Explaining organizational diseconomies of scale in R&D: agency problems and the allocation of engineering talent, ideas, and effort by firm size. *Management Science*, 40, 708–729.
- Zenger, T. R., & Lawrence, B. S. (1989). Organizational demography: the differential effects of age and tenure distributions on technical communication. *Academy of Management Journal*, 32, 353–376.
- Zucker, L. G. (1986). Production of trust: institutional sources of economic structure, 1840–1920. *Research in Organizational Behavior*, 8, 53–111.

This Page Intentionally Left Blank

‘TESTS TELL’: CONSTITUTIVE LEGITIMACY AND CONSUMER ACCEPTANCE OF THE AUTOMOBILE: 1895–1912

Hayagreeva Rao

ABSTRACT

Constitutive legitimacy can be created through evangelical appeals, the efforts of social movement organizations, the enactment of laws that authorize new products, and advertising by firms. This paper investigates these parallel routes to the legitimacy of the car in the early American automobile industry. The results show that evangelical appeals in the form of reliability contests organized in a focal state and social movement organizations in the shape of automobile clubs significantly increased automobile sales in the focal state. The positive effects of reliability contests on automobile sales increased with the number of automobile clubs. However, the effects of auto clubs and reliability contests declined as advertising by firms grew, and fell with the passage of time since legislation authorizing the car in the focal state. Taken together, these results suggest complex interdependencies among the parallel routes to constitutive legitimacy in the case of new industries.

INTRODUCTION

A common proposition in organizational theory is that new industries based on embody unfamiliar and strange products, and are sources of uncertainty for consumers, potential employees, financiers and governmental authorities (Stinchcombe, 1965; Meyer & Rowan, 1977; Aldrich, 1999). New industries flourish when they possess constitutive legitimacy, that is, they become understood and taken-for-granted by consumers, investors, and potential employees (Carroll & Hannan, 2000; Baum & Powell, 1995; Zucker, 1983). A key issue in organizational theory centers around the sources of constitutive legitimacy.

Organizational ecologists have proposed that new industries become taken-for-granted as the number of organizations in the industry (density) increases until a point, after which, growing density unleashes competition. An impressive number of studies have shown an inverted U-shaped relationship between population density and foundings, and a U-shaped relationship between density and mortality (see: Hannan & Freeman, 1989; Carroll & Hannan, 2000 for a review). Although the density-dependence formulation is exceptionally useful for its simplicity and generality, it looks at the intermediate outcomes such as foundings and failures rather than direct outcomes such as consumer acceptance and/or acceptance by financiers. Moreover, it implies that constitutive legitimacy is an automatic positive spill-over that arises as managers of firms go about the business of building and prolonging their enterprises.

However, there are many other parallel routes to constitutive legitimacy. Thus, new industries can secure constitutive legitimacy through an institutional project in which evangelists justify the product, and social structures exist to mobilize participation of actors (DiMaggio, 1988; Suchman, 1995; Carroll & Hannan, 2000; Rao, Morrill & Zald, 2000; Swaminthan & Wade, 2000). For example, consumer concerns about the morality of pricing lives impeded the early growth of the life insurance industry until industry activists undertook a publicity campaign that framed the purchase of life insurance as a morally responsible and prudent act (Zelizer, 1979). For instance, the patronage of hobbyists who assembled kit computers was crucial to the acceptance of the personal computers, and clubs where hobbyists could exchange information spread awareness of personal computers. In a related vein, other writers suggest that laws authorizing the use of a product and stipulating the conditions of its use not only have a facilitative effect but reflect understandings of the product (Edelman & Suchman, 1997). For instance, the sperm banking industry became unquestioned as family laws legally separating the donor and the mother became enacted and regulations authorizing tissue banking became codified.

Furthermore, new industries can be legitimated by the mass media through the medium of advertising. In this case, firm-level activity produces a positive externality in the shape of a legitimacy for the industry. Finally, new industries can also be legitimated through trade associations and professional societies. For example, the Harry Benjamin International Gender Dysphoria Association, a society of gender-reassignment specialists, played a significant role in establishing worldwide standards and creating the gender reassignment industry.

No one study investigates the roles of evangelism, mobilizing structures, law, advertising, and density dependence as sources of constitutive legitimacy. Ingram and Clay (2000) provide a helpful point of departure when they argue that institutional forms can be categorized into a two-by-two matrix with the public-private dimension on one side, and the centralized-decentralized dimension on the other side.

The resultant four quadrants accommodate the parallel sources of constitutive legitimacy. Thus, laws are in the public-centralized quadrant, evangelism and mobilizing structures are in the public-decentralized quadrant, advertising and density-dependence are in the private decentralized quadrant, and trade association activity could be situated in the private-centralized quadrant. There may be complex interdependencies among these parallel routes; thus, the effects of advocacy by enthusiasts in the public-decentralized quadrant may be modified by advertising and law.

This study seeks to uncover the role of evangelism, mobilizing structures, laws, and advertising on the constitutive legitimacy of the car as measured by car sales in the early American automobile industry. At its birth in 1895, with the advent of the Duryea Car Company, the automobile industry lacked constitutive legitimacy. It was unclear if the car was a reliable means of transportation. Observers indicate that the automobile was the target of jokes and derision. The automobile gained acceptance through mobilizing structures in the shape of automobile clubs peopled by motor car enthusiasts. These clubs lobbied for laws authorizing the use of the car, and also evangelized on behalf of the car. Evangelism took the shape of auto clubs organizing reliability contests to demonstrate the dependability of the car. Cars likely to be bought by consumers were pitted against each other and evaluated in endurance, hill climbing, and fuel economy runs. Reliability contests were 'demonstration events' that blended the practice of racing with the logic of product testing. Each contest was widely viewed as a 'test' that 'told' audiences that the automobile was reliable. Additionally, auto clubs also played an important role in lobbying state governments for laws licensing motor car usage, and these laws also codified understanding of the car. Auto producers also advertised

their cars, and grew in numbers. However, there was no consequential trade association activity of note in the early history of the auto industry.

I study the industry from 1895 until 1912 because the first reliability contest started in 1895, and by 1912, reliability contests were discontinued because by then the automobile was no longer an artifact but a social fact. Soon, thereafter, Henry Ford initiated the mass production of cars in 1912.

NEW INDUSTRIES AND CONSTITUTIVE LEGITIMACY: PARALLEL ROUTES

Several writers have suggested that the formation of new industries resembles a social movement because activists seek to mobilize support for their cause and evangelize (Fligstein, 1996; Rao, Morrill & Zald, 2000; Carroll & Hannan, 2000). Social movements underlie the creation of new industries because constitutive legitimacy is a public good, and the emergence of collective action hinges on political opportunity, mobilizing structures, and framing processes.

New industries may be legitimated by conflict-oriented social movements when political threats are salient, and by consensual social movements when political opportunity is high (Rao, Morrill & Zald, 2000). Conflict-oriented social movements are likely when a new industry faces opposition from vested interests or is created in explicit opposition to vested interests. Organizers of the new industry who have little power and are excluded from existing social structures (McCarthy & Zald, 1977) challenge vested interests by framing the antagonist in negative terms, and their own industry in positive terms. For example, the craft-brewing industry arose in opposition to industrial brewers, and activists organized fairs to mobilize interest in craft-brewing, framed industrial brewers as poor-quality producers, and depicted craft brewing as authentic, traditional, and artisanal (Carroll & Swaminthan, 2000). After the craft-brewing industry was promoted by third parties, producers exploited their rhetoric to launch advertising campaigns.

Social movements may also be relevant when there are minimal political threats facing the new industry. Consensus-oriented social movements arise when there is a failure of market processes and normal market incentives are inadequate. For example, although technical standards benefit both producers and consumers, neither a coalition of producers nor a coalition of consumers took the lead in creating standard-setting organizations in the United States. Activists such as James Chase and Frederick Schlink (who were employees of the National Bureau of Standards) railed against the evils of wasteful variety, preached the virtues of standards, recruited converts from the ranks of private corporations, and set up standard-setting bodies.

Mobilizing Structures

Institutional activists play a key role in mobilizing legitimacy and other resources in social movements (McCarthy & Zald, 1977). Some writers suggest that institutional activists can exclude others from the *psychological benefit* of contributing to a large and important cause and therefore, overcome the free-riding problem (Moe, 1980). Other writers propose that activism is not the product of rational calculation and is instead, driven by personal pride, community affiliations, and inter-group rivalry (Knight & Ensminger, 1998; Ingram & Inman, 1996).

In new industries, activists may be executives in producing firms, experts in trade associations or professional societies, or clubs of enthusiastic consumers. Executives of producing firms can play a key role in building the constitutive legitimacy of new industries: for example, Steve Jobs played a crucial role in the development of the personal computer industry. Similarly, trade association personnel or members of professional societies can play an important role in establishing the new industry (Aldrich, 1999). For example, the Harry Benjamin International Gender Dysphoria Association, a society of gender-reassignment specialists, played a significant role in establishing worldwide standards and creating the gender reassignment industry. Consumer activists may also organize campaigns to improve the cognitive status of new industries. For instance, bicycle clubs peopled by cycling enthusiasts played a more important role in establishing the bicycle as source of health than individual producers who bankrolled campaigns to advertise the bicycle as a valued product (Smith, 1972).

Activists gain access to the political system and mobilize resources through formal and informal mobilizing structures. Such structures include formal social movement organizations or SMOs (McCarthy & Zald, 1977), extra-movement structures such as work and neighborhood organizations, and informal friendship networks (Tilly, 1978). In general, committed activists adhering to a cause are effective when they are formally organized in SMOs. SMOs mobilize their constituency with a view to obtaining a collective good or preventing a collective ill, and may take the form of local volunteer groups as in grassroots movements at one extreme (Lofland, 1985), or may be manifested as local professional organizations such as public interest groups (McCarthy, 1996). These local organizations may be connected to each other through informal linkages, formal committees, or more enduring federal structures that mobilize diffuse networks of supporters (Staggenborg, 1991).

The greater the number of SMOs, the easier it is for activists to disseminate information, gain attention, and to recruit other activists (Kriesi, 1996). The sheer density of SMOs enhances constitutive legitimacy for new industries on three counts. First, it promotes face-to-face communication and enables activists

to develop a sense of collective identity (Staggenborg, 1991). Second, close contact facilitates the exchange of specialized political knowledge necessary to mobilize a group and enables network formation (Klandermans, 1992). Third, an increase in the density of SMOs enables activists to thwart organized opposition in conflict-oriented movements, and to build social consensus when there is a lack of explicit opposition (McAdam, McCarthy & Zald, 1996). The larger the number of SMOs, the higher is their ability to enhance consumer awareness and acceptance of the new product. As a result, sales of the new product are likely to increase. Therefore:

H1: SMOs have a positive effect on sales of the new product.

Evangelism

Resources can only be mobilized when activists engage in evangelical “meaning-work” wherein they frame an issue, and array events and experiences in cognitive packages (Snow, Rochford, Worden & Benford, 1986). Just as a picture frames our understanding, evangelical claims made by activists also frame the new product as valid, useful, and appropriate. Evangelical claims become potent when they are premised on events that can be construed as evidence by audiences, and which reflect prevalent social beliefs (Snow & Benford, 1992, p. 150).

Evangelism can subsume a variety of strategies (Suchman, 1995). At one extreme, activists can rely on ‘demonstration events’ to communicate the viability and dependability of the radical product underlying the new industry. For example, bicycle clubs composed of enthusiasts organized races to establish the bicycle as a font of health and vitality (Smith, 1972). In some cases, activists may launch publicity campaigns: in the life insurance industry, activists portrayed the purchase of a life insurance policy as a responsible act to defuse opposition to the idea of pricing lives (Zelizer, 1979). Activists may also formulate performance measures to legitimate new industries (Meyer, 1994). For instance, in 1991 the National Committee on Quality Assurance, a group consisting of producer and consumer representatives, defined sixty performance measures and integrated them into a ‘report card’ for managed-care organizations (Cerna, 1993). These report cards were plausible claims of quality that allowed individual HMO’s to stand out, thereby allowing the industry to legitimate itself as a customer-centered and cost-focused innovation. Activists may also rely on scientific research to establish the viability of new industries. Van de Ven and Garud (1989) document how rival bands of researchers in the cochlear ear implant industry favoring single and multiple channels relied on peer-reviewed articles to show that their technology was better.

Such strategies of evangelism reduce doubts in the eyes of consumers. The greater the evangelism, the more consequential are its priming and framing effects. Evangelical claims such as demonstration events and report cards make the new product 'available' to potential consumers (Schudson, 1989), and place a narrative framework that may resonate with the life of the audience. The frequency of evangelical claims is important because audiences make sense of the unfamiliar through cognitive shortcuts such as the availability heuristic in which judgement depends on what comes to mind (Kinder, 1998). Audiences arrive at opinions by averaging across occurrences that are accessible, and in turn, accessibility hinges on frequency (Iyengar & Kinder, 1987). The greater the frequency of evangelical claims such as demonstration events that provide social proof of a new product's viability, the lower is the uncertainty, and the more favorable are consumer evaluations of the new product, and the higher is consumer acceptance, and sales of the new product. Therefore:

H2: Evangelism has a positive effect on sales of a new product.

The effect of evangelism on sales of a new product is likely to be augmented by the social organization of institutional activists. The greater the number of SMOs, the easier it is for them to mobilize resources to build constitutive legitimacy. The more frequent evangelical appeals are in a focal state, the more primed are audiences of the benefits of the new product, and the lower is their uncertainty. Thus, the strength of activism is likely to magnify the effect of evangelical appeals on consumer acceptance. Therefore:

H3: The SMOs × evangelism interaction has a positive effect on sales of a new product.

Advertising as a Source of Constitutive Legitimacy

A ubiquitous alternative to activism and claim-making by institutional activists is advertising by producers. If evangelism and SMO's are in the public-decentralized sphere of institutional work, advertising is in the private-decentralized realm. Berger and Luckmann (1966) suggest that new objects can be integrated into the prevalent cultural order through rudimentary ideas purveyed via mass media channels. Sociologists of consumption suggest that mass media advertising exerts powerful effects on the popular understandings of new products and commodities, and connects them to the prevalent cultural order (Schudson, 1989). Advertising primes associations between an object and other products, and thereby, heightens awareness and integrates a new product with the prevalent cultural order (Schudson, 1989; Iyengar & Kinder,

1987). By situating the new product in established genres, advertising makes new products comprehensible and understood by consumers and investors.

Although firms undertake advertising primarily to build their reputations and increase demand for their products, there may still be spill-over benefits such as enhanced consumer awareness and knowledge about the product. The literature on pioneer burnout in new industries suggests that pioneers bear the costs of advertising, but the benefits of advertising can be appropriated by other potential founders entering the market later (Mowery & Rosenberg, 1998). The greater the extent of advertising, the more exposed and informed are audiences to the new product. Since advertising increases consumer familiarity and may also reduce questions about the product, the social organization of activists and their evangelism should have less consequential effects on sales. Therefore:

H4: Advertising × SMOs interaction has a negative effect on sales of a new product.

H5: Advertising × evangelism interaction has a negative effect on sales of a new product.

Law as a Source of Legitimacy

Neo-institutional researchers have shown that institutionalization is a temporal process by which understandings of structures, policies, and programs become codified (Meyer & Rowan, 1977, p. 341; Zucker, 1983; Tolbert & Zucker, 1988). An important marker of institutionalization is legal recognition of the new industry (Stinchcombe, 1965). Although legal recognition is commonly viewed as a source of socio-political legitimacy, laws also embody constitutive understandings of a radical new product that become even more codified as time since the enactment of legislation increases (Edelman & Suchman, 1997). Laws authorizing the use of the product and stipulating the conditions of its use have a facilitative effect and reflect understandings of the product. The greater the elapsed time since the enactment of a law authorizing the use of a product, the more taken-for-granted are understandings of the product, and the more likely are they to become the default rules by which consumers and other audiences make sense of the world (Edelman, 2000). If so, the enactment of laws should lead to the substitution of institutional forms of legitimacy such that the public-centralized realm weakens the effects of public-decentralized sources of constitutive legitimacy. The greater the elapsed time since the onset of legal recognition, and the more codified the understandings of the product, the less consequential is the role of activists and evangelism in the accretion of constitutive legitimacy. Therefore:

- H6:* The time since legal recognition $\times$ SMOs interaction has a negative effect on sales of a new product.
- H7:* The time since legal recognition $\times$ evangelism interaction has a negative effect on sales of a new product.

AUTO CLUBS, RELIABILITY CONTESTS AND THE LEGITIMACY OF THE AUTO: 1895–1912

These hypotheses were investigated in the early American automobile industry for four reasons. First, the early American automobile industry constituted a radical departure from its precursor, the horse carriage industry. Unlike the horse carriage industry that relied on animal power, automobile firms used steam, gasoline, and electric power to provide customers with horseless carriages. Horse carriage firms were one-man operations, but early automobile firms were assemblers who put together components (Flink, 1970). Above all, the automobile was a radically new artifact that promised to transform the experience of transportation. Second, activists who banded together to establish auto clubs played an important role in promoting a favorable image of the automobile. Third, activists faced a favorable political opportunity structure to organize and make evangelical claims. Finally, activists evangelized by organizing demonstration events in the form of reliability contests (Epstein, 1928; Flink, 1970). Demonstration events were a potent claim-making strategy because they were credible, experiential, and reflected master logics of society. Reliability contests were credible because they were interpreted as evidence by the targets of claim-making. Since they intruded into the daily lives of targets, they derived experiential validation. Reliability contests possessed narrative fidelity to the extent to which they were connected with extant master logics of testing in society.

Some writers trace the origins of the American automobile industry to the Selden two-stroke engine design developed in 1879, to William Morrisons's electric car of 1892, or to Ransom Olds's steam vehicle, purportedly sold to an Indian firm in Bombay. However, the first firm to make automobiles was set up by the Duryea brothers in 1895 (Flink, 1970). Some accounts of the automobile industry reduce the birth of the early automobile industry to the question of how gasoline-powered cars began to dominate the industry. Although the technological development of the automobile industry is interesting in its own right, it can deflect attention from the larger question of how the automobile came to be legitimated.

Even though the automobile was a substitute for the horse-drawn carriage, it did not arouse opposition from manufacturers of horse-drawn carriages, livery

stable owners, and horse-drawn vehicle driver associations (Flink, 1970, p. 64). The reason was that prominent makers of horse-drawn carriages (e.g. Studebaker & Flint Wagon) began to produce cars, and owners of livery stables opened garages. Some opposition to the automobile was organized by a handful of vigilante anti-speed organizations in New York and rural states such as Minnesota and Indiana, but it was not persistent, and petered out. Thus, there was substantial political opportunity for activists to evangelize on behalf of the automobile, and such evangelism was necessary because of doubts about the automobile.

Consumer Uncertainty About the Automobile

A problem facing the new industry was consumer uncertainty. At the origins of the automobile industry, the automobile was an unfamiliar product to consumers and its reliability was suspect. There was no standardized vocabulary to refer to the car and it was variously called a gasocar, a motocle, a motorcar, a motor-car, or a horseless carriage.

Many of the early manufacturers were remotely connected to making cars. For example, the Smith Automobile Company of Topeka, Kansas, was in the business of making hernia trusses before taking up automobile manufacturing. Almost equally offbeat was the origin of the renowned Pierce-Arrow. George N. Pierce began as a manufacturer of bird cages, which led him to the manufacture of spokes for bicycle wheels. This was followed by the production of complete bicycles and finally, automobiles (Volti, 1996).

Given the unorthodox origins of entrepreneurs, it is not surprising that many cars were unable to complete a drive successfully and had to be hauled back by a team of horses. Quite a few vehicles were designed with whip sockets and harness hitches (Epstein, 1928). Consumers were hesitant to purchase a car given uncertainty about whether it was a superior alternative to the horse, and a dependable means of transportation (Epstein, 1928, pp. 89–92).

Failure of Collective Action by Producers

Firms within the new industry could not successfully band together to legitimate the automobile. In the first four years, there were no trade associations to advance the cause of the automobile. The National Association of Automobile Manufacturers was established in 1900 in a bid to assure product quality, but it was superseded by the Association of Licensed Automobile Manufacturers (ALAM), which was formed in 1903. ALAM was a trade association formed to license the Selden patent and was set up ostensibly to prevent the incursion of imitators. But the Selden patent was widely disregarded, and due to internal divisions, ALAM was unable to secure quality by enforcing its threat of

litigation. A rival association called the American Motor Car Manufacturer's Association (AMCMA) was established in 1905, and also proved to be an ineffective mechanism of collective action. Both trade associations disintegrated during the period 1909–1911, as a result of legal battles. Since there were no effective trade associations and the Federal government was inactive, it was not possible to formulate a national license and registration policy to defuse opposition to the car from anti-speeding interests.

Lack of Professional Infrastructures

Professional societies did little to legitimate the new industry. The Society of Automobile Engineers (SAE) began in 1905 with a small group of journalists and automobile engineers, and established a standards committee by 1910. In 1910, the SAE diagnosed the lack of inter-company standardization of components as a major cause of production problems and expenses, and initiated a program that eventually resulted (in 1921) in the formation of 224 standards.

Professionals also did not play an important role in training personnel. In the early years of the industry, some makers of high-priced cars trained drivers at special schools (e.g. Packard and Locomobile). Nearly a decade after the appearance of the first automobile firm between 1903 and 1904, the YMCA instituted courses for drivers, and advanced training for engineers and draftsmen. In 1905, the New York School of Automobile Engineers was established by a professor of engineering from Columbia, and other affiliates arose in a few large cities to disseminate mechanical expertise relevant to the automobile. However, the professionalization of automobile engineering really began when the SAE began a large training effort after 1912.

Automobile Clubs as SMOs

It was in this context that automobile fans banded together into automobile clubs, and served as the automobiling movement which sought to promote the car, and to lobby for better roads. Flink (1970, p. 144) notes that the “automobile club became the most important institution championing the diffusion of the automobile in the United States. Voluntary associations of motorists propagandized to encourage a favorable image of the automobile and automobilists.”

The American Motor League, set up in 1895, was the first attempt to organize a national-level organization, but it foundered for lack of support. By contrast, local clubs of motor car enthusiasts organized in cities and towns flourished. By 1901, 22 local clubs had mushroomed in different cities such as Boston, Newark and Chicago. By 1904, it was “as difficult to find a number of motorists who have not formed a club, as it is to find an individual motorist who is not a member of some such body” (*Horseless Age*, February 17, 1904, pp. 196–197).

This surge of automobile clubs also induced attempts to form national associations. As early as 1902, there were two rival attempts to form national organizations. The American Motor League (revived by its 1895 organizers), led by Charles Duryea, sought to enlist individual motorists as members and aimed to establish branches. By contrast, the American Automobile Association (AAA) began as an association of local clubs, and after some early disagreements with the New York clubs, emerged as the primary representative of car owners in America.

By 1910, there were 225 local clubs affiliated with the AAA. The local clubs promoted the image of the automobile in numerous ways. Some clubs organized tours for underprivileged children to dissuade them from throwing stones on passing cars in places such as Chicago and New York. Automobile clubs supported state ordinances requiring tags and mandating speed limits to prevent a maze of city-specific regulations and to defuse opposition to the car. More importantly, auto clubs codified rules for reliability races and provided the personnel for scheduling and supervising these contests.

Attempts to demonstrate the capabilities of the automobile began in 1895 when the first firm to produce cars was established by the Duryeas. The first contest was the *Times-Herald* race held on Thanksgiving Day in 1895. The organizers stated they were influenced by the “desire to promote, encourage and stimulate the invention, development and perfection and general adoption of motor vehicles” (quoted in Thomas, 1977, p. 21).¹ Five of the eleven entrants participated, and only two vehicles were able to complete the race. The first prize of \$10,000 was won by a Duryea car powered by gasoline, which had a winning speed of eight miles an hour. The *Times-Herald* report the next day stated that the race had been run in 30-degree temperatures “through deep snow and along ruts that would have tried horses to the utmost,” and implied that automobiles were practical.

Shortly thereafter, *Cosmopolitan* magazine offered a prize of \$3,000 and held a contest on May 30, 1896 that was won by a Duryea. Subsequently, the Rhode Island State Fair Association organized a competition and offered \$5,000 in prize money; this was won by an electric car. The Riker Electric car won the race, but spectators found the contest to be so dull that they originated the cry “get a horse” (Flink, 1970, p. 42). Local automobile clubs quickly jumped into the fray to sponsor reliability contests, and soon newspapers were reduced to the role of covering races rather than organizing them.

Reliability runs consisted of hill-climbing, endurance, and on occasion, fuel economy runs. These runs featured cars that were likely to be used by ordinary consumers. A writer commented about the Glidden Tour, a reliability contest, saying that it had “proved the automobile is now almost foolproof. It has proved

that American cars are durable and efficient . . . it has strengthened our belief in the *permanence* of the motor car” (*Horseless Age*, July 26, 1905, p. 153, italics mine). If reliability contests established the ordinary motor car likely to be used by consumers as a viable mechanism, speed contests included featured specialized monstrosities unlikely ever to be bought by consumers. Beach, track and road races that “placed primary emphasis on speed were more important for their contribution to automotive technology as tests for weaknesses in design than as publicity for the motor vehicle” (Flink, 1970, p. 42).

Firms had incentives to enter reliability contests because winning firms reaped substantial publicity due to press coverage, and proclaimed these victories in their advertising campaigns. Losers were not penalized by negative publicity, and could always enter another competition. Manufacturers sponsored cars directly in these contests. A few contests (especially the Glidden tour) stipulated that cars were to be driven by their owners; however, auto firms circumvented this because any executive could drive a car himself (Flink, 1970, p. 41).

Automobile producers could not organize reliability contests and simultaneously enter their own cars because of conflicts of interest. If producers had sponsored a contest and entered their own cars, other producers would have been less willing to participate in the contest. A focal producer had little incentive to publicize winning a race organized by itself. Rather, automobile producers had incentives to enter cars in reliability contests organized by third parties, and to win them.

Automobile clubs played an important role in organizing reliability contests. As early as 1901, the Automobile Association of America in New York City formulated a set of racing rules and assisted local auto clubs in scheduling contests. Local clubs were minimalist organizations composed of motor car enthusiasts and car owners drawn from the community. Since club members were typically car-owning enthusiasts and lead users, club involvement enabled members to construct an identity built around a new consumer role (Charles, 1993). As community organizations, auto clubs were forums for the exchange of information about cars, lobbying devices to get good roads and speed laws, and sought to promote the acceptance of the car within the community by organizing reliability contests. Moreover, since reliability contests were spectacles that aroused public interest, it was also a useful mechanism for clubs to boost membership. Activities such as lobbying for roads and organizing reliability contests enabled club members to feel that they were contributing to the community and to derive a sense of pride and collective identity. Thus, automobile clubs were SMOs who overcame collective action problems to organize reliability contests.

Reliability Contests as Evangelism

Each reliability race was a claim that the car was reliable and safe. Reliability contests were potent claims because they possessed empirical credibility, experiential commensurability, and narrative fidelity. Reliability contests were credible because each race was an event that could be interpreted as evidence of the dependability of cars by the public. Since reliability contests were events that could easily be watched by the public, they possessed experiential commensurability. Finally, reliability contests had narrative fidelity because they combined the logic of testing with the practice of racing. Swidler (1986) suggests that entrepreneurs treat pre-existing cultural resources as items in a menu and blend them creatively to fashion new repertoires. Organizers of automobile clubs borrowed the practice of racing from the bicycle industry, and blended it with the nascent logic of testing to fashion a potent strategy of claim-making. Individual contests generated winners who quickly acquired a reputation for reliability, but also strengthened the claim of the automobile as a safe and reliable device.

Bicycle racing appeared in 1878, soon after the introduction of the bicycle in America. Early bicycle races were road races that sprang out of cycling tours conducted by bicycle clubs, and almost every city with pretensions of being important had a road race (Smith, 1972, p. 144). Participants were weekend cyclists of varying ability, and were ‘handicapped’ by clubs such as the League of American Wheelmen or the Associated Cycling Clubs. The prize went to the man who finished first, and newspapers featured extensive coverage of many races, especially prominent ones such as the Pullman Road Race held in Chicago, which originated in 1883. Since it was difficult to tell who was winning a road race, track races first appeared in 1883. As cycle manufacturers realized the publicity that accrued to them from winning, they began to attract the best racers, and such “maker’s amateurs” made track racing a contest among manufacturers (Smith, 1972, pp. 149–150). Firms that won bicycle races advertised victories as signals of quality.² Thus, the prevalent practice of bicycle races being organized by clubs served as a building block for car enthusiasts and automobile clubs as they grappled with the problem of product quality. Additionally, numerous car races organized in France also increased the attractiveness of races as assessment mechanisms (Epstein, 1928).

Concurrently, a nascent logic of testing was also becoming established with the rise of standards and testing bodies that were extensions of trade associations and professional societies seeking to promulgate common metrics to assist business organizations.³ In 1894, an association of insurance underwriters (*Underwriters Laboratory*) received a charter to certify wires and light fixtures as fire-resistant in order to build insurable real estate. Other trade associations

established standard nomenclatures and performance specifications in the wool blanket and laundry industries. Materials'-testing experts promulgated standards for paint, and electrical engineers developed standards for electrical components for large business enterprises (Chase & Schlink, 1928). The testing and standards ideas received encouragement from the state when the National Bureau of Standards was founded in 1901, and soon instituted annual national conferences on weights and measures. Later, a '*Journal of Weights and Measures*' was established in 1908 for the 'benefit of Dealers, Sealers and the Purchasing Public'.

Car enthusiasts in automobile clubs found it easy to blend the material practice of racing with the logic of testing. Reliability contests became 'tests' that 'told' consumers of the quality of cars; wins in these contests were small cues, which were embellished by the media and through advertising by firms, into a 'story' of quality. Buick, after winning several contests, proclaimed "Tests tell – Could you ask for more convincing evidence?" Thus, advertising campaigns planned by automobile producers were mechanisms to inform the public of their winning record in tests (Epstein, 1928).

As the public watched reliability contests and learned about them through the media, knowledge of the automobile per se diffused across different sections of American society. Specialized trade journals arose to disseminate information about the automobile, and many of them dedicated resources to the coverage of contests. By 1909, the novelty of the automobile had largely worn off. Even in the prestigious Glidden tour, the number of entrants had dropped to 13 in 1909, and after 1912, the Glidden Tour was discontinued. Soon, the Ford Motor Company stopped participating in reliability contests because they were deemed to be unnecessary. After 1912, reliability contests ended, but speed contests continued to flourish because they had become a sport (Epstein, 1928).

Advertising

If reliability contests organized mainly by auto clubs played an important role in shaping the understanding of the automobile, producers also sought to purvey an image of cars through advertising. The popular press and trade journals accepted advertisements by producers who touted their own products. For example, the Peerless Company advertised its cars as "rapid and powerful hill climbers." Similarly, St. Louis Motor Carriage Company touted its cars as "rigs that run." Oldsmobile grandly proclaimed that the Olds "runs everywhere" and Cadillac coined the slogan that "when you buy a Cadillac you buy a round trip" (Thomas, 1977, p. 47). Such advertisements may have reinforced positive evaluations of the car.

Legal Authorization by State Governments

Some opposition to the automobile was organized by a few vigilante anti-speed organizations in cities such as New York and in rural states such as Minnesota. Municipal governments wanted to arrest speeders and insisted that each car should have a numbered tag, and also contemplated the passage of city ordinances specifying speed limits. Some local automobile clubs initially challenged these city ordinances, but quickly realized that a maze of municipal regulations could only be checked if there were state-wide rules for registering and licensing automobiles. Although the National Association of Automobile Manufacturers (an abortive trade association), sought to have Federal legislation providing a national license, it made little headway because Congress and the Federal government were apathetic to the automobile until 1909. Indeed, the Federal government was not even a customer of the automobile until 1909 when the War Department and the Post Office began to acquire specialized automobiles (Flink, 1988).

In these circumstances, individual state-level governments were the key actors who conferred recognition on the automobile. New York took the lead in 1901, and required that all cars have a numbered tag and mandated a 20-mph speed limit. By 1903, eight other states had followed, and by 1906, fifteen states required car tags and speed limits of 20 mph.

Thus, the average speed limit in a state was substantially higher than the average speed limit of 5 mph for horseless carriages. By 1910, thirty-six states required motor vehicle registration, and by 1912, forty-three states had motor vehicle registration, where automobiles had to possess a numbered tag. These laws reflected an understanding of the automobile as a convenient and fast method of transportation.

DATA AND METHODS

The window of observation began in 1895, when the industry was born with the Duryea car and the first reliability competition was organized. The window of observation ended in 1912, because after that reliability contests disappeared, even as speed races flourished, and the car was widely assumed to be legitimate (Flink, 1970). It is noteworthy that Henry Ford also commenced mass production of cars in 1912, and 43 states had laws with liberal speed limits.

Dependent Variable

States were defined as the units of analysis because data on the dependent variable, sales of cars, were available at the state level and unavailable at

the municipality level. Additionally, consumers in a focal state faced a homogeneous institutional environment because laws which stipulated licensing of drivers and speed limits were enacted by state authorities rather than by municipalities. Moreover, clubs in a focal state also tended to work together to lobby for roads and laws, and reliability contests organized by a club were also publicized by newspapers. Hence, I constructed a dataset where each observation was a state-year. The dependent variable, consumer acceptance, was operationalized as the sales of cars in a focal state. Data on car sales were obtained from the censuses of various states and from historical statistics of the United States.

Independent Variables

The number of social movement organizations in a focal state was defined as the number of automobile clubs in the focal state. Data on the number of automobile clubs was obtained from the American Automobile Association historical archives, state censuses, *Horseless Age*, and histories of the automobile industry. Where data were missing for a year, linear interpolation techniques were used.

The extent of evangelism was measured through the relative frequency of reliability contests. Hill-climbing competitions, fuel economy contests, and endurance runs were treated as reliability competitions because they emphasized the dependability and durability of cars likely to be bought by ordinary consumers. Speed contests comprised of road, beach, and speedway races were excluded because they featured specially-designed cars to test technical ideas and to shatter speed records. In some cases, they also evoked opposition (Flink, 1970; Thomas, 1977). I focused on in-state reliability contests because they were proximal events likely to affect consumer uncertainty rather than out-of-state contests. Evangelism was defined as the proportion of in-state reliability contests/all in-state contests because it was necessary to capture the relative frequency of reliability contests vis-à-vis speed contests in a focal state. Thus, the frequency of reliability contests was divided by the frequency of all contests (reliability and speed) in a focal state. Data on the exact dates and location of these contests were gathered from *Horseless Age*.

Data on the amount of money spent by producers on advertising during the period 1895–1912 were not available for all producers. Instead, I devised a proxy that relied on the number of pages of advertising in the premier national automobile publication, *Horseless Age*. I restricted myself to *Horseless Age* because it was prohibitively costly to gather data from multiple newspapers in individual states. I inspected each issue for every year from 1895–1912, computed the total number of pages devoted to advertising, and standardized it by the number of pages in the issue of the magazine. The

total number of advertising pages/total number of pages was computed for each year.

Elapsed time since legal recognition was defined as the number of days since a law was enacted in the focal state that required that cars be registered and conform to a speed limit. Data on the timing of legislation enacted by various states were obtained from the U.S. Department of Commerce, *Highway Statistics, Summary to 1955*.

Additionally, I also included some controls to account for other causal influences on car sales in a focal state. Since affluent and more urbanized states more likely to have a greater number of car buyers, I controlled for state domestic product and urbanization. State gross product was defined as the total value of farm and industrial production, and urbanization was defined as the percentage of urban population. Data were obtained from the ICPSR databases on historical statistics of the United States. Since the data are decennial in nature, I used linear interpolation to get annual figures. Instead of using 1890, 1900 and 1910 data and then obtaining the interpolated annual values, I used the data from 1790–1930, and then arrived at the interpolated yearly values.

I also used the number of car firms within a focal state as a control variable because the number of producers is likely to impact sales. Density-dependence theorists propose that new industries become taken-for-granted as the number of organizations (density) increases until a point, after which growing density unleashes competition. This formulation suggests that there is an inverted U-shaped relationship between density and growth rates of individual firms in an industry (see Carroll & Hannan, 2000, for a review). However, density dependence has been shown to affect the foundings and growth rates of individual firms, and not industry-level outcomes such as sales. Moreover, the inverted U-shaped density-dependence relationship holds when the complete history of a new industry is observed, and researchers contend that legitimation effects predominate in the early history of any industry (Carroll & Hannan, 2000). Since I was studying an industry-level outcome, especially in an industry's early history, I included only the first order effect of density as a control. Data on the number of car companies were obtained from Kimes and Clark (1989). All independent variables were lagged by a year. Table 1 shows the descriptive statistics.

Since some of the correlations are between 0.50 and 0.60, I conducted tests to assess the issue of multi-collinearity in exploratory analyses. Belsley, Kuh and Welsch (1980) suggest that if the conditioning number is greater than 100, then estimates are likely to be erroneous. My analysis of collinearity diagnostics indicated that the conditioning number was 67.91. Additionally, Belsley, Kuh and Welsch (1980) propose that the variance inflation factor (VIF)

Table 1. Correlation among independent variables.

	1	2	3	4	5	6	7
1. State Gross Product		0.58	0.40	0.10	0.17	0.32	0.24
2. Urbanization			0.60	0.11	0.30	0.46	0.44
3. No. of in-state firms				0.12	0.31	0.60	0.50
4. Advertising					0.52	0.19	0.26
5. Time since legal recognition						0.40	0.47
6. No. of SMOs							0.41
7. Evangelism							

should not exceed 10, and my analyses revealed a mean VIF of 2.02. Thus, both tests mitigated any concerns about collinearity.

Modeling Strategy

Studies of sales as a dependent variable (conducted principally by marketing researchers) presume that there are diminishing returns to promotional efforts on sales, and a common strategy is to implement a double-log specification, where the dependent and independent variables are log-transformed before estimating regression models (Leeflang, Wittink, Weder & Naert, 2000). Since the effects of SMOs, evangelism, state gross product, urbanization, number of firms and time since legislation are likely to encounter diminishing returns, I used the double-log specification. In order to ensure that state-years where sales or an independent variable with a value of zero were treated as valid observations, I added the value of 1 before log-transforming the data.

Since the data set consisted of state-years, the clustering of observations by states violates the assumption of the independence of observations and also introduces problems of serial correlation. In such cases, ordinary least-squares estimates are biased; I therefore estimated a random effects regression which had the following specification:

$$Y[i, t] = \alpha + \beta * X[i, t] + v[i] + \varepsilon[i, t]$$

where $v[i] + \varepsilon[i, t]$ is the residual, and $v[i]$ is the cluster-specific residual which differs between clusters, but its value for any cluster is constant. I used Maximum-Likelihood Estimation (MLE) techniques to build random-effects models because Generalized Methods of Moments techniques do not consider the predicted group means as one of the moments and therefore do not match observed values.

RESULTS

Table 2 displays the results obtained from the random-effects MLE regression models of car sales. Model 1 is a baseline model. State gross product, time since legislation, and the density of car firms significantly increase car sales. The effects of advertising are not significant. Model 2 includes the effects of SMOs (automobile clubs) and evangelical claims. The effect of time since regulation is now insignificant, and the number of in-state firms has significant and negative effects. Urbanization has significant positive effects. The number of SMOs (auto clubs) significantly increases car sales, and there is support for H1. Evangelism (reliability contests) significantly boosts car sales, and H2 is endorsed. There is a significant increase in the chi-square, thereby indicating that this model fits the data better than Model 1. (In unreported analyses, I also estimated models without in-state firms and found an identical pattern of support for H1 and H2. Additionally, I also estimated subsequent models reported in Table 2 without in-state firms and the support for predictions was identical.)

Model 3 includes an interaction between the number of SMOs and evangelism. The main effect of SMOs is positive and significant, but the main effect of evangelism is insignificant. However, the $\text{SMO} \times \text{evangelism}$ interaction term is significant and positive. Thus, there is support for the prediction that activist organization enhances the effect of evangelical claims on car sales. Model 4 inserts an interaction between the number of SMOs and advertising. The main effect of advertising is now significant and positive as is the main effect of SMOs. The effect of $\text{SMO} \times \text{advertising}$ interaction is significant and negative, and there is confirmation of the prediction that advertising weakens the effect of SMOs on sales (H4).

Model 5 includes an interaction between evangelism and advertising. The main effect of advertising is insignificant but the main effect of evangelism is positive and significant. The $\text{evangelism} \times \text{advertising}$ interaction is significant and negative, thereby supporting the hypothesis (H5) that the effect of evangelical claims on car sales wanes with advertising. Model 6 adds an interaction between the number of SMOs and time since regulation in the focal state. The main effect of SMOs is positive and significant, as is the main effect of time since regulation. The $\text{SMO} \times \text{time since legislation}$ interaction term has significant and negative effects and there is endorsement of the prediction (H6) that the effects of SMOs would decline as time since regulation rose. Model 7 inserts an interaction between evangelism and time since regulation in the focal state. While the main effect of evangelism is significant and positive, the main effect of time since regulation is insignificant. The $\text{evangelism} \times \text{time since regulation}$ interaction effect is significant and negative, albeit at the 0.10 level, and H7 is confirmed.

Table 2. Random Effects MLE Models of State-Level Car Sales.

Variables	Model Numbers							
	1	2	3	4	5	6	7	8
Constant	-7.85*** (3.70)	-10.28*** (3.52)	-10.15*** (3.41)	-9.50*** (3.84)	-10.14*** (3.51)	-9.78*** (3.50)	-10.21*** (3.51)	-8.54*** (3.43)
State Gross Product	0.323*** (0.130)	0.228** (0.123)	0.229** (0.123)	0.208* (0.122)	0.226* (0.123)	0.234* (0.122)	0.227* (0.123)	0.215* (0.121)
Urbanization	2.44 (1.64)	3.56** (1.55)	3.54** (1.52)	3.22** (1.53)	3.49** (1.55)	3.27* (1.56)	3.54** (1.52)	2.79* (1.51)
No. of in-state firms	0.215*** (0.049)	-0.151*** (0.059)	-0.168*** (0.059)	-0.125*** (0.058)	-0.150*** (0.058)	-0.115*** (0.059)	-0.147*** (0.058)	-0.118*** (0.054)
Advertising	-0.871 (1.47)	-1.25 (1.38)	-1.37 (1.38)	10.06*** (2.85)	0.485 (1.63)	-1.19 (1.38)	-1.21 (1.38)	9.71*** (3.15)
Time since legal recognition	0.139*** (0.045)	0.006 (0.047)	-0.014 (0.047)	0.026 (0.047)	-0.008 (0.047)	0.333** (0.122)	0.035 (0.053)	0.460*** (0.143)
No. of SMO's		0.960*** (0.095)	0.944*** (0.095)	1.09*** (0.098)	0.955** (0.095)	0.988*** (0.095)	0.944*** (0.096)	1.08*** (0.099)
Evangelism		0.188** (0.085)	-0.507 (0.309)	0.230*** (0.084)	0.315*** (0.106)	0.229*** (0.082)	0.327*** (0.122)	-1.38*** (0.352)
SMO's × Evangelism			0.184*** (0.078)					0.451*** (0.090)
SMO's × Advertising				-3.77*** (0.835)				-4.09*** (1.12)
Evangelism × Advertising					-1.25** (0.620)			0.820 (8.43)
SMO's × Time since legal recognition						-0.097*** (0.032)		-0.118* (0.042)
Evangelism × Time since legal recognition							-0.034* (0.021)	-0.031 (0.025)
p	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1
Log Likelihood	-2098.96	-2049.47	-2046.74	-2039.37	-2047.48	-2044.94	-2048.22	-2025.98
Chi-Square	118.95***	217.93***	223.39***	238.12***	221.92***	226.98**	220.22***	264.90***

Note: The p statistic for each equation has been rounded off to the first decimal. The chi squares for all models are vis-à-vis a baseline model with only the intercept. *** $p < 0.01$, ** $p < 0.05$, * $p > 0.10$ (Two tailed tests for all variables).

When all interactions are simultaneously examined in Model 8, there is no support for H5 and H7, but support for the other predictions endures in the fully-saturated model. Note that the inclusion of several interaction terms generates collinearity in Model 9, and therefore, the individual models are used to interpret the results.

Robustness Checks

In analyses not reported for the sake of brevity, I conducted several tests to assess the robustness of the results, and sought to rule out counter-arguments and specification problems.

I began by exploring whether the results survived the inclusion of additional controls such as population, industrialization, and lagged sales in the focal state. I found that population was highly correlated with urbanization, and industrialization was highly correlated with state gross product, and hence, did not merit inclusion. The insertion of lagged sales improved the fit of the models, albeit marginally, but did not alter the basic pattern.

An important counter argument is that the presence of automobile clubs, the number of reliability contests, the number of car firms, and elapsed time since regulation in a focal state are endogenously related to the dependent variable – car sales in a focal state. When a variable is suspected to be endogenous, Davidson and MacKinnon (1993) suggest an augmented regression test (Durbin-Wu-Hausman test), in which the residuals of each suspected endogeneous variable are first obtained as a function of all exogenous variables, and then are included in a regression of the original model. If the residuals are significantly different from 0, then instrumental variables are to be used, and a simultaneous equation approach is to be employed. I ran regressions that explicitly controlled for heteroscedasticity of state-year observations, obtained the residuals of all the suspected variables, and included them in a regression of sales. I found that endogeneity was not an issue since none of the residuals were significantly different from zero. Among these variables, the residuals for auto clubs had the lowest p level. Even though the effect was insignificant, as a precaution, I employed a system of structural equations and developed an instrumental variable for auto clubs. I used three-stage least-squares techniques and re-estimated Model 2, and found that the results were identical.

I checked whether the results would differ if each reliability contest was weighted by its visibility. Since inspecting each state's newspapers and then calculating a measure of visibility was time consuming and expensive, I computed the visibility of each contest in the most popular automobile magazine – *Horseless Age*. The visibility of reliability contests was operationalized as the frequency of reliability contests, weighted by the media exposure, and divided by the frequency of all contests (reliability and speed) weighted by their media exposure. When I checked the correlation between reliability contests/all contests in a focal state and visibility of reliability contests/visibility of all contests in a focal state, it was 0.94. Therefore, it is unlikely that would one obtain vastly different results if the effects of reliability contests were weighted by visibility.

I also sought to ascertain whether the advent of the gasoline engine as the dominant design influenced car sales. Industry observers indicate that the gasoline engine became accepted in 1903, but the inclusion of an explicit control for period effects did not appreciably alter the results. Moreover, some writers on the automobile industry refer to anti-speed vigilante organizations, and imply that they may have dampened sales (Flink, 1970). I searched the *New York Times* and *Horseless Age* but was only able to turn up a handful of ephemeral organizations, and hence, was unable to include them in the analysis.

Finally, a potential counter-argument is that sales are truncated at zero, and cannot be non-negative and therefore, should be modeled using a tobit specification. If this were so, then the main effects of SMOs and evangelism should be biased. I used random-effects tobit models to re-estimated Model 2 and did not find any appreciable change; I found that additional models indicated a similar pattern of results to those reported in Table 2. On balance, these checks imply that the results reported in Table 2 are robust.

DISCUSSION

This study illuminates the parallel routes to the constitutive legitimation of new industries. The findings show that the public-decentralized realm (Ingram & Clay, 2000) is an important domain for the formation of constitutive legitimacy. The results indicate that as the number of SMOs (auto clubs) in a focal state increased, sales of cars also significantly increased in the focal state. Hence, one implication of the results is that activism matters even when the establishment of new industries resembles a consensus-oriented social movement. A unique attribute of this study was that it also measured the effect of evangelism directly. The results show that evangelism had substantial positive effects on constitutive legitimacy and sales of the new industry. Evangelical claims that took the form of reliability contests in a focal state exposed consumer audiences to the car in experiential ways, allayed doubts about the viability of the car, and led to greater consumer purchases. The results also suggest that the public-centralized realm is an important arena for the creation of constitutive legitimacy. Time since enactment of laws authorizing the car significantly increased car sales. However, advertising by firms did not have spill-over effects on sales, and so cast doubt on industry-level legitimacy being a spill-over from advertising. The implication is that the private-decentralized realm may not be significant when public-decentralized and public-centralized domains are consequential.

More importantly, the study suggests that there are interesting interdependencies among the parallel routes to constitutive legitimacy. Thus, the number of

SMOs augments the effect of evangelism, and their joint effect increases sales. The growth of auto clubs supplied a social infrastructure for legitimation that augmented the persuasive impact of evangelism, thereby pointing to synergy in the public-decentralized realm of action. These findings connect neo-institutional accounts of activism with cognitive accounts of framing in the social movement literature (Snow & Benford, 1992), and underline how activists are cognitive entrepreneurs who mine available cultural resources to persuade audiences.

The findings also provide concrete evidence that advertising and laws dampen the effect of SMOs and evangelism. As advertising increased and promoted consumer awareness, the impact of auto clubs and reliability contests on consumer sales in a focal state declined, thereby suggesting that producer-sponsored advertising renders activism and evangelism superfluous. Moreover, the effects of SMOs and evangelism also witnessed a decline as conceptions of the automobile became stabilized over time. As time since legal recognition of the car in a focal state increased, the effects of auto clubs and reliability contests on sales in the focal state also diminished. Thus, as constitutive understandings of the automobile became crystallized by law, the effects of SMOs and evangelism weakened. These results suggest that the social organization of activists and evangelism matter primarily in the social movement phase of constructing new industries, but decay as industries gain advertising exposure and legal footholds. Thus, there may be a process of succession where public-decentralized domains of action are critical initially but decline in importance as public-centralized and private-decentralized realms become consequential.

The findings also speak to the field of organizational ecology even if this study was not designed to test ecological predictions. On a general level, the findings impart empirical substance to speculations by ecologists that social movement entrepreneurs and evangelism may play an important role in early phases of new industries (Carroll & Hannan, 2000; Swaminathan & Wade, 2000). The findings also address a debate between density-dependence theorists and institutional ecologists. On the one hand, density-dependence theorists hold that legitimation and competition are density-dependent processes, and have shown that there is an inverted U-shaped linkage between density and foundings/growth rates of firms, and a U-shaped relationship between density and mortality (Carroll & Hannan, 2000). The great benefit of the density-dependent approach is that it is simple, generalizable and replicable. On the other hand, institutional ecologists have argued that density dependence researchers should explicitly consider the effect of endorsements to individual organizations (Baum & Oliver, 1992; Baum & Powell, 1995). However, both points of view emphasize *firm-level outcomes* such as foundings and failures of firms, and subscribe to a *producer-centric* conception of the legitimation of new industries.

This study suggests that it may be useful to directly study how new industries are legitimated in the eyes of important audiences such as consumers or investors, and complements ecological research on firm-level outcomes. More importantly, an implication of this study is that consumer activists may play key roles in the legitimation of new industries. To date, organizational sociology depicts consumers as “lumped together in an aggregate, in a passive role” (White, 1981, pp. 522–523) and defocalizes their role in building the identity of new industries. Cultural analyses of consumption suggest that consumers can and do play an important role in shaping the meanings of products and the definition of product quality (Kopytoff, 1988). User groups play an important role in defining the identity of operating systems and software in diverse sectors of the economy. Similarly, marijuana consumers are playing an important role in a bid for its legalization as a medicine. This study shows how consumer activists located in auto clubs played an important role in situating the automobile as a cultural object and integrating it in the prevalent culture through their evangelism. It suggests that future work on legitimation needs to not only consider the role of consumer activists but also be attentive to the role of suppliers in enhancing constitutive legitimacy of a new industry, and the role of technological rivals in delegitimizing the new industry.

The findings are also of relevance to sociologists of consumption. To date, sociological analyses of consumption have shown how advertising and mass media play an important role in the social construction of new products (Schudson, 1989; Marchand, 1985). But sociological analyses of consumption have overlooked how “market interfaces” are created between producers and consumers (Frenzen, Hirsch & Zerrillo, 1996). The study of how novel artifacts are turned into social facts is essential in order to understand the market interface between consumers and producers, and this study shows that social movement activists and their evangelism influenced meanings of the automobile.

The limitations of this study point to directions for profitable research. This study analyzed how social movement processes contributed to consumer acceptance but did not analyze effects on other audiences such as investors and potential employees. Future research needs to show how social movement processes reduce uncertainty in capital and labor markets and enhance the generality of findings. This study analyzed the role of activists in new industries that emerged from a consensus-oriented social movement, but future research needs to study how activists and evangelism affect constitutive legitimacy in industries born through conflict-oriented movements such as craft brewing or alternative dispute resolution.

Moreover, this study focused on the number of SMOs but did not study how the social composition of the membership of SMOs influenced constitutive

legitimacy and consumer sales. A useful complement consists of exploring how membership of SMOs and their demographic structure influences the legitimation of new industries. Moreover, this study unraveled the effect of one type of evangelism – demonstration events. A valuable extension consists of analyzing whether the results are generalizable to other types of demonstration events such “bakeoffs” and benchmark tests, and studying report cards or public interest advertising designed to promote the image of the industry.

This study examined evangelism by consumer activists when professional societies were moribund and trade associations were unsuccessful. A useful elaboration would be to study the effects of evangelism by activists drawn from the ranks of producers and professionals. A profitable direction is to compare the evangelical repertoires of claim-making used to legitimate and delegitimize organizational forms. For example, the conglomerate form of organization was partly delegitimated through ‘bust-up acquisitions’ initiated by take-over specialists, and by denunciations of the firm-as-portfolio model by business professors in magazines targeting senior managers (Davis, Diekmann & Tinsely, 1994).

Finally, this study only looked at the actions of activists in the early stage of an industry, but research on the effects of activists and evangelism in the later stages of industries is required. After the automobile was constitutively legitimated, institutional activists turned to the question of improving roads in the mid-1920s, and then at the height of the industry’s growth they highlighted the issues of product safety. Thus, institutional entrepreneurs ceased to be evangelists for the car and instead, became critics of the industry. Critics may lobby for the creation of new laws as temperance activists did in the case of the beer industry (Swaminathan, Wade & Saxon, 1998). Future research needs to show how activist-critics degrade investor and consumer perceptions, and induce counter-responses from producers. Research into the ideological competition between producers and critics sponsoring social movements is essential for advancing the frontiers of economic sociology.

ACKNOWLEDGMENTS

I am grateful to Paul Ingram and Brian Silverman for incisive suggestions that greatly improved the paper. I am also grateful to Emory University for providing funds for the research, and to Ron Harris for superb data assistance. Farah Mihoubi did an outstanding job copy-editing the paper.

NOTES

1. As the quote implies, reliability contests were opportunities to demonstrate the capabilities of the horseless vehicle. This demonstration provided firms a chance to win and to acquire a reputation. Note that it was firm-level reputation that was the positive spill-over and not the legitimacy of the automobile. The latter was the object of deliberate action by auto clubs.

2. For example, the Remington Arms Company in 1895 (coincidentally, the year in which the auto industry was born with the creation of the Duryea firm) proudly announced that the winner of the Irvington-Millburn Road race had used one of their "scientifically constructed bicycles" (Smith, 1972, p. 34).

3. Other writers urged for the testing of products before the 1890s, but it was only after 1890 that the logic became institutionalized when it was embodied in testing and standards organizations. These organizations catered to producers rather than consumers. Special-purpose testing agencies for consumers appeared in the late 1920s in the form of non-profit consumer watchdogs.

REFERENCES

- Aldrich, H. E. (1999). *Organizations evolving* (forthcoming). Beverly Hills: Sage.
- Baum, J. A. C., & Oliver, C. (1992). Institutional embeddedness and the dynamics of organizational populations. *American Sociological Review*, 57, 540–549.
- Baum, J. A. C., & Powell, W. W. (1995). Cultivating an institutional ecology: Comment on Hannan, Carroll, Dundas and Torres. *American Sociological Review*, 50, 529–538.
- Berger, P., & Luckmann, T. (1966). *The social construction of reality*. New York: Free Press.
- Belsley, E. K., & Welsch, R. E. (1980). *Regression Diagnostics*. New York: Wiley.
- Caroll, G. R., & Hannan, M. T. (2000). *The demography of corporations and industries*. Princeton, NJ: Princeton University Press.
- Cerna, M. (1993). New performance measures will yield comparative data on HMOs. *Hospitals & Health Networks*, 67, 48.
- Charles, J. (1993). *Service clubs in American society*. Chicago: University of Illinois Press.
- Davidson, R., & MacKinnon, J. G. (1993). *Estimation and inference in econometrics*. New York: Oxford University Press.
- Davis, G. F., Diekmann, K., & Tinsley, C. (1994). The decline and fall of the conglomerate firm in the 1980s: The deinstitutionalization of an organizational form. *American Sociological Review*, 59, 547–570.
- DiMaggio, P. (1988). Interest and agency in institutional theory. In: L. Zucker (Ed.), *Institutional Patterns and Organizations: Culture and Environment* (pp. 3–21). Cambridge, MA: Ballinger.
- Edelman, L., & Suchman, M. (1997). The legal environment of organizations. *Annual Review of Sociology*, 23, 479–515.
- Edelman, L. (2000). Constructed legalities: The endogeneity of law. In: W. W. Powell & D. Jones (Eds), *Bending the Bars of the Iron Cage*. Chicago: University of Chicago Press.
- Epstein, R. C. (1928). *The automobile industry: Its economic and commercial development*. Chicago: A. W. Shaw & Co.
- Fligstein, N. (1996). Markets as politics: A political cultural approach to market institutions. *American Sociological Review*, 61, 656–673.

- Flink, J. J. (1970). *America adopts the automobile, 1895–1910*. Cambridge: MIT Press.
- Flink, J. J. (1988). *The automobile age*. Cambridge: MIT Press.
- Hannan M. T., & Freeman, J. (1989). *Organizational ecology*. Cambridge, MA: Belknap Press.
- Horseless Age* (various issues).
- Ingram, P., & Inman, C. (1996). Institutions, intergroup rivalry and the evolution of hotel populations around Niagara Falls. *Administrative Science Quarterly*, 41, 629–658.
- Ingram, P., & Clay, K. (2000). The choice-within-constraints new institutionalism and implications for sociology. *Annual Review of Sociology*, 20, 525–546.
- Iyengar, S., & Kinder, D. (1987). *News that matters*. Chicago: University of Chicago Press.
- Kimes, B. R., & Clark, H. (1989). *Standard catalogue of American cars* (2nd ed.). Iola, WI: Krause Publications.
- Kinder, D. (1998). Communication and public opinion. *Annual Review of Political Science*, 1, 167–197.
- Klandermans, B. (1992). The social construction of protest and multi-organizational fields. In: A. Morris & C. Mueller (Eds), *Frontiers in Social Movement Theory* (pp. 77–103). New Haven, CT: Yale University Press.
- Knight J., & Ensminger, J. (1998). Conflict over changing social norms: Bargaining, ideology and enforcement. In: V. Nee & M. Brinton (Eds), *The New Institutionalism in Sociology*. New York: Russell Sage.
- Kriesi, H. (1996). The organizational structure of new social movements in a political context. In: D. McAdam, J. D. McCarthy & M. N. Zald (Eds), *Comparative Perspectives on Social Movements: Political Opportunities, Mobilizing Structures, and Cultural Framings* (pp. 152–184). Cambridge, England: Cambridge University Press.
- Leeflang, P., Wittink, D., Wedel, M., & Naert, P. (2000). *Building models for marketing decisions*. Leiden: Kluwer Publishers.
- Lofland, J. (1985). *Protest: Studies of collective behavior and social movements*. New Brunswick: Transaction Books.
- McAdam, D., McCarthy, J. D., & Zald, M. N. (1996). Introduction: Opportunities, mobilizing structures, and framing processes-Toward a synthetic, comparative perspective on social movements. In: D. McAdam, J. D. McCarthy & M. N. Zald (Eds), *Comparative Perspectives on Social Movements: Political Opportunities, Mobilizing Structures, and Cultural Framings* (pp. 1–20). Cambridge, England: Cambridge University Press.
- McCarthy, J. D. (1996). Constraints and opportunities in adopting: Adapting and inventing. In: D. McAdam, J. D. McCarthy & M. N. Zald (Eds), *Comparative Perspectives on Social Movements: Political Opportunities, Mobilizing Structures, and Cultural Framings* (pp. 142–184). Cambridge, England: Cambridge University Press.
- McCarthy, J. D., & Zald, M. N. (1977). Resource mobilization and social movements: A partial theory. *American Journal of Sociology*, 82, 1212–1241.
- Meyer, J. W., & Rowan, B. (1977). Institutionalized organizations: Formal structures as myth and ceremony. *American Journal of Sociology*, 83, 340–363.
- Meyer, M. W. (1994). Measuring economic performance. In: N. Smelser & R. Swedberg (Eds), *Handbook of Economic Sociology* (pp. 556–578). New York: Russell Sage.
- Moe, T. M. (1980). *The Organization of Interests*. Chicago: University of Chicago.
- Mowery, D., & Rosenberg, N. (1998). *Paths of innovation: Technological change in 20th-century America*. New York: Cambridge University Press.
- Rao, H., Morrill, C., & Zald, M. (2000). Power plays: Social movements, collective action and new organizational forms. In: B. Staw & R. Sutton (Eds), *Research in Organizational Behavior*, 22 (forthcoming).

- Schudson, M. (1989). How culture works: Perspectives from media studies on the efficacy of symbols. *Theory and Society*, 19, 153–180.
- Smith, R. A. (1972). *A social history of the bicycle: Its early life and times in America*. New York: Heritage Press.
- Staggenborg, S. (1991). *The pro-choice movement*. New York: Oxford University Press.
- Snow, D. A., Rochford, E. B., Jr., Worden, S., & Benford, R. D. (1986). Frame alignment processes, micromobilization, and movement participation. *American Sociological Review*, 51, 464–481.
- Snow, D. A., & Benford, R. D. (1992). Master frames and cycles of protest. In: A. Morris & C. Mueller (Eds), *Frontiers in Social Movement Theory* (pp. 133–155). New Haven, CT: Yale University Press.
- Stinchcombe, A. L. (1965). Organizations and social structure. In: J. G. March (Ed.), *Handbook of Organizations* (pp. 193–253). Chicago: Rand McNally.
- Suchman, M. (1995). Managing legitimacy: Strategic and institutional approaches. *Academy of Management Review*, 20, 571–610.
- Swaminathan, A., & Wade, J. B. (2000). Social movement theory and the evolution of new organizational forms. In: C. B. Schoonhoven & E. Romanelli (Eds), *The Entrepreneurship Dynamic in Industrial Evolution*. Stanford, CA: Stanford University Press.
- Swidler, A. (1986). Culture in action: Symbols and strategies. *American Sociological Review*, 51, 273–286.
- Tilly, C. (1978). *From mobilization to revolution*. Reading, MA: Addison-Wesley.
- Thomas, R. P. (1977). *An analysis of the patterns of growth of the automobile industry, 1895–1929*. New York: Arno Press.
- Van de Ven, A. H., & Garud, R. (1989). A framework for understanding the emergence of new industries. In: R. Rosenbloom & R. Burgelman (Eds), *Research in Technological Innovation, Management and Policy* (pp. 195–225). Greenwich, CT: JAI Press.
- Zelizer, V. (1979). *Morals and markets, the development of life insurance in the United States*. New York: Columbia University Press.
- Zucker, L. G. (1983). Organizations as institutions. In: S. Bachrach (Ed.), *Research in the Sociology of Organizations* (pp. 1–42). Greenwich, CT: JAI Press.

This Page Intentionally Left Blank

PART 4:
WHAT DETERMINES SUCCESS
AND FAILURE IN INTERNATIONAL
COMPETITION?

This Page Intentionally Left Blank

LEARNING ABOUT THE INSTITUTIONAL ENVIRONMENT

Witold J. Henisz and Andrew Delios

ABSTRACT

Cross-national variation in institutional environments adds uncertainty to foreign operations, which in turn affects international strategy decisions such as when to enter a market, the entry mode used if entering, as well as the performance of foreign entries. Although all firms are exposed to the influence of a host country's institutional environment, firms exhibit differential responses to this influence based on resident knowledge and capabilities. Managers in a multinational firm must therefore work to align their strategies with both the hazards and opportunities they face in a given institutional environment, as well as with the firm-specific knowledge and capabilities at their disposal. Rather than taking institutions as an immutable constraint when making decisions, a firm can cultivate and exploit its ability to successfully manage diverse institutional hazards in its host country environments.

INTRODUCTION

Whether a multinational firm expands abroad for purposes of exploitation of its existing resources or exploration for new resources, a subsidiary's operations and its success in managing those operations are encumbered by cross-national and inter-temporal variation in institutional environments. The

institutional environment, as we define it here, includes political institutions such as the national structure of policy making, regulation and adjudication; economic institutions such as the structure of the national factor markets and the terms of access to international factors of production; and socio-cultural institutions such as informal norms.

Research examining the link between these various components of a nation's institutional environment and multinational entry strategies has focused on multinationals' responses to the varying levels of hazards and opportunities presented by the institutional environment of a potential host country. This research has, however, tended to treat firms as homogeneous in their characteristics and responses to specific configurations of institutional environments. Yet, a key development in strategic management has been the recognition of the strategic and performance implications of knowledge and capability diversity across firms. By the same token, the developing research tradition on the strategic and performance influences of knowledge and capability heterogeneity among populations of firms has not examined how multinationals with different sets of knowledge and capability respond to variance in the institutional environment.

In this paper, we contend that both streams of research have much to gain from closer collaboration. The framework we outline in this study develops from a review of literature on national institutional environments, and on experiential learning in multinational firms. In combining the insights from these two literatures, we argue a firm's investment experience contributes to the development of knowledge that can be used to mitigate hazards encountered when making direct investments abroad. In this sense, we conceptualize knowledge as a firm-specific resource that moderates the influence that country-level and inter-temporal variations in institutional environments have on a multinational's strategic decisions. In describing these influences, we make reference to several market entry decisions including whether to invest in a market, the mode of investment and scale of investment. We also discuss the performance implications of these decisions.

Underlying our analysis is the presumption that a firm's managers will act with foresight to mitigate hazards and maximize the opportunities present in a multinational's multiple institutional environments. We develop this discussion around entry, entry mode and investment size decisions because these are highly visible forms of a firm's international investment strategy. Likewise, pursuing arguments to the level of firm performance permits us to draw stronger implications about the antecedents and effects of these decisions although, as we note, empirical modeling of performance influences should proceed with caution.

We acknowledge there are many fruitful research questions that examine the strategic and performance influences of firm heterogeneity, including knowledge

and international capabilities, as well as of cross-national and intertemporal variation in institutional environments. However, our discussion also features several extensions to this literature we think are worthy of attention. First, researchers in economics, political science, sociology and other fields have different conceptions of the institutional environment and the relative importance of these various components. We present some of this diversity in perspectives that seek to stimulate further research into the operationalization of the institutional environment. We likewise recognize, and believe, that the mechanisms of knowledge and capability development are poorly understood especially in the international context. A multinational firm has multiple points of learning about institutional environments including learning within various business networks and learning from the actions of competitors. Where learning does occur, it sometimes remains localized, and fails to flow throughout the organization. Hence, the potential relevance of knowledge or capabilities developed in one context for application in another is often unrealized. We also highlight fruitful extensions that further delineate the mechanisms and limits to learning in the international context.

While we present our arguments regarding the role of knowledge and capabilities in moderating hazards and securing advantage from diversity in institutional environments in the context of the market-based strategies of entry, entry mode and investment scale, we recognize that firms can take many other steps to mitigate hazards and improve their performance outcomes. Prominent among these are non-market strategies that involve activities in the political, regulatory or legal arenas designed to secure competitive advantage for the firm. We also discuss how non-market strategies such as lobbying, litigating and other influence activities can function as complementary means of hazards mitigation.

In describing our perspective on the intersection of research on institutional environments, experiential learning and multinational strategy, we hope to stimulate new research in this field. We believe that our straightforward extension of the arguments surrounding the strategic importance of the institutional environment and the role of firm knowledge and capabilities has received insufficient attention from theorists and empiricists alike.¹ While the data collection challenges required to demonstrate support for these arguments are substantial, the payoff in terms of refining our theoretical understanding of these phenomena and enhancing the practical applicability of these theories warrants further effort.

INSTITUTIONAL ENVIRONMENT

To successfully deploy its resources in a new country, a multinational firm must identify and contend with numerous differences between the host country market

and the markets in which it has previously operated (Beamish, 1988; Hymer, 1976; Martin, Swaminathan & Mitchell, 1998; Zaheer, 1995). Laws and regulations surrounding the acquisition of property, licensing of new businesses, hiring of workers, importing factors of production, exporting output or capital, contracting with suppliers for needed inputs, payment of taxes, government licenses and fees, and the means and feasibility of exit all vary widely across countries. Even where laws and regulations appear similar, differences in legal systems can have important differences in such relevant outcomes as the protection afforded to shareholders vs. creditors or minority investors (La Porta et al., 1998, 1999). A recent study that examined regulations of entry in 75 countries found that the official procedures and costs required to start a new business varied from as few as two procedures (Canada) taking two days (Canada) and 0.4% of average per capita income (New Zealand) to as many as 20 procedures (Bolivia) requiring up to 174 days (Mozambique) and 263% of average per capita income (Bolivia) (Djankov et al., 2000).

In addition to the costs of starting a new business, political and regulatory hurdles can impose substantial costs on the day-to-day operations of an overseas subsidiary. Where policy makers can act unilaterally or have high certainty that a subservient or allied legislature and judicial branch will support their actions, future policies are likely to be more volatile in response to either exogenous shocks, changes in the identity of policymakers or even changes in the preferences of the existing policy makers. Such changes that are the result of direct lobbying by host country competitors or incumbents are of particular concern to multinationals.

Cultural differences between nations can also influence multinational entry strategies. For example, substantial evidence of cross-national differences in such factors as "administrative practices and employee expectations" (Kogut & Singh, 1988, p. 414) or the adoption of three paradigms of organizational management: scientific management, human relations and structural analysis (Guillén, 1994), illustrate the difficulty with which models of management are transferred from one nation to another (Guillén, 2001). Empirical evidence supporting these differences is found in the studies of Hamilton and Biggart (1988) on the industrial arrangements and organizational strategies employed in the economies of South Korea, Taiwan and Japan, and in Biggart and Guillén (1999) who looked at the evolution of the auto industry in South Korea, Taiwan, Spain and Argentina.

Institutional differences between nations magnify difficulties in collecting, interpreting and organizing the relevant information necessary to mount a successful entry. Markets that are similar in political structure, factor market structure or culture pose less uncertainty, relatively lower costs of entry and,

therefore, lower hurdle rates of return. Investors are hence more likely to enter countries where the future policy regime is relatively easy to predict (Bennet & Green, 1972; Gastanaga, Nugent & Pashamova, 1998; Green & Cunningham, 1975; Loree & Guisinger, 1995; Root & Ahmed, 1978; Stobaugh, 1969; Vernon, 1977; Wei, 2000). Relatedly, investors are more likely to enter countries that are culturally similar, and have similar organizational structures (Hanson, 1999; Loree & Guisinger, 1995).

The impact of institutions on entry strategies can extend beyond the choice of where to invest to include the design of the subsidiary. For example, firms that perceive hazards emanating from policy uncertainty can take hazard-mitigating actions designed to shift the decision calculus of the potential expropriating government. Such a shift should seek to either raise the political and/or economic costs (lost revenue, employment and future investment) to asset or revenue expropriation, or lower the benefits (the value of seized assets or revenue stream and the nationalist political reaction) from expropriation.

As an example of a hazard mitigating strategy, a multinational can form a partnership with a host country firm. In such a partnership, the host country firm would also suffer in the event of an expropriation of a foreign subsidiary's returns or its assets because of the subsidiary's dependence on a continuing relationship with its parent firms for its complementary assets. In exchange for ownership in the foreign subsidiary, a host country partner provides a valuable service. Host country firms tend to use, on average, a greater percentage of domestic inputs. Because of superior information regarding the availability of, terms of and procedures for acquiring goods in the domestic market, host country partners rely more heavily than the multinational on domestically-sourced labor, intermediate products and trading partners. While a multinational may pay to acquire this form of local information, pursuing such a strategy raises production costs relative to domestic firms.

Depending on the size of this cost wedge, a multinational shifts some positive quantity of inputs from domestic to international sourcing. Expropriation of the assets or revenue stream of a joint venture between a multinational and a host country partner will therefore result in a greater expropriation of assets or revenue streams owned by domestic constituents than expropriation of a solely foreign venture. As more domestic constituents are implicated in the expropriation, a partnership between a multinational and a host country firm is, on average, politically more costly to expropriate for the government than a solely foreign venture.² Gatignon and Anderson (1988), Hill, Hwang and Kim (1990), Agarwal and Ramaswami (1992), Oxley (1999), Delios and Beamish (1999), Smarzynska (2000) and Henisz (2000) provide empirical evidence supporting this link between market entry mode choice and the degree of policy

uncertainty.³ Similarly, Goodnow and Hansz (1972), Davidson and McFetridge (1985), Kogut and Singh (1988) and Erramilli (1996) find an analogous link between cultural proximity and the choice of entry mode.

In addition to the choice of entry mode, the type of assets chosen for overseas investment may vary across different institutional structures. Empirical work that has adopted a case study approach provides strong support for the hypothesis that long-lived and/or politically visible investments such as those in infrastructure sectors will be particularly sensitive to a country's institutional environment (Bergara Duque, Henisz & Spiller, 1998; Caballero & Hammour, 1998; Crain & Oakley, 1995; Dailami & Leipziger, 1998; Daniels & Trebilcock, 1994; Grandy, 1989; Levy & Spiller, 1994; 1996 ; Ramamurti, 1996; Savedoff & Spiller, 1997; Spiller & Vogelsang, 1996; Williamson, 1976). Two recent efforts to extend this logic to panel datasets in telecommunications (Henisz & Zelner, 2001) and electricity (Henisz & Zelner, 2002) have also found strong support for the hypothesis that political institutions that fail to constrain arbitrary behavior by political actors dampen the incentive for infrastructure providers to deploy capital and, *ceteris paribus*, yield lower levels of per capita infrastructure investment.

Unstable policy regimes and culturally dissimilar markets are likely associated with performance penalties as well. Carroll and Delacroix (1982) find higher mortality rates for newspapers in Argentina and Ireland during periods of political unrest. Zaheer and Mosakowski (1997) find that the failure rates of foreign firms in the financial service sector are higher in more tightly regulated and less globally integrated markets. Li and Guisinger (1991) find an analogous relationship between cultural proximity and survival as do Barkema, Bell and Pennings (1996) and Barkema and Vermeulen (1997).

While the empirical literature broadly supports hypotheses regarding the effect of institutional differences on the timing and location of entry, the choice of entry mode, the magnitude of investment, the dynamics of market entry strategies and the probability of survival by the average multinational firm, a wide variety of international expansion strategies and performance outcomes are observed in any given population within a given institutional environment. One reason for this observed heterogeneity in entry strategies is variation in the extent to which a given host country market is "similar" to those previously encountered by a firm. Each firm makes decisions about the timing and location of entries based not only upon the extent of similarity between the host country market and the previous markets in which it has operated but also on the firm's knowledge of the host country. To the extent that this knowledge base differs across firms, so should observed entry strategies.

Examination of the variation in the institutional environment in the absence of a similar treatment of heterogeneity in the population of firms is necessarily incomplete. Studies examining entry, entry mode or investment scale decisions as function of institutional similarity should account for within-population variation in the familiarity with a given country or institutional profile. We turn to an elaboration of these within population differences.

EXPERIENTIAL LEARNING

The assumption that all firms in a country are equally influenced by institutional factors is inconsistent with the research tradition on the effects of international experience on market entry strategy (Eriksson et al., 1997; Huber, 1991; Yu, 1990). The internationalization literature contends that differences in firms' knowledge bases extend from differing profiles of international experience. Differences in firms' knowledge bases are also thought to be an important source of observed heterogeneity in firm strategies, especially in the international arena.

The experience literature is rooted in Johanson and Vahlne's (1977) postulated expansion path in which a firm moves sequentially from operations that entail relatively low levels of resource commitment in relatively similar markets to operations with higher levels of resource commitment. This shift occurs as a firm gains experience and knowledge of the intricacies in its host country markets. Over time, experience and knowledge accretion mean that a multinational firm becomes less foreign and less dependent on its local partners. At the same time, it begins to operate in host countries as it does in the home country or other similar host country markets. Experiential knowledge – country specific knowledge based on a firm's activities in a host country – is the linchpin in this process because it provides substantially higher benefits than objective knowledge to firms seeking to collect, interpret and use incoming information obtained from a dissimilar environment. The speed with which firms enter countries and progress down this commitment chain is positively correlated with the similarity of the host country to the firm's home environment.

Aharoni (1966), Kindleberger (1969), Wilkins (1970) and Johanson and Wiedersheim-Paul (1975) provide qualitative support for these arguments about the sequence of entry for U.S., British and Swedish multinational firms, respectively. Statistical analysis has been less supportive of the basic assertions of the stages model. While Davidson (1980) and Yu (1990) show that U.S. foreign direct investment is most likely in large "similar" markets where a firm had prior experience, Sullivan and Bauerschmidt (1990), Benito and Gripsrud (1992) and Pedersen and Shaver (2000) did not find support using samples

of European forestry firms, Norwegian manufacturing firms and Danish manufacturing firms, respectively.

Building on these insights, researchers have argued that a firm's international expansion strategy consists of a series of integrated choices that capitalize on the best market opportunities and take advantage of experiential learning (Kogut, 1983). Experience acquired by prior investments helps to develop new capabilities, and these new capabilities affect the way a firm evaluates its ownership position when making its next investments (Chang, 1995). The motivations for subsequent entries may be qualitatively different than those for a firm's initial investment (Chang & Rosenzweig, 2001). Hence, one motivation for a firm to enter a host country market might be to acquire experiential knowledge that can foster the development of capabilities used to support future entries. This idea underlies the internationalization literature in which low commitment investments to a market serve as knowledge stepping stones for developing a better understanding about local culture and local consumer preferences (Johanson & Vahlne, 1977). The same logic could be applied to investment stepping stones used to acquire knowledge about the policy environment, with the added caveat that low commitment investments might not provide a significant enough deployment of senior managers to enable such learning to occur (Delios & Henisz, 2002). The dilemma then arises within a firm that to learn about a market a firm must make a significant commitment to that market, yet it must make that commitment when its experience in that market is comparatively low.

Extant literature examining the relationship between a firm's experience levels and its commitment to a market, in terms of equity ownership positions assumed in subsidiaries, highlights the level of commitment aspect of this dilemma. This literature identifies that when a firm enters new product or geographic markets, the capabilities required to compete successfully in the new market can differ significantly from the ones required for success in the firm's existing markets. Under such conditions, it becomes incumbent on the firm to develop new capabilities suited to the market into which it has expanded. This need stems from the specific nature of a firm's routines (Nelson & Winter, 1982) and the bounded rationality of its managers (Simon, 1945), both of which impede deployment of a firm's capabilities outside of its current market contexts.

Because of the difficulty found in deploying existing capabilities in new markets, a firm may try to acquire required capabilities via partnerships with other firms that have experience in the market in which it is investing. In the context of the foreign direct investment decision, a firm inexperienced in a given host country may partner with local firms (Inkpen & Beamish, 1997; Teece, 1977) or, more broadly, may use alliances as a platform for knowledge development prior to making an entry decision (Kogut, 1991; Kogut &

Kulatilaka, 1994; Mitchell & Singh, 1992). Similarly, when a firm makes an investment outside of its existing product-markets, other firms with experience in that product area may be sought as partners to augment existing capabilities (Hennart, 1988; Mowery, Oxley & Silverman, 1996).

Results regarding the choice and evolution of entry mode have generally been supportive of hypotheses regarding escalating commitment in experience. Erramilli (1991) finds that after overcoming an initial reluctance to partner, firms with more experience tend to have more confidence in their own abilities to manage and evaluate market risks. Gatignon and Anderson (1988), Agarwal and Ramaswami (1992), Barkema and Vermeulen (1998), Mutinelli and Piscitello (1998), Delios and Beamish (1999), Oxley (1999) and Guillén (1999) all similarly find a positive relationship between experiential learning and the probability of adopting an entry mode characterized by greater commitment.

Turning to the link between experience and measures of performance, international experience can diminish a firm's liability of foreignness (Hymer, 1976; Zaheer, 1995) by expanding the scope of its knowledge of the institutions, culture, markets and language that constitute the local market environment. International experience that is garnered in the host country contributes to the development of new capabilities, and this development influences a firm's strategy and performance (Barkema, Bell & Pennings, 1996; Pennings, Barkema & Douma, 1994). As host country and international experience generate general knowledge and capabilities applicable to a host country environment, multinational firms that have accumulated host country experience have a reduced scope of competitive disadvantages compared to local firms and face fewer operational difficulties in the local market (Delios & Beamish, 2001).

This reduction in the liability of foreignness has specific performance benefits. Li (1995) finds a positive relationship between experience and survival. Steensma and Lyles (2000) report a similar relationship between learning and survival as do Barkema, Bell and Pennings (1996) who also report that the effect is stronger for joint ventures than wholly owned subsidiaries. As argued by Lorenzoni and Lipparini (1999) and Simonin (1997) and demonstrated by Anand and Khanna (2000) and Delios and Beamish (2001), the capabilities developed by operating alliances in the past may enhance the value to a firm of subsequent alliances.⁴ Delios and Beamish (2001) also found a positive relationship between experience and survival, but experience did not have a direct effect with an alternate measure of performance, namely profitability. Instead, experience had an indirect effect in which it influenced subsidiary profitability by providing a means by which a firm's existing capabilities could be adapted to its subsidiary's new environment. Similarly, Mitchell, Shaver and Yeung (1993) report that firms with greater international experience are more

likely to survive an industry shake-out, but, the performance benefits of internationalization only accrue over time (Mitchell, Shaver & Yeung, 1992).

While these firm-level differences are clearly important determinants of multinational entry strategies, their elaboration and analysis has been developed in isolation from cross-national differences in institutional environments as outlined above and are thus subject to the critique of failing to treat cross-national heterogeneity with the same sophistication as within-population heterogeneity in experience. For example, studies of entry considered the aggregate level of experience but not the applicability of that experience to a given country's institutional environment. Similarly, studies of entry mode examined the prevalence of adoption of entry modes characterized by greater commitment as a function of experiential learning but failed to examine the extent to which the knowledge amassed by the parent firm was relevant to the entry mode of choice in a given host country. Finally, analyses of performance outcomes based on experience have often failed to take into account the appropriateness of entry and entry mode choices as well as the type of experience. Doing so, creates the risk of a selection bias when analyzing performance (Shaver, 1998).

THE INSTITUTIONAL ENVIRONMENT AND EXPERIENTIAL LEARNING

If experience does provide firms with competitive advantages in managing institutional differences, we should expect that the importance of experiential learning is especially acute in the most dissimilar environments. Learning is likely contextual (Brown & Digid, 1991). While firms might follow an expansion path from the most "similar" to the least "similar" institutional environments, similarity is a function not just of the institutional environment of the home country but also of the knowledge available to the parent firm regarding a potential host country. Entry mode choices will similarly be influenced not just by the similarity of the institutional environments of home and host countries but also by the extent to which the experiential learning of a multinational allows its managers to collect and interpret information from the potential host country in a way that makes the market more "similar." Finally, the performance effects of experiential learning will be greatest in those host countries with the largest *prima facie* differences in institutional environments.

While little research to date has examined these complex interrelationships, several studies examine the manner in which cross-national differences and population differences in experiential learning interact to influence multinational entry strategies. Using the metaphor of stepping stones, Barkema and Drogendijk

(2002) examine the incremental nature of internationalization from the home country to similar countries and then to countries that share institutional characteristics with the first wave of countries entered. Relatedly, Rhee and Cheng (2000) find that Korean manufacturing firms with greater international experience were less likely to expand into a new host country incrementally. Further, firms with higher levels of resource recoverability were less sensitive to their lack of experience in a host country market especially when entering countries with labor cost advantages relative to Korea.

Henisz and Delios (2001) find that Japanese multinationals tend to eschew politically hazardous countries, with the effect strongest for firms without experience in the host country and without experience gathered from international investments in other countries. These results demonstrate the importance of both the institutional environment and experiential learning as determinants of entry. Perhaps more importantly, the results suggest a causal mechanism by which experiential learning provides a competitive advantage to firms.

In the context of political hazards, one benefit enjoyed by experienced firms is a relative advantage in detecting and safeguarding against detrimental changes to the policy environment by host country governments. A specific government or a government with a specific institutional configuration can exhibit patterns in its behavior that a firm with prior experience in the same or similar countries may observe and react more expediently to than inexperienced counterparts. With the development of knowledge of these patterns, an experienced foreign firm can pursue appropriate hazard mitigating strategies, such as developing the ability to become more integrated into local factor markets with the gain of information about local buyers and suppliers. Should the government then consider implementing a policy that has deleterious effects on a foreign firm, a larger number of domestic constituents would also be harmed. This collateral effect would alter the political decision calculus of the host country government.

Aside from partnering with local firms, a second hazard mitigation strategy is identifying key government officials and developing better working relationships with these officials or with local partners who have access to and influence with them. A multinational might also develop the capability to choose local partners that have political influence and will use that influence to support the interests of the multinational. Finally, firms may employ non-market strategies including lobbying and influence tactics as discussed below.

While external sourcing, such as hiring a political risk consultant, is another feasible strategy to overcome political hazards, experienced firms have the additional lower cost option of drawing from their internal capabilities to implement an appropriate strategy (Burke & Casson, 1998). Furthermore, given industry and project-level variance in political hazards (Kobrin, 1979), the value

of externally-sourced information is questionable. Even under the assumption that such information is valuable to a firm, senior management may not be able to replicate the knowledge transfer that occurs among the operational management team via experience in similar countries.

Chung and Alacer (2002) develop another set of causal arguments that allows for heterogeneity among firms in response to institutional considerations. They find that the R&D intensity of a multinational is an important source of heterogeneity among foreign investors as it concerns the relative attractiveness of local wages for multinationals investing in the United States. In contrast to much of the literature that emphasizes the exploitation of existing capabilities, Chung and Alacer (2002) focus on the drivers of exploration. While the authors consider only one host country, their insight regarding the importance of heterogeneity in multinationals' technological capabilities parallels many of the arguments developed above.

Just as experiential learning or a multinational's resource profile can affect its entry decision, those same characteristics can influence the choice of entry mode. One specific example of this concerns the capability to detect and safeguard against opportunistic behavior on the part of host country governments. As mentioned earlier, patterns in behavior can emerge in specific governments or governments with specific institutional configurations, and a multinational can learn about these patterns. With the development of host country knowledge, a multinational can become more integrated in local factor markets as it becomes familiar with local buyers and suppliers, and it can develop capabilities more suited to dealing with local political actors. Hence, a firm with host country specific experience or experience in "similar" institutional environments should be less likely to share equity with a local partner in a politically hazardous environment than a less experienced counterpart. Delios and Henisz (2000) provide strong empirical support for this hypothesis in a sample of 2,827 manufacturing subsidiaries of Japanese multinational firms in 18 emerging markets.

Turning to the performance implications of these entry and entry mode decisions, there is no empirical evidence of which we are aware that links characteristics of the institutional environment to firm-level characteristics in the determination of subsidiary profitability or survival.⁵ Yet, at least two propositions emerging from this review could be subject to empirical testing to begin to build empirical evidence for these assertions. First, performance in relatively dissimilar markets should be higher for firms with greater experiential learning in, and more knowledge of, those markets. Firms that have actively participated in a given country through economic and political cycles and shocks should have a better understanding of the formal and informal mechanisms that link the institutional environment to market outcomes. This knowledge means

these firms are better able to influence those processes and thus should be expected to be relatively more profitable or more likely to survive.

Proposition 1: In host countries with dissimilar institutional environments compared to a firm's home country, subsidiary performance along dimensions such as survival and profitability will be greater, the greater a firm's experiential knowledge of the dissimilar market.

Second, this same set of knowledge, however, could become an encumbrance should the institutional environment change radically. Those firms with a broad range of experience should outperform their counterparts with relatively narrower country-specific experience in times of rapid institutional change. In a dynamic institutional environment, country-specific contacts and country-specific knowledge about how to get things done and influence processes may, in fact, become the wrong contacts and the wrong political levers to pull when there is radical change in the institutional environment. The capability to respond rapidly to change without being too closely tied to the old institutional environment is more likely a key determinant of performance.

Proposition 2: In dynamic institutional environments undergoing rapid institutional change, subsidiary performance along the dimensions of survival and profitability will be greater, the greater the breadth of a firm's experience across dissimilar national institutional environments.

FUTURE RESEARCH

While empirical research demonstrating differential sensitivity to variation in the institutional environment within a population of firms is relatively scant and deserving of further attention, there also exist a range of fruitful extensions to the research questions outlined above. These questions call for microanalytic measurement of the underlying constructs or an expansion of scope of the strategic choices examined.

Measurement of the Institutional Environment

As noted above, research on the institutional environment and multinational strategy has examined such diverse factors as political systems, legal systems,

labor market structures, and “culture.”⁶ Each of these measures is part of the institutional environment that has been defined as “the set of fundamental political, social and legal ground rules that establishes the basis for production, exchange and distribution. Rules governing elections, property rights, and the right of contract are examples . . .” (Davis & North, 1971, pp. 6–7). While these institutional factors are well-defined and conceptually discrete, we have a poor understanding of the relative importance of any one of these factors controlling for others, as well as the linkages across each factor. As described above, much work has already been done on individual factors of the institutional environment in isolation from one another; however, we believe it is important to begin to treat the institutional environment as the multi-faceted concept it is. Research for example, could explore the question of which elements of the institutional environment can be altered, holding others constant, and still yield substantive changes in multinational strategies in that country. Alternatively, a broader approach could be taken to begin to identify the extent to which the impact of political and legal systems on strategy and performance are themselves determined by the colonial heritage of the host country (Acemoglu, Johnson & Robinson, 2000). Similarly, there is the question of whether the strategic and performance implications of culture or informal norms are determined by the structure of the political and legal system (Ades & Tella, 1997, 1999; Broadman & Recanatini, 2000; Collier, 1999; Friedman et al., 1999; Goel & Nelson, 1998; Johnston, 1999; Larrain & Tavares, 1999; Tanzi, 1998; Treisman, 1999, 2000).

Further, if one factor or a set of factors emerges as being relatively more important than others for multinational entry strategies, what are the most important components of variation in that factor?

- If democracy influences strategy, what is the relative importance of electoral rules vs. checks and balances vs. the stability of the political institutions or party preferences in that democracy vs. the level of satisfaction of the citizens of that democracy with its institutions and choice of parties?
- If checks and balances are implicated, what is the relative importance of the existence of various veto points, the homogeneity or heterogeneity of their preferences and the information that they possess (Holburn & Vanden Bergh, 2002), the process by which they interact (agenda setting and last-mover powers) and the means by which actors arrive at membership in these veto points?
- What elements of French or Spanish legal systems pose the greatest difficulty for multinationals from alternate legal systems?

- What elements of labor market structures may be altered across firms within a country or across countries within a firm and which are fixed at one level or another or both?
- What do we mean by “culture” and where does it come from?

Addressing these questions will provide researchers, managers and policy makers with substantial guidance as to what dimensions of the institutional environment warrant their attention. If, as seems likely, the relatively easily obtained information on the existence of well-structured electoral rules and checks and balances must be supplemented by more subjective measures of political sentiment and political process, quantitative studies using available or readily obtainable data will have to be supplemented by detailed qualitative research examining specific institutional contexts in more detail (Spicer & Pyle, 2002). Such efforts in the academic, strategic or policy realms will require the composition of effective interdisciplinary teams that are able to both combine existing knowledge and then provide a conduit by which this new information can feed back into their respective areas of functional expertise.

Aside from examining jointly multiple facets of the institutional environment, future research should seek to do more than demonstrate that a measure of the institutional environment has an impact on multinational strategy. Rather, it should set forth mechanisms by which certain components of the institutional environment, and not others, influence a firm's strategic choices. Such evidence would not only deepen our understanding of the role of these forces on multinational strategy but it would also reassure skeptics concerned that any country-specific effect is not merely a proxy for unobserved country heterogeneity. Country-case studies or panel designs in which one or several facets of the institutional environment are observed to vary while other country-specific factors are held constant would serve similar purposes.

Sources of Knowledge

While conventional wisdom in the literature on multinationals suggests firms do learn about institutional environments in ways that influence multinational entry strategies, the mechanisms by which that learning occurs have not been sufficiently explored in the international context.⁷ As outlined in our review of the internationalization literature, learning can occur in an experiential expansion process that involves successfully higher levels of commitment to a country. Yet, it might also occur via alternate mechanisms such as by contracting for feasibility studies or by senior management's personal ties to a host country (Virany, Tushman & Romanelli, 1992).

A firm can also look outside its organizational boundaries for information. When formulating and implementing strategy, managers act with the knowledge that their competitors are facing similar choices. In international strategy, each individual actor faces uncertainty regarding the magnitude of hazards and opportunities in a potential host country. Yet the action of an initial firm or set of firms to choose a given country, entry mode, or investment level signals that at least one competitor believes that the opportunities outweigh the hazards. Given that a successful early entrant can use their incumbent status to advantage, competitors may follow rapidly. Such rational bandwagon effects (Abrahamson & Rosenkopf, 1993) have been observed in the behavior of such diverse populations as coal miners (Conell & Cohn, 1995), radio stations choosing market position (Greve, 1996), corporate boards contemplating adopting the M-form of governance (Fligstein, 1985; Palmer et al., 1987) or the form of a merger (Amburgey & Miner, 1992), countries deciding whether to nationalize foreign oil companies (Kobrin, 1985) or to seek independence from an Empire (Strang, 1991), and cities adopting civil service reform (Tolbert & Zucker, 1983).

The influence a firm's referent groups' choices exert on that firm's strategy has been extended to the international literature. Several recent studies point to a strong imitative pull on a firm's market entry and mode choices by its competitors and its business group members. Guillén (2002) has established that South Korean firms entered China at greater rates when other business group members had previously set up operations in China, or competitors had entered China. Related to the idea about firm heterogeneity in actions, mimetic responses were greatest among South Korean firms with minimal experience in China. Henisz and Delios (2001) produce an analogous finding but for the international investment choices of Japanese firms worldwide. The expansion of the market context to more than one host country permits Henisz and Delios (2001) to show that firm-level heterogeneity along dimensions of a firm's experience, both in the host country and in other international markets, and market-level heterogeneity, along the dimension of political hazards jointly influence market entry preferences and a firm's predilection to engage in mimetic behavior. Importantly, unlike a firm's level of uncertainty about a market, policy uncertainty cannot be bridged by market imitation. Lu (2002) represents a further extension to this work by showing how a firm's reliance on different forms of imitation – trait-based, frequency-based and outcome-based – are conditional on a firm's experience levels in a market. The more social (less technical) an imitation mode, the less an experienced firm relies on that imitation mode for making decisions about entering by a joint venture or by a wholly owned subsidiary.

These three studies are each firmly grounded in neoinstitutional theory, which identifies the important cues for a firm's action that a firm's reference

groups provide for strategic actions. The merging of neoinstitutional theory and the literature on national institutional environments found in Henisz and Delios (2001) also brings to attention differences in the definition of institutions across literatures. Neoinstitutional theory in the organizational domain typically concerns regulative, normative and cognitive issues. The preceding paragraphs, and much of our subsequent discussion, concerns the cognitive domain, in which the actions of other organizations provide mimetic pressures which can lead to a conformity in decision making and organizational forms. The regulative domain is captured by our discussion of the institutional environment. The normative domain can emerge in studies of multinational firms as well, as one explores issues such as the pressures and norms within the home country industry environment to compete internationally. Normative pressures can also emerge in the host environment, as local interest groups and local norms exert pressures on a multinational to conform to local standards and practices. Responding effectively to such normative pressures builds social legitimacy and perhaps eases the constraints a multinational encounters when operating in a new environment (Dowell, Swaminathan & Wade, 2002).

Returning to the cognitive dimension and the issue of imitation and social cues for action, the weight or value attached to an information cue may vary depending on the source (Guillén, 2002; Henisz & Delios, 2001). Firms that are formally or informally tied can appeal to one another for information or jointly identify, incorporate and share information (Kogut, 1998; Powell, Koput & Smith-Doerr, 1996). Tighter linkages between firms may also allow for a more rapid transverse of information across organizational boundaries and the readier acceptance of information even if it involves new insights or challenges to a firm's prevailing assumptions about a market. The number of contacts across firms in a network may also increase the probability of information sharing and new knowledge creation (Ingram & Simons, 1999; Saxenian, 1991) as could networks that span important knowledge or sub-group boundaries (Burt, 1992; Liebeskind et al., 1996; Tushman, 1977). The possibility of knowledge creation is further enhanced if the magnitude of the overall network allows for specialization by its member entities and the development of a comparative advantage in some form of knowledge development that is then shared among member firms (Sakakibara & Serwin, 2000; Tan & Vertinsky, 1996).

Within the Japanese firm context, vertical keiretsu, or vertically-oriented production groups, are one example where specialization has led to cost-based advantages for member firms (Dyer, 1996). Membership in a vertical keiretsu can also provide benefits to newly internationalizing firms (Belderbos & Sleuwaegen, 1996). Horizontal keiretsu in Japan, chaebol in Korea, family-centered industrial groups in India and Chinese family business groups are other

network forms in which member firms are integrated by such mechanisms as cross-appointments of directors and executives, cross-shareholdings, and joint projects (Khanna & Rivkin, 2000), or by social ties between member firms' owners (Keister, 2000).

One consequence of the close relationships between member firms in a network is that it fosters good information flows between companies (Gerlach, 1992; Weinstein & Yafeh, 1995). These relationships have been characterized as networks of knowledge (Imai, 1987) or "pathways for the exchange of resources" (Stuart, 1999). In such a network, information about foreign markets, along with resources related to finance, technology or other fields, is also a pooled resource among member firms (Helou, 1991). In these networks, firms gain information and capabilities from one another through ongoing trading relationships and collaborative projects (Lincoln, Gerlach & Ahmadjian, 1998). Personnel exchanges are an important source of information flow. For example, the major bank in a horizontal keiretsu will send staff to assume senior management positions in member companies. Personnel in keiretsu also tend to be interchangeable within staff functions across group companies (Gerlach, 1992; Weinstein & Yafeh, 1995).

Another more general form of organization through which learning may occur is the relationship between an acquiring and acquired firm (Mitchell et al., 2000; Vermeulen & Barkema, 2000) or between the parent firm in a chain or franchise relationship (Winter & Szulanski, 2002). These relationships may be further strengthened to the extent that the knowledge stocks of the two firms are complementary (Baum, Calabrese & Silverman, 1999; Dussuage, Garrette & Mitchell, 1999).

Each of these mechanisms appears viable for the purpose of learning about the institutional environment in a host country but we raise the question about which mechanisms are relatively cost effective under what conditions? While researchers have begun to address this question in the domestic context, there has been relatively little comparable research in the international arena. Two sets of questions that could be examined in multinational and domestic contexts could yield substantive insights into the mechanisms of learning about institutional environments.

The first set of questions deals with the nexus of organizational experience, institutional environments, learning and knowledge transfer. Approaching these questions could provide new insight into long-standing debates about knowledge-based (Kogut & Zander, 1993) and transaction-cost based theories of the firm (Williamson, 1996) by highlighting the issue of the efficacy of knowledge transfer across organizational and national boundaries.

- What is the relative effectiveness with which organizational experience is transferred across organizational and/or national boundaries (Simonin, 1999; Szulanski, 1996, 2000)? On what does it depend (Levinthal & March, 1993)?
- When does location matter (Tyre & Hippel, 1997) such that learning is more readily transferred within national boundaries than across them (Irwin & Klenow, 1994)?
- When do the types of knowledge matter such that some forms of learning are more readily transferred under certain environmental conditions (Steensma & Lyles, 2000)?
- When does the mechanism of knowledge transfer matter (Hatch & Mowery, 1998; Inkpen & Dinur, 1998)?
- How fast does learning depreciate (Argote, Beckman & Eppele, 1990; Darr, Argote & Eppele, 1995) and does this rate differ across different mechanisms of learning (Ingram & Baum, 1997)?

Given recent advances in the experiential learning literature in studies that drew on archival data (Barkema, Bell & Pennings, 1996; Barkema & Vereulen, 1998; Barkema & Vermeulen, 1997; Delios & Henisz, 2000; Henisz & Delios, 2001), inquiries based on primary data collection (Steensma & Lyles, 2000), in which the mechanisms of experiential knowledge gain and cross-border transfer are explored, are likely to have a substantial impact on our understanding of these issues. This form of data would be valued particularly in examinations involving learning about the policy environment, where little is known about the acquisition, retention and transmission of this knowledge within multinational firms.

The second set of questions concerns the intersection between national institutional environments, the inter-organizational environment and knowledge development. Given the intense demands on data for constructing time-varying profiles of competitor and partner activity, inquiry for this set of questions must be based on an archival data analysis. The advantage of approaching questions, such as the ones below, is that debates concerned with whether social cues provide information valuable to a decision or legitimization to a decision (DiMaggio & Powell, 1983; Haunschild, 1993; Haunschild & Miner, 1997; Podolny, 1994) can be broached.

One way to approach this debate is to focus on the outcomes of decisions made in response to social cues. To the extent that following social cues provides new information valuable to a particular decision, the organizational performance of firms that follow an imitative strategy should be higher. Better organizational performance implies that by social imitation an inexperienced organization in a particular institutional environment actually did acquire

knowledge and learn how to respond to a given situation. An important contrast to show in establishing this evidence is to compare the performance-related benefits of social imitation among a set of firms that plausibly have a need for such learning, vs. those that do not. In effect, this narrows to a comparison of social imitation across firms experienced and inexperienced in an institutional environment. To the extent that there is not a difference in outcomes between experienced and inexperienced firms, explanations of social imitation as legitimization receive greater support.

Making this form of investigation also entails establishing appropriate collections of firms across which social imitation might occur. We outline two such collections – business groups and competitive groups.

- How effective is learning within network forms of organization (Argote, 1993; Baum & Ingram, 1998)? Are there differences among different forms of business groups (chaebol, keiretsu, grupos, etc.) (Miyashita & Russell, 1994)?
- How much can be inferred from the behavior of competitors given the differences between firms (Irwin & Klenow, 1994; Shaver, Mitchell & Yeung, 1997)?

Non-market Strategies

In addition to increasing the precision with which researchers operationalize the institutional environment and the mechanisms by which firms learn, it would also be fruitful to extend the scope of strategic activities in which a multinational firm may engage. The majority of research on multinationals has delimited the scope of strategic choices to those that fall in the market environment – which markets to enter, what entry modes to employ, how much to invest, what types of investment to undertake, whether to exit and via what mechanism to exit. Given the centrality of the institutional environment to questions of multinational strategy, it is surprising that international strategy researchers lag behind domestic counterparts in considering strategic choices in the non-market environment. Choices about non-market activities involve resolving issues about what sort of policy compromises a firm should be willing to accept (Mahon, 1983); the allocation of scarce firm resources to achieve desirable non-market outcomes; and issues about organization, whether a firm should use a local partner, a host or home country trade association, its embassy in the host country, or other governmental arms. Non-market strategies also include questions about which public actors should be targeted, for example,

international organizations, informal actors, and host or home country executive, legislative and administrative branches (Holburn & Vanden Bergh, 2002).

Several sets of findings have begun to emerge from this nascent literature. Many of these findings concern the determinants of non-market activity, and the success of such activities. The prevalence of non-market activity has been linked to organizational attributes such as firm size and profits (Dickie, 1984; Masters & Keim, 1986; Salamon & Sigfried, 1977), organizational slack and a physical presence in the nation's capitol (Lenway & Rehbein, 1991; Schuler, 1996) as well as relatively poor environmental conditions (slumps in domestic demand or surges in imports) (Schuler, 1996). Meanwhile, research has shown that the number and diversity (including geographic dispersion) of supporting coalitions influences the success of a non-market strategy (Esty & Caves, 1983; Rehbein & Lenway, 1994; Yoffie, 1988). Other determinants of success include internalization of the non-market activity, the share of exports relative to production and the size of the industry in which the activity is occurring (Rehbein & Lenway, 1994).

In pursuing non-market activities, firms must contend with several problems. These include free-riding⁸ (de Figueiredo & Tiller, 2001; Esty & Caves, 1983; Pittman, 1988; Salamon & Sigfried, 1977; Schuler, 1996), the complexity of the relevant strategies (de Figueiredo & de Figueiredo, 2002), the specificity of knowledge which may "leak" to other firms should management decide to deploy a trade association or external lobbyist (de Figueiredo & Tiller, 2001) and the specificity of plant and equipment which make market adjustment strategies more difficult relative to their non-market counterparts as these assets have substantially lower values in their next best use (Alt et al., 1999). Additional disaggregated studies of the determinants of the organization, incidence and success of non-market strategies are clearly needed to better understand how non-market strategic activities are pursued within a single country.

Compared to research on the domestic front, the international analysis of non-market activity has proceeded even more hesitantly (Hillman & Keim, 1995; Murtha & Lenway, 1994). Given the importance of the institutional environment and of relevant organizational learning for multinational strategic behavior, the rewards to well-designed, non-market activity should be substantial. Of course, cross-national differences in the form of lobbying and lack of publicly available archival data hinder analysis of these effects. Nevertheless, given the apparent importance of non-market activities, future empirical research should seek to examine the determinants of their incidence and organization.

In particular, the following empirical questions seem worthy of additional study.

- Do firms originating in or possessing experience in similar institutional environments possess an advantage when requesting advantageous policy changes from government officials?
- What is the relative effectiveness of various strategies employed by firms to appear more “local” thus avoiding unfavorable policy changes? Strategies that could be considered include hiring local managers, employees and lobbyists and funding local economic and social infrastructure and political campaigns.
- How does the effectiveness of these “local” strategies compare to strategies based on the use of external leverage in the form of the home country embassy, multilateral agencies and international financial institutions? What is the relative effectiveness of these leverage strategies?
- In the case of either “local” or “external leverage” strategies, which political actor or agency is the most effective target for lobbying or informational strategies (Holburn & Vanden Bergh, 2002) and how important are broader classes of informal actors or belief systems (Dowell, Swaminathan & Wade, 2002) in achieving non-market strategy objectives?

Answering these questions would allow academics, managers and policy makers to focus their scarce resources on those strategies of highest return to their respective goals. Should non-market capabilities be transferable across national boundaries, they likely reside in corporate or regional headquarters or staff members that are rotated across individual facilities. If such capabilities are purely domestic, firms entering new institutional environments must advance slowly as they develop the specific non-market knowledge and capabilities of most use in that national context. The distinction between local and external leverage strategies similarly speaks to the location of the relevant capabilities and the speed with which they can be applied to new markets. The relative weight of formal political actors and informal actors and belief systems speaks to the relevance of formal analysis such as employed in Holburn and Vanden Bergh (2002) and sociological perspectives such as employed in Dowell, Swaminathan and Wade (2002). Given the preliminary nature of our understanding of these complex phenomena, it seems likely that managers, academic and policy makers are each making costly “mistakes” (de Figueiredo & de Figueiredo, 2002).

In addition to the complexity of these problems, an additional explanation for the wide variety in strategies and outcomes that we observe in the non-market domain is consistent with the thrust of our essay. One should expect substantial variation within a population of firms in their ability to transfer non-market capabilities across national boundaries, the types of non-market

strategies that they employ, their ability to identify effective targets for their strategies and, thus, the ultimate effectiveness of these strategies. Evidence in support of such firm-level heterogeneity would further enrich our understanding of the importance of firm-level knowledge and capabilities in a new dynamic sector critical to firm performance.

CONCLUSION

A prominent influence on a multinational firm's strategy is the range of institutional environments it encounters in the host countries in which it has invested, or in which it is considering investment. We have made the point that cross-national variation in the institutional environment adds uncertainty to new foreign operations. Uncertainty raises the hurdle rate of return and thus discourages entry. For entries that do occur, multinational firms are more likely to enter slowly with readily redeployable assets and rely upon host country partners to assist with understanding relevant domestic policies, factor markets and cultural norms. Survival probabilities will be higher to the extent that the institutional environment remains relatively stable or shares common characteristics with the environment in the home country, which reduces the extent of uncertainty.

To these basic but essential relationships between the institutional environment and dimensions of multinational strategy, we add the concept of a firm's international experience. We have developed arguments that show why firms with extensive international experience in similar institutional environments should be advantaged relative to their less experienced peers. Compared to less-experienced competitors, experienced firms will enter more rapidly, choose more appropriate investment levels and technologies, and have better success managing their alliances, or be less dependent upon them. Ultimately, we expect experienced firms' subsidiaries will enjoy higher survival probabilities. We also believe that the effects of experience will be highest when firms are entering the most dissimilar, and most hazardous, institutional environments.

The core idea in this framework is that a firm's profile of capabilities, as delineated by its investment experience in foreign markets, identifies the extent to which it is exposed to, and can manage hazards extending from, the institutional environments in which it operates. This framework explicitly integrates research at the firm-level and country-level and it reinforces the notion that research in international strategy is inherently multi-level. We have described a framework that does not consider these multi-level effects as independent effects; rather, we show how variance in one level, affects predictions about another level's influence on multiple aspects of a firm's

international strategy. This basic approach could readily be adopted to the integration of other levels of research, including the level of the inter-organizational environment which we also posit has a role to play in influencing a firm's responses to differing institutional environments.

While we explicitly considered several aspects of a firm's strategy, undoubtedly, numerous additional decisions, performance and firm characteristics could be similarly analyzed. As part of this, we expanded on our three core concepts – the institutional environment, experiential learning and market entry strategy – to show the richness of the phenomena being examined. In this expansion, we stress that the institutional environment consists of many more aspects than just political and cultural institutions. The benefits of better definition and measurement of the richness of the institutional environment rest in an improved understanding of the features of, and mechanisms by which, the institutional environment influences a multinational's entry strategy. Complementing this is the idea that a multinational has multiple routes to acquire knowledge and capabilities in foreign markets. By establishing the comparative effectiveness of forms of knowledge acquisition in foreign markets, particularly against different elements of the institutional environment, we can better understand how multinationals can overcome the liabilities they have in foreign markets. As the third part of this expansion, we also suggest that research begin to explore alternative strategies for addressing infirmities in the institutional environment to discern how non-market activities can complement or supersede traditional actions, such as the design of the foreign subsidiary.

Even though we offer these expansions on the conceptualizations of our core concepts, we maintain that the essence of our argument remains the same. Managers in a multinational firm must work to align strategies with both the hazards and opportunities they face in a given institutional environment, as well as with the firm-specific knowledge and capabilities at their disposal. Undoubtedly, multinational managers perceive the institutional environment as a fundamental influence on market strategy. Rather than taking institutions as an immutable constraint when making decisions, a firm can cultivate and exploit its ability to successfully manage diverse institutional hazards in its host country environments. This in itself is a resource of potentially high strategic value. Our research is an attempt to better understand this resource, and how its effective management can contribute to better performance in multinational firms.

NOTES

1. For a similar critique see Arias and Guillén (1998).

2. This hypothesis is strongly supported by the only available empirical study (Bradley, 1977). The author finds that expropriation of joint ventures exclusively between foreign multinationals is eight times as likely as expropriation of joint ventures that involve local partners.
3. Henisz (2000) demonstrates that this relationship is conditional upon the nature of the assets under the purview of the overseas subsidiary.
4. The importance of similarity or comparability for learning has also been explored in a domestic context. Halebian and Finkelstein (1999) find that experience in performing acquisitions that are “similar” improves the performance of subsequent acquisitions. Dussauge, Garrett and Mitchell (1999) find that alliance partners with greater overlap in their capabilities outperform their counterparts with respect to learning.
5. Ingram and Simons (2000) demonstrate a similarly complex and related relationship between political institutions, networks, ideology and survival among Israeli worker cooperatives over seven decades.
6. For further elaboration upon the conceptualization and measurement of cross-national differences in institutions see Guillén and Suarez (2001).
7. For notable exceptions within the marketing literature see Ganesh, Kumar and Subramaniam (1997) and Ganesh and Kumar (1996).
8. Damania and Frederikkson (2000) argue that the collusive potential of industry lobbying is higher in concentrated industries. At the aggregate level, this effect dominates the free rider effect. Zardkhooi (1985) and Munger (1988) find no relationship between industry concentration and the magnitude of PAC (political action committee) contributions.

REFERENCES

- Abrahamson, E., & Rosenkopf, L. (1993). Institutional and Competitive Bandwagons: Using mathematical modeling as a tool to explore innovation diffusion. *Academy of Management Review*, 18, 487–517.
- Acemoglu, D., Johnson, S., & Robinson, J. A. (2000). The Colonial Origins of Comparative Development: An empirical investigation. *Mimeo*.
- Ades, A., & Tella, R. D. (1997). National Champions and Corruption: Some unpleasant interventionist arithmetic. *Economic Journal*, 107, 1023–1042.
- Ades, A., & Tella, R. D. (1999). Rents, Competition and Corruption. *American Economic Review*, 89, 982–993.
- Agarwal, S., & Ramaswami, S. N. (1992). Choice of Foreign Market Entry Mode: Impact of ownership, location and internalization factors. *Journal of International Business Studies*, 23, 1–27.
- Aharoni, Y. (1966). *The Foreign Investment Decision Process*. Graduate School of Business, Harvard University.
- Alt, J. E., Carlsen, F., Huem, P., & Johansen, K. (1999). Asset Specificity and the Political Behavior of Firms: Lobbying for subsidies in Norway. *International Organization*, 53, 99–116.
- Amburgey, T. L., & Miner, A. S. (1992). Strategic Momentum: The effects of repetitive, positional and contextual momentum on merger activity. *Strategic Management Journal*, 13, 335–348.
- Anand, B., & Khanna, T. (2000). Do Firms Learn to Create Value? The case of alliances. *Strategic Management Journal*, 21, 295–315.

- Argote, L. (1993). Group and Organizational Learning Curves: Individual, system and environmental components. *British Journal of Psychology*, 32, 31–51.
- Argote, L., Beckman, S. L., & Eppele, D. (1990). The Persistence and Transfer of Learning in Industrial Settings. *Management Science*, 36, 140–154.
- Arias, M. E., & Guillén, M. F. (1998). The Transfer of Organizational Management Techniques Across Borders. In: J. L. Alvarez (Ed.), *The Diffusion and Consumption of Business Knowledge* (pp. 110–137). London: Macmillan.
- Barkema, H. G., Bell, J. H. J., & Pennings, J. M. (1996). Foreign entry, cultural barriers, and learning. *Strategic Management Journal*, 17, 151–166.
- Barkema, H. G., & Drogendijk, R. (2002). A New Internationalization Process Model: Theory and Evidence. *Mimeo*.
- Barkema, H. G., & Vereulen, F. (1998). International Expansion through Start-up or Acquisition: A learning perspective. *Academy of Management Journal*, 41, 7–27.
- Barkema, H. G., & Vermeulen, F. (1997). What Differences in the Cultural Backgrounds of Partners are Detrimental for International Joint Ventures. *Journal of International Business Studies*, 28, 845–864.
- Baum, J. A. C., Calabrese, T., & Silverman, B. S. (1999). Don't Go it Alone: Alliance network composition and startups' performance in Canadian biotechnology. *Strategic Management Journal*, 21, 267–294.
- Baum, J. A. C., & Ingram, P. (1998). Survival-enhancing Learning in the Manhattan Hotel Industry. *Management Science*, 44, 996–1016.
- Beamish, P. W. (1988). *Multinational Joint Ventures in Developing Countries*. New York: Routledge.
- Belderbos, R., & Sleuwaegen, L. (1996). Japanese Firms and the Decision to Invest Abroad: Business Groups and Regional Core Networks. *Review of Economics and Statistics*, 78, 214–220.
- Benito, G. R. G., & Gripsrud, G. (1992). The Expansion of Foreign Direct Investments: Discrete rational location choices or a cultural learning process? *Journal of International Business Studies*, 23, 461–476.
- Bennet, P. D., & Green, R. T. (1972). Political Instability as a Determinant of Direct Foreign Investment in Marketing. *Journal of Marketing Research*, 19, 182–186.
- Bergara Duque, M. E., Henisz, W. J., & Spiller, P. T. (1998). Political Institutions and Electric Utility Investment: A cross-nation Analysis. *California Management Review*, 40, 18–35.
- Biggart, N. W., & Guillén, M. F. (1999). Developing Difference: Social Organization and the Rise of the Auto Industries of South Korea, Taiwan, Spain and Argentina. *American Sociological Review*, 64, 722–747.
- Bradley, D. (1977). Managing Against Expropriation. *Harvard Business Review*, 75–83.
- Broadman, H. G., & Recanatini, F. (2000). Seeds of Corruption: Do market institutions matter? *Mimeo of the World Bank, European and Central Asia Regional Operations Poverty Reduction and Economic Management Department*.
- Brown, J. S., & Digid, P. (1991). Organizational Learning and Communities of Practice: Toward a unified view of working, learning and organization. *Organization Science*, 2, 40–57.
- Burke, S. P., & Casson, M. (1998). Information Strategies in Foreign Market Entry: Investing in sales and distribution facilities. *Mimeo*.
- Burt, R. (1992). *Structural Holes: The Social Structure of Competition*. Cambridge: Harvard University Press.
- Caballero, R. J., & Hammond, M. L. (1998). The Macroeconomics of Specificity. *Journal of Political Economy*, 106, 724–768.

- Carroll, G. R., & Delacroix, J. (1982). Organizational Mortality in the Newspapers Industries of Argentina and Ireland. *Administrative Science Quarterly*, 27, 169–198.
- Chang, S. J. (1995). International expansion strategy of Japanese firms: Capability building through sequential entry. *Academy of Management Journal*, 38, 383–407.
- Chang, S. J., & Rosenzweig, P. M. (2001). The Choice of Entry Mode in Sequential Foreign Direct Investment. *Strategic Management Journal*, 22, 747–776.
- Chung, W., & Alcacer, J. (2002). Heterogeneous Investment Motives and Location Choice of Foreign Direct Investment in the United States. *Mimeo*.
- Collier, M. W. (1999). Explaining Political Corruption: An institutional-choice approach. *Mimeo*.
- Conell, C., & Cohn, S. (1995). Learning from Other People's Actions: Environmental variables and diffusion in French coal mining strikes: 1890–1935. *American Journal of Sociology*, 101, 366–403.
- Crain, W. M., & Oakley, L. K. (1995). The Politics of Infrastructure. *Journal of Law and Economics*, 38, 1–17.
- Dailami, M., & Leipziger, D. (1998). Infrastructure Project Finance and Capital Flows: A new perspective. *World Development*, 26, 1283–1298.
- Damania, R., & Frederiksson, P. G. (2000). On the Formation of Industry Lobby Groups. *Journal of Economic Behavior and Organization*, 41, 315–335.
- Daniels, R., & Trebilcock, M. J. (1994). Private Provision of Public Infrastructure: The next privatization frontier? Paper presented at IDS-270.
- Darr, E. D., Argote, L., & Eppler, D. (1995). The Acquisition, Transfer and Depreciation of Knowledge in Service Organizations: Productivity in franchises. *Management Science*, 41, 1750–1762.
- Davidson, W. (1980). The Location of Foreign Direct Investment Activity: Country Characteristics and Experience Effects. *Journal of International Business Studies*, 11, 9–22.
- Davidson, W. H., & McFetridge, D. G. (1985). Key Characteristics in the Choice of International Technology Transfer Mode. *Journal of International Business Studies*, 16, 5–21.
- Davis, L. E., & North, D. C. (1971). *Institutional Change and American Economic Growth*. Cambridge: Cambridge University Press.
- de Figueiredo, J., & de Figueiredo, R. J. (2002). Managerial Decision Making in Non-market Strategy: An Experiment. *Advances in Strategic Management*, 19.
- de Figueiredo, J. M., & Tiller, E. H. (2001). Organizing and Engaging in Non-market Activity: A transactional analysis of corporate lobbying at the Federal Communication Commission. *Journal of Economics and Management Strategy*, 10, 91–122.
- Delios, A., & Beamish, P. (1999). Ownership Strategy of Japanese Firms: Transactional, institutional and experience influences. *Strategic Management Journal*, 20, 711–727.
- Delios, A., & Beamish, P. W. (2001). Survival and Profitability: The roles of experience and intangible assets in foreign subsidiary performance. *Academy of Management Journal*, 44 (forthcoming).
- Delios, A., & Henisz, W. J. (2000). Japanese Firms' Investment Strategies in Emerging Economies. *Academy of Management Journal*, 43, 305–323.
- Delios, A., & Henisz, W. J. (2002). Political Hazards, Experience and Sequential Investment Strategies: The International Expansion of Japanese Firms: 1980–1998. *Mimeo*.
- Dickie, R. (1984). Influence of Public Affairs Office on Corporate Planning and of Corporations on Government Policy. *Strategic Management Journal*, 5, 15–34.
- DiMaggio, P. J., & Powell, W. W. (1983). The Iron Cage Revisited: Institutional Isomorphism and Collective Rationality in Organizational Fields. *American Sociological Review*, 48, 147–160.

- Djankov, S., Porta, R. L., Silanes, F. L. d., & Shleifer, A. (2000). The Regulation of Entry. *NBER Working Paper*, 7892.
- Dowell, G., Swaminathan, A., & Wade, J. (2002). Pretty Pictures and Ugly Scenes: Political and Technological Maneuvers in High Definition Television. *Advances in Strategic Management*, 19.
- Dussauge, P., Garrette, B., & Mitchell, W. (1999). Learning From Competing Partners: Outcomes and duration of scale and link alliances in Europe, North America and Spain. *Strategic Management Journal*, 21, 99–126.
- Dyer, J. H. (1996). Specialized Supplier Networks as a Source of Competitive Advantage: Evidence from the auto industry. *Strategic Management Journal*, 17, 271–291.
- Eriksson, K., Johanson, J., Majkgard, A., & Sharma, D. D. (1997). Experiential Knowledge and Cost in the Internationalization Process. *Journal of International Business Studies*, 28, 337–360.
- Erramilli, M. K. (1991). The Experience Factor in Foreign Market Entry Behavior of Service Firms. *Journal of International Business Studies*, 22, 479–501.
- Erramilli, M. K. (1996). Nationality and Subsidiary Ownership Patterns in Multinational Corporations. *Journal of International Business Studies*, 27, 225–248.
- Esty, C., & Caves, R. E. (1983). Market Structure and Political Influence: New data on political expenditures, activity and success. *Economic Inquiry*, 21, 24–38.
- Fligstein, N. (1985). The Spread of the Multidivisional Form Among Large Firms: 1919–1979. *American Sociological Review*, 50, 377–391.
- Friedman, E., Johnson, S., Kaufmann, D., & Zoido-Lobaton, P. (1999). Dodging the Grabbing Hand: The determinants of unofficial activity in 69 countries. *Journal of Public Economics*, 76, 459–493.
- Ganesh, J., & Kumar, V. (1996). Capturing the Cross-national Learning Effect: An analysis of industrial technology diffusion. *Journal of the Academy of Marketing Science*, 24, 328–337.
- Ganesh, J., Kumar, V., & Subramaniam, V. (1997). Learning Effect in Multinational Diffusion of Consumer Durables: An exploratory investigation. *Journal of the Academy of Marketing Science*, 25, 214–228.
- Gastanaga, V. M., Nugent, J. B., & Pashamova, B. (1998). Host Country Reforms and FDI Inflows: How much difference do they make? *World Development*, 26, 1299–1314.
- Gatignon, H., & Anderson, E. (1988). The Multinational Corporation's Degree of Control over Foreign Subsidiaries: An empirical test of a transaction cost explanation. *Journal of Law, Economics and Organization*, 4, 305–336.
- Gerlach, M. L. (1992). *Alliance Capitalism: The Social Organization of Japanese Business*, 351. Berkeley, CA: University of California Press.
- Goel, R. K., & Nelson, M. A. (1998). Corruption and Government Size: A disaggregated analysis. *Public Choice*, 97, 107–120.
- Goodnow, J. D., & Hansz, J. E. (1972). Environmental Determinants of Overseas Market Entry Strategies. *Journal of International Business Studies*, 3, 33–50.
- Grandy, C. (1989). Can the Government Be Trusted to Keep its Part of a Social Contract? New Jersey and the railways: 1825–1888. *Journal of Law, Economics and Organization*, 5, 249–269.
- Green, R. T., & Cunningham, W. (1975). The Determinants of U.S. Foreign Investment: An empirical examination. *Management International Review*, 15, 113–120.
- Greve, H. R. (1996). Patterns of Competition: The diffusion of a market position in radio broadcasting. *Administrative Science Quarterly*, 41, 29–60.
- Guillén, M. (1994). *Models of Management: Work authority and organization in comparative perspective*. Chicago: University of Chicago Press.

- Guillén, M. (1999). Organizational Learning and the Choice Between Wholly-Owned and Joint-Venture Foreign Manufacturing: South Korean firms and business groups in China. *Mimeo*.
- Guillén, M. (2001). *The Limits of Convergence: Globalization and Organizational Change in Argentina, South Korea, and Spain*. Princeton, NJ: Princeton University Press.
- Guillén, M. (2002). Structural Inertia, Imitation and Foreign Expansion: South Korean firms and business groups in China: 1987–1995. *Academy of Management Journal* (forthcoming).
- Guillén, M. F., & Suarez, S. L. (2002). The Institutional Context of Multinational Activity. In: S. Ghoshal & E. Westney (Eds), *Organization Theory and the Multinational Corporation*. New York: St. Martin's Press.
- Haleblian, J., & Finkelstein, S. (1999). The Influence of Organization Acquisition Experience on Acquisition Performance: A behavioral learning theory perspective. *Administrative Science Quarterly*, 44, 29–56.
- Hamilton, G. G., & Biggart, N. W. (1988). Market, Culture and Authority: A Comparative Analysis of Management and Organization in the Far East. *American Journal of Sociology*, 94, 52–94.
- Hanson, J. R. I. (1999). Culture Shock and Direct Investment in Poor Countries. *Journal of Economic History*, 59, 1–16.
- Hatch, N. W., & Mowery, D. C. (1998). Process Innovation and Learning by Doing in Semiconductor Manufacturing. *Management Science*, 44, 1461–1477.
- Haunschild, P. R. (1993). Interorganizational Imitation: The impact of interlocks on corporate acquisition activity. *Administrative Science Quarterly*, 38, 564–592.
- Haunschild, P. R., & Miner, A. S. (1997). Models of Interorganizational Imitation: The effects of outcome salience and uncertainty. *Administrative Science Quarterly*, 42, 472–500.
- Helou, A. (1991). The Nature and Competitiveness of Japan's Keiretsu. *Journal of World Trade*, 25, 99–131.
- Henisz, W. J. (2000). The Institutional Environment for Multinational Investment. *Journal of Law, Economics and Organization*, 16, 334–364.
- Henisz, W. J., & Delios, A. (2001). Uncertainty, Imitation, And Plant Location: Japanese Multinational Corporations: 1990–1996. *Administrative Science Quarterly*, 46, 443–475.
- Henisz, W. J., & Zelner, B. A. (2001). The Institutional Environment for Telecommunications Investment. *Journal of Economics & Management Strategy*, 10, 123–147.
- Henisz, W. J., & Zelner, B. A. (2002). Interest Groups, Political Institutions and Electricity Investment. *Mimeo*, 2001.
- Hennart, J.-F. (1988). A Transaction Cost Theory of Joint Ventures. *Strategic Management Journal*, 9, 361–374.
- Hill, C. W. L., Hwang, P., & Kim, W. C. (1990). An Eclectic Theory of the Choice of International Entry Mode. *Strategic Management Journal*, 11, 117–128.
- Hillman, A. J., & Keim, G. (1995). International Variation in the Business-Government Interface: Institutional and Organizational Considerations. *Academy of Management Review*, 20, 193–214.
- Holburn, G., & Vanden Bergh, R. (2002). Policy and Process: A Framework for the Design of Non-Market Strategy. *Advances in Strategic Management*, 19.
- Huber, G. (1991). Organizational Learning: The contributing processes and literatures. *Organization Science*, 2, 88–115.
- Hyer, S. (1976). *The International Operations of National Firms*. Cambridge, MA: MIT Press.
- Imai, K. (1987). The Corporate Network in Japan. *Japanese Economic Studies*, 16, 1–37.

- Ingram, P., & Baum, J. (1997). Opportunity and Constraint: Organizations' learning from the operating and competitive experience of industries. *Strategic Management Journal*, 18, 75–98.
- Ingram, P., & Simons, T. (1999). The Exchange of Experience in a Moral Economy: Thick ties and vicarious learning in kibbutz agriculture. *Working Paper*.
- Ingram, P., & Simons, T. (2000). State Formation, Ideological Competition, and the Ecology of Israeli Workers' Cooperatives, 1920–1992. *Administrative Science Quarterly*, 45, 25–53.
- Inkpen, A. C., & Beamish, P. W. (1997). Knowledge, Bargaining Power and the Instability of International Joint Ventures. *Academy of Management Review*, 22, 177–202.
- Inkpen, A. C., & Dinur, A. (1998). Knowledge Management Processes and International Joint Ventures. *Organization Science*, 9, 454–468.
- Irwin, D. A., & Klenow, P. J. (1994). Learning-by-Doing Spillovers in the Semiconductor Industry. *Journal of Political Economy*, 102, 1200–1227.
- Johanson, J., & Vahlne, J.-E. (1977). The Internationalization Process of the Firm – A Model of Knowledge Development and Increasing Foreign Market Commitments. *Journal of International Business Studies*, 8, 23–32.
- Johanson, J., & Wiedersheim-Paul, F. (1975). The Internationalization Process of the Firm – Four Swedish cases. *Journal of Management Studies*, 12, 305–322.
- Johnston, M. (1999). Party Systems, Competition and Political Checks against Corruption. Prepared for a Conference on “Political Corruption and Parties” at the Robert Schuman Center of the European University Institute in Florence, Italy.
- Keister, L. A. (2000). *Chinese Business Groups*. New York: Oxford University Press.
- Khanna, T., & Rivkin, J. W. (2000). Ties that Bind Business Groups: Evidence from an emerging economy. *Harvard Business School Division of Research Working Paper*.
- Kindleberger, C. P. (1969). *American Business Abroad*. New Haven, CT: Yale University Press.
- Kobrin, S. (1979). Political Risk: A review and reconsideration. *Journal of International Business Studies*, 10, 67–80.
- Kobrin, S. J. (1985). Diffusion as an Explanation of Oil Nationalization (Or the domino effect rides again). *Journal of Conflict Resolution*, 29, 3–32.
- Kogut, B. (1983). Foreign direct investment as a sequential process. In: C. P. Kindleberger (Ed.), *The Multinational Corporations In the 1980s* (pp. 35–56). Cambridge: MIT Press.
- Kogut, B. (1991). Joint Ventures and the Option to Expand and Acquire. *Management Science*, 37, 19–33.
- Kogut, B. (1998). The Network as Knowledge. *Strategic Management Journal*, 21, 405–425.
- Kogut, B., & Kulatilaka, N. (1994). Operating Flexibility, Global Manufacturing and the Option Value of a Multinational Network. *Management Science*, 40, 123–139.
- Kogut, B., & Singh, H. (1988). The Effect of National Culture on the Choice of Entry Mode. *Journal of International Business Studies*, 19, 411–432.
- Kogut, B., & Zander, U. (1993). Knowledge of the Firm and the Evolutionary Theory of the Multinational Corporation. *Journal of International Business Studies*, 24, 625–645.
- La Porta, R., Lopez-De-Silanes, F., Shleifer, A., & Vishny, R. W. (1998). Law and Finance. *Journal of Political Economy*, 106, 1113–1155.
- La Porta, R., Lopez-De-Silanes, F., Shleifer, A., & Vishny, R. W. (1999). Investor Protection: Origins, Consequences, Reform. *NBER Working Paper*, 7248.
- Larrain, F. B., & Tavares, J. (1999). Can Openness Deter Corruption? *Mimeo*.
- Lenway, S. A., & Rehbein, K. (1991). Leaders, Followers and Free Riders: An empirical test of variation in corporate political involvement. *Academy of Management Journal*, 34, 893–905.
- Levinthal, D. A., & March, J. G. (1993). The Myopia of Learning. *Strategic Management Journal*, 14, 95–112.

- Levy, B., & Spiller, P. T. (1994). The Institutional Foundations of Regulatory Commitment: A Comparative Analysis of Telecommunications Regulation. *Journal of Law, Economics and Organization*, 10, 201–246.
- Levy, B., & Spiller, P. T. (1996). *Regulations, Institutions and Commitment*. Cambridge: Cambridge University Press.
- Li, J. (1995). Foreign Entry and Survival: Effects of strategic choices on performance in international markets. *Strategic Management Journal*, 16, 333–351.
- Li, J., & Guisinger, S. (1991). Comparative Business Failures of Foreign-controlled Firms in the U.S. *Journal of International Business Studies*, 22, 209–224.
- Liebeskind, J. P., Oliver, A. L., Zucker, L., & Brewer, M. (1996). Social Networks, Learning and Flexibility: Sourcing scientific knowledge in new biotechnology firms. *Organization Science*, 7, 428–443.
- Lincoln, J., Gerlach, M., & Ahmadjian, C. (1998). Evolving Patterns of Keiretsu Organization and Action in Japan. In: B. M. Staw & L. L. Cummings (Eds), *Research in Organizational Behavior* (Vol. 20, pp. 307–343). Greenwood, NJ: JAI Press.
- Loree, D. W., & Guisinger, S. E. (1995). Policy and Non-policy Determinants of U.S. Equity Foreign Direct Investment. *Journal of International Business Studies*, 26, 281–299.
- Lorenzoni, G., & Lipparini, A. (1999). The Leveraging of Interfirm Relationships as a Distinctive Organizational Capability: A longitudinal study. *Strategic Management Journal*, 20, 317–338.
- Lu, J. W. (2002). Intra- and Inter-Organizational Imitative Behavior: Institutional Influences on Japanese Firms' Entry Mode Choice. *Journal of International Business Strategy*, 33 (in press).
- Mahon, J. (1983). Corporate Political Strategies: An empirical study of chemical firm responses to superfund legislation. In: L. Preston (Ed.), *Research in Corporate Social Performance and Policy* (Vol. 5, pp. 143–182). Greenwich, CT: JAI Press.
- Martin, X., Swaminathan, A., & Mitchell, W. (1998). Organizational Evolution in the Interorganizational Environment: Incentives and Constraints on International Expansion Strategy. *Administrative Science Quarterly*, 43, 566–601.
- Masters, M. S., & Keim, G. (1986). Variation in Corporate PAC and Lobbying Activity: An organizational and environmental analysis. In: J. E. Post (Ed.), *Research in Corporate Social Performance and Policy* (pp. 249–272). Greenwich, CT: JAI Press.
- Mitchell, W., Baum, J. A. C., Banaszak-Holl, J., Berta, W. B., & Bowman, D. (2000). Opportunity and Constraint: Chain-to-Component Transfer Learning in Multitunit Chains of U.S. Nursing Homes, 1991–1997. In: N. Bontis & C. W. Choo (Eds), *Strategic Management of Intellectual Capital and Organizational Knowledge*. New York: Oxford University Press.
- Mitchell, W., Shaver, J. M., & Yeung, B. (1992). Getting There in a Global Industry: Impacts on performance of changing international presence. *Strategic Management Journal*, 13, 419–432.
- Mitchell, W., Shaver, J. M., & Yeung, B. (1993). Performance Following Changes of International Presence in Domestic and Transition Industries. *Journal of International Business Studies*, 24, 647–669.
- Mitchell, W., & Singh, K. (1992). Incumbents Use of Pre-entry Alliances before Expansion into New Technical Subfields of an Industry. *Journal of Economic Behavior and Organization*, 18, 347–372.
- Miyashita, K., & Russell, D. W. (1994). *Keiretsu: Inside the hidden Japanese conglomerates*. New York: McGraw Hill, Inc.
- Mowery, D. C., Oxley, J. E., & Silverman, B. S. (1996). Strategic Alliances and Interfirm Knowledge Transfer. *Strategic Management Journal*, 17, 77–91.

- Munger, M. C. (1988). On the Political Participation of the Firm in the Electoral Process: An update. *Public Choice*, 56, 295–298.
- Murtha, T., & Lenway, S. A. (1994). Country Capabilities and the Strategic State: How national political institutions affect multinational corporations strategies. *Strategic Management Journal*, 15, 113–129.
- Mutinelli, M., & Piscitello, L. (1998). The Entry Mode Choice of MNEs: An evolutionary approach. *Research Policy*, 27, 391–506.
- Nelson, R. R., & Winter, S. G. (1982). *An Evolutionary Theory of Economic Change*. Cambridge: Belknap Press.
- Oxley, J. E. (1999). Institutional Environment and the Mechanisms of Governance: The impact of intellectual property protection on the structure of inter-firm alliances. *Journal of Economic Behavior and Organization*, 38, 283–310.
- Palmer, D., Friedland, R., Hennings, P. D., & Powers, M. E. (1987). The Economics and Politics of Structure: The multi-divisional form and the large U.S. corporation. *Administrative Science Quarterly*, 32, 25–48.
- Pedersen, T., & Shaver, J. M. (2000). Internationalization Revisited: The ‘big step’ hypothesis. *Mimeo*.
- Pennings, J. M., Barkema, H., & Douma, S. (1994). Organizational Learning and Diversification. *Academy of Management Journal*, 37, 608–640.
- Pittman, R. (1988). Rent-seeking and Market Structure: A comment. *Public Choice*, 58, 173–185.
- Podolny, J. M. (1994). Market uncertainty and the social character of economic exchange. *Administrative Science Quarterly*, 39, 458–483.
- Powell, W. W., Koput, K. W., & Smith-Doerr, L. (1996). Interorganizational Collaboration and the Locus of Innovation: Networks of learning in biotechnology. *Administrative Science Quarterly*, 41, 116–145.
- Ramamurti, R. (Ed.) (1996). *Privatizing Monopolies: Lessons from the telecommunications and transport sectors in Latin America*. Baltimore: Johns Hopkins University Press.
- Rehbein, K., & Lenway, S. (1994). Determining an Industry’s Political Effectiveness With the U.S. International Trade Commission. *Business & Society*, 33, 270–292.
- Rhee, J. H., & Cheng, J. L. C. (2000). Re-examining the Incremental Model of International Expansion: The impact of firm, industry and host-country factors. University of Illinois at Urbana-Champaign College of Commerce and Business Administration CIBER Working Paper.
- Root, F. R., & Ahmed, A. A. (1978). The Influence of Policy Instruments on Manufacturing Direct Foreign Investment in Developing Countries. *Journal of International Business Studies*, 9, 81–93.
- Sakakibara, M., & Serwin, K. (2000). U.S. Distribution Strategy of Japanese Manufacturing Firms: The role of Keiretsu. *Journal of the Japanese and International Economies*, 14, 43–72.
- Salamon, L. M., & Sigfried, J. J. (1977). Economic Power and Political Influence: The impact of industry structure on public policy. *American Political Science Review*, 71, 1026–1043.
- Savedoff, W., & Spiller, P. (1997). Commitment and Governance in Infrastructure Sectors: Manuscript prepared for an IDB Conference on *Private Investment, Infrastructure Reform and Governance in Latin America and the Caribbean* (September 15–16).
- Saxenian, A. L. (1991). The Origins and Dynamics of Production Networks in Silicon Valley. *Research Policy*, 20, 423–437.
- Schuler, D. A. (1996). Corporate Political Strategy and Foreign Competition: The case of the steel industry. *Academy of Management Journal*, 39, 720–737.

- Shaver, J. M. (1998). Accounting for Endogeneity When Assessing Strategy Performance: Does entry mode choice affect FDI survival? *Management Science*, 44, 571–585.
- Shaver, J. M., Mitchell, W., & Yeung, B. (1997). The Effect of Own-Firm and Other-Firm Experience on Foreign Direct Investment Survival in the United States: 1987–1992. *Strategic Management Journal*, 18, 811–824.
- Simon, H. (1945). *Administrative Behavior*. New York: The Free Press.
- Simonin, B. L. (1997). The Importance of Collaborative Know-How: An empirical test of the learning organization. *Academy of Management Journal*, 40, 1150–1174.
- Simonin, B. L. (1999). Ambiguity and the Process of Knowledge Transfer in Strategic Alliances. *Strategic Management Journal*, 20, 595–623.
- Smarzynska, B. K., & Wei, S.-J. (2000). Corruption and Composition of Foreign Direct Investment: Firm-level evidence. *Mimeo, The World Bank*.
- Spicer, A., & Pyle, W. (2002). Institutions and the Vicious Circle of Distrust in the Russian Market for Household Deposits: 1992–1999. *Advances in Strategic Management*, 19.
- Spiller, P. T., & Vogelsang, I. (1996). The Institutional Foundations of Regulatory Commitment in the U.K. (with special emphasis on telecommunication). In: B. Levy & P. Spiller (Eds), *Regulations, Institutions and Commitment* (pp. 121–144). Cambridge University Press.
- Steensma, H. K., & Lyles, M. A. (2000). Explaining IJV Survival in a Transitional Economy Through Social Exchange and Knowledge-Based Perspectives. *Strategic Management Journal*, 21, 831–851.
- Stobaugh, R. (1969). How to analyze foreign investment climates. *Harvard Business Review*, 100–108.
- Strang, D. (1991). Adding Social Structure to Diffusion Models: An event history framework. *Sociological Methods and Research*, 19, 324–353.
- Stuart, T. E. (1999). Interorganizational Alliances and the Performance of Firms: A study of growth and innovation rates in a high-technology industry. *Strategic Management Journal*, 21, 791–811.
- Sullivan, D., & Bauerschmidt, A. (1990). Incremental Internalization: A test of Johanson and Vahlne's thesis. *Management International Review*, 30, 19–30.
- Szulanski, G. (1996). Exploring Internal Stickiness: Impediments to the transfer of best practice within the firm. *Strategic Management Journal*, 17, 27–43.
- Szulanski, G. (2000). When Credible Sources Share Complex Knowledge. *Mimeo*.
- Tan, B., & Vertinsky, I. (1996). Foreign Direct Investment by Japanese Electronics Firms in the United States and Canada: Modeling the timing of entry. *Journal of International Business Studies*, 27, 655–681.
- Tanzi, V. (1998). Corruption Around the World: Causes, Consequences, Scope and Cures. *IMF Staff Papers*, 45, 559–594.
- Teece, D. J. (1977). Technology Transfer by Multinational Firms: The resource cost of transferring technological know-how. *The Economic Journal*, 87, 242–261.
- Tolbert, P. S., & Zucker, L. G. (1983). Institutional Sources of Change in the Formal Structure of Organizations: The diffusion of civil service reform: 1880–1935. *Administrative Science Quarterly*, 28, 22–39.
- Treisman, D. (1999). Decentralization and Corruption: Why are federal states perceived to be more corrupt? *Mimeo*.
- Treisman, D. (2000). The Causes of Corruption: A cross-national study. *Journal of Public Economics*, 76, 399–457.
- Tushman, M. L. (1977). Special Boundary Roles in the Innovation Process. *Administrative Science Quarterly*, 22, 587–605.

- Tyre, M. J., & Hippel, E. v. (1997). The Situated Nature of Adaptive Learning in Organizations. *Organization Science*, 8, 71–83.
- Vermeulen, F., & Barkema, H. (2000). Learning Through Acquisitions. *Mimeo*.
- Vernon, R. (1977). The strain on national objectives: The developing countries, *Storm over the Multinationals: The Real Issues*, 139–173. Cambridge, MA: Harvard University.
- Virany, B., Tushman, M. L., & Romanelli, E. (1992). Executive Succession and Organization Outcomes in Turbulent Environments: An Organizational Learning Approach. *Organization Science*, 3, 72–91.
- Wei, S.-J. (2000). How Taxing is Corruption on International Investors. *Review of Economics and Statistics*, 82, 1–11.
- Weinstein, D. E., & Yafeh, Y. (1995). Japan's corporate groups: Collusive or competitive? An empirical investigation of keiretsu behavior. *Journal of Industrial Economics*, 43, 359–376.
- Wilkins, M. (1970). *The Emergence of Multinational Enterprise: American Business Abroad from the Colonial Era to 1914*. Cambridge, MA: Harvard University Press.
- Williamson, O. E. (1976). Franchise Bidding with Respect to CATV and in General. *Bell Journal of Economics*, 73–104.
- Williamson, O. E. (1996). *The Mechanisms of Governance*. New York: Oxford University Press.
- Winter, S., & Szulanski, G. (2002). Replication as Strategy. *Organization Science* (forthcoming).
- Yoffie, D. B. (1988). How an Industry Builds Political Advantage. *Harvard Business Review*, 82–89.
- Yu, C.-M. J. (1990). The Experience Effect and Foreign Direct Investment. *Weltwirtschaftliches Archiv*, 126, 561–580.
- Zaheer, S. (1995). Overcoming the Liability of Foreignness. *Academy of Management Journal*, 38, 341–363.
- Zaheer, S., & Mosakowski, E. (1997). The Dynamics of the Liability of Foreignness: A global study of survival in financial services. *Strategic Management Journal*, 18, 439–464.
- Zardkhooi, A. (1985). On the Political Participation By Firms in the Electoral Process. *Southern Economic Journal*, 58, 187–191.

INSTITUTIONS AND THE VICIOUS CIRCLE OF DISTRUST IN THE RUSSIAN HOUSEHOLD DEPOSIT MARKET, 1992–1999

Andrew Spicer and William Pyle

ABSTRACT

Can private firms produce the trust needed to develop and sustain new markets in the absence of government intervention? In our analysis of the Russian household deposit market during the 1990s, we show how the initial conditions of market emergence contributed to a vicious circle in which private commercial banks progressively lost the trust of potential depositors. The Russian case highlights the importance of institutions as collective goods whose successful construction are critical to market success.

INTRODUCTION

When the gains to trust production have merited it, firms have been shown to develop strategies for building their reputation, improving their image and signaling the goodness of their intentions (Klein, 2000). Moreover, researchers have demonstrated how private, third-party organizations often arise to generate market-supporting trust, with successful cases ranging from medieval law

merchants (Milgrom, North & Weingast, 1990) to modern credit rating agencies (Newman, 2000). The progressive evolution of trust, however, is not inevitable. Indeed, the recent case of the Russian deposit market provides us with an example of how unregulated market competition may lead to a self-reinforcing cycle of trust erosion and market destruction. Despite potential gains from trade, the Russian population's distrust of financial institutions led them to withdraw from the deposit market. And in the face of this withdrawal, commercial banks grew both more unwilling and unable to generate the trust necessary to jumpstart market development.

The roots of this destructive dynamic lay in the initial conditions of market emergence. Initial experiences of fraud and financial loss led Russian households to distrust financial organizations as well as the rules and norms that governed their marketplace. Although this period of "wild capitalism" during the early 1990s may be characterized as having provided a necessary, albeit harsh, lesson for inexperienced investors about the need for diligence in market exchange, we see the effects as less benign and more permanent. As distrust grew and became more ingrained, the competitive conditions in the deposit market changed in a way that further increased the gains to opportunism and decreased the returns to trust production. In a self-reinforcing process, fraud beget more fraud.

We draw on Zucker's (1986) distinction between "process-based" and "institutional-based" trust to examine both the cause and persistence of distrust in the Russian deposit market. Process-based trust, Zucker explains, is tied to the behavior of a specific actor. An individual or organization may either build a reputation for trustworthiness through repeated exchange or undertake some action, such as branding, that credibly signals to others one's reliability. Institutional-based trust, on the other hand, is defined as the product of formalized social structures. This form of trust originates not from the characteristics or history of a particular actor but from confidence in third-party organizations and political structures to monitor and enforce common standards of good conduct.¹ Since institutional-based trust does not require that partners to an exchange know about one another's past behavior, its development allows for an expansion of impersonal market-based transacting. Or conversely, as we show in the case of the Russian deposit market, growing distrust in the collective institutions of market exchange contributes to a contraction of impersonal exchange.

After outlining the origins of the Russian household deposit market, we examine and explain the macro-level data that demonstrate its fate from 1992 to 1999. We then explore the micro-institutional processes that underpinned the decline of the Russian market. We conclude with an analysis of why the initial

erosion of institutional trust in early market emergence can lead to a self-reinforcing process of market decline.

THE RISE AND FALL OF THE RUSSIAN HOUSEHOLD DEPOSIT MARKET

An understanding of why households grew to distrust private savings organizations must begin with a review of the rapid transformation of Russia's banking sector just prior to the dissolution of the Soviet Union's centrally planned economy. Until the late 1980s, the Soviet government prohibited all private financial organizations and *Sberbank* (the state-owned savings bank) was the only institution allowed to accept the savings of households. *Sberbank* operated thousands of branches and service counters across the country, offering both savings and demand deposits. After 1957 and the discontinuation of forced government bond purchases, households quickly began to increase their savings in *Sberbank* as state-run media advertised the confidentiality and liquidity of its deposits. Many branches offered payroll deduction plans, served as local collection points for telephone bills and distributed pension checks (Garvy, 1977). By the 1960s, *Sberbank* had developed into a trusted institution and an important presence in the daily lives of Soviet citizens.

The rest of the state banking system had no direct contact with households. It was a highly centralized organization, fully subordinated to the political leadership and its production plan for real goods and services. But in the late 1980s, it was restructured in a manner that would greatly influence the banking environment confronting households in the early 1990s. In 1988, as a part of Mikhail Gorbachev's *perestroika* reform program, the state bank was divided into two tiers, a central bank and several specialized banks charged with serving the financial needs of state-owned firms in specific sectors of the economy. Legislation passed in 1990 formally advanced the dissolution of the old state monobank by calling for the break-up of the specialized banks and their transformation into joint-stock, commercial banks. But changes were already occurring more rapidly at the sub-national level. As early as 1989, the government of the Russian republic required banks on its territory to transform themselves into commercial banks (Abarbanell & Meyendorff, 1997). In the next couple of years, as the Soviet state weakened and the government of the Russian republic grew in power and influence, hundreds of branches of the specialized banks split off from their parent organizations to form autonomous banking units, licensed by the Central Bank of Russia (CBR).

The *perestroika* reforms of the late 1980s also included measures allowing for the development of a private, cooperative sector. To support its development,

the ban on non-state banks was lifted, allowing new credit organizations to be established. The Soviet central bank set few formal requirements for their registration, gave them the right to access cheap financial resources from the state and established only modest capital requirements. Despite having set up little in the way of regulatory structures to oversee them, hundreds of new banks with no institutional roots in the old system were allowed to register in just a couple of years.

Because of these developments, in January 1992, just after the Soviet Union's break-up and concurrent to Russia's rapid economic liberalization, there were 1360 banks registered in Russia. The policies of the late Soviet period had left Russia the most heavily banked country in the former socialist bloc. Cheap credit from the CBR and a continuation of the lax policies on capital standards contributed to additional growth. By January 1995, the number of banking institutions in Russia climbed to 2527 (*Bulletin of Banking Statistics*).

In holding over 99.7% of all household deposits, Russia's *Sberbank* had a de facto monopoly in the household savings market when the institutional basis for private savings was laid down in January 1992. Almost immediately thereafter, many of the recently created commercial banks, as well as non-bank credit organizations, entered this new market. None of these new entrants had a reputation, either for good or bad, among potential depositors. Even though some had been hived off from the state bank and could thus trace their origins back to the Soviet system, they had not developed reputations as savings institutions. In other words, any sort of repeated interaction that could support "process-based trust" between households and the new institutions was largely absent. *Sberbank* was the only established player. And though it began the reform era with several distinct advantages, like its competitors, it had no real experience in honoring liabilities within a market setting.

Effective public regulatory bodies were also missing in the new household savings market. These would have to be created from scratch. The CBR's licensing procedures, for instance, were extremely lax; and capital requirements were negligible, allowing basically any group with the right political connections to open a bank. Furthermore, the CBR's capacity to monitor bank behavior, including the riskiness of loan portfolios, was virtually non-existent in the early 1990s. Although formal rules had been put in place (e.g. minimum reserve requirements and capital adequacy standards), most analysts believed that their effectiveness was limited, in part due to an insufficient number of well-trained personnel.

On the other side of the market, Russian households were quite naïve about these new financial organizations. They had had very little experience with making investment decisions under socialism and little or no exposure to

assessing financial risk. A population reared under centralized planning, after all, had limited exposure to bankruptcies and other stories of financial failure.

Ironically, perhaps because of households' ignorance, the market for their deposits grew quickly in the months following January 1992. According to official data, the net inflow of their ruble savings into bank deposits expanded quickly in the years following market liberalization. Figure 1 shows how households nearly doubled their net bank-based savings rate from the first half of 1992 to the first half of 1994.² This initial surge in the flow of savings into the banking system could be attributed exclusively to the increased rate of saving in new commercial banks. Indeed, by June 1994, after having been in operation for effectively two-and-a-half years, the private banking sector had accumulated 49.3% of the stock of all bank-based household savings. These developments were particularly noteworthy when considering the significant competitive advantages possessed by *Sberbank*. Not only did it have an extensive pre-existing branch network and an explicit government guarantee to stand behind its deposits, *Sberbank* also was able to offer many services not available from its competitors, such as transfers for utility fees, taxes and pensions.

By the summer of 1994, it appeared as though a robust market for household deposits had developed. By and large, the new market pitted the higher deposit rates (see Table 1) and heavier advertising of the new private banks against the services, locational convenience and apparent institutional stability of *Sberbank*. The private savings market seemed to be thriving in spite of the relatively short period in which the stock of impersonal trust had time to grow.

As they were increasing their savings in private banks, households also rapidly expanded their investments in new non-bank financial institutions. The Russian government estimated that up to 2,000 unlicensed companies operated on the financial markets between 1992–1994, attracting the equivalent of \$5–7 billion from 80 million Russian investors (Federal Commission, 1996). Although some were legitimate intermediaries, many were no more than pyramid schemes. MMM, the largest and most notorious of these, sold shares that promised to yield absurdly high returns in relatively short periods of time. Indeed, the initial wave of investors did make out well, since the company was able to pay them off with the investments of those that arrived later. The scheme could not last long, however. The company crashed in July 1994, its share price dropping from sixty-four dollars to fifty cents in the course of a single day. Investors lost between one and two billion dollars in personal savings.

Another company, Tibet, was not a bank in any legal sense, yet it was able to engage in a number of activities that mimicked retail banking. In nationally broadcast commercials in March 1993, it offered 30% monthly deposit rates, a nominal annual return of over 2000%. Over 600,000 investors from cities

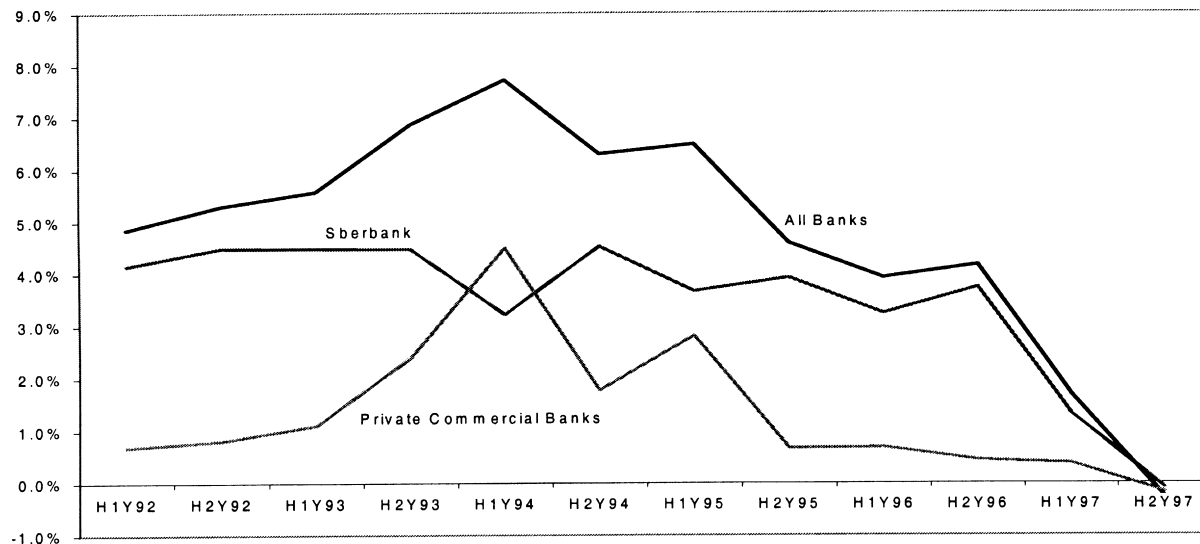


Fig. 1. Percentage of After Tax Personal Income Held in Bank Deposits, 1992–1997.

Note: The horizontal axis portrays time in half-year increments (e.g. H2Y93 refers to the second half of 1993). The percentages were calculated by taking the difference in the stock of deposits between January 1st and July 1st of the given time periods and dividing this by after-tax household income earned over the same period. The one exception is the H2Y97 data; the percentages there were calculated using changes in deposit data from July 1st to December 1st; this is due to a re-definition of the series (noted in the text) on January 1st 1998.

Source: *Bulletin of Banking Statistics*, Central Bank of Russia.

Table 1. Interest Rates on Time Deposits at Leading Banks in the Retail Market, March 1995.

Name of Bank	1-month	3-months	6-months	9-months
Sberbank	7%	22%	50%	83%
Inkombank	—	30%	60%	90%
Promstroibank Rossii	7%	24%	—	—
Mezhregionbank	9%	35%	—	—
Unikombank	9%	27%	54%	82%
Vozrozhdenie	—	36%	87%	153%
Tver'universalbank	7.9%	26%	—	—
Natsional'nyi kredit	10%	35.5%	86%	108%
Stolichnyi Bank	8.7%	33%	65%	98%

Note: Time series data on bank deposit rates are not available.

Source: Bazhenova et al. (1995).

throughout Russia placed the equivalent of \$3 million into the company. When Tibet went bankrupt after eleven months of operation, a mob of depositors stormed its headquarters to appropriate office equipment and other fixed assets to compensate for their lost savings.³

Although Tibet and MMM were among the most high profile cases, the general story line was not uncommon. According to Russian law, only organizations with banking licenses could offer deposit accounts for household savings. But in 1993 and 1994, over a hundred unlicensed financial companies in Moscow alone offered promissory contracts that were structurally equivalent to deposits at commercial banks (Kogut & Spicer, 2001). Their average life-span was seven months. By August 1995, only seven remained.

The inability of many of these un-licensed institutions to live up to their outrageous promises clearly had a negative impact on the ability of commercial banks to attract deposits. In fact, the collapse of MMM in July 1994 coincided precisely with the abrupt reversal of savings trends in private commercial banks. At this time, there had been no significant failures of licensed commercial banks.⁴ But households, perhaps understandably, did not draw clear distinctions between private banks and un-licensed pyramid schemes. Both, after all, were competing for the same household deposits. Looking back at Fig. 1, we see that the second half of 1994 was the first period in the market era during which there was a decline in the bank-based savings rate. Having saved in net terms over four percent of their income in private commercial banks in the first half of the year, households restricted the increase in deposits to under two percent of their income in the second half. In *Sberbank*, however, the rate of net inflows actually increased during the same period in which it was decreasing in the

private sector. *Sberbank*, that is, seemed to be the immediate beneficiary of the scandals in the private sector. But the failures of some private financial institutions seemed to have had a negative impact on all. In the three years following the MMM collapse, the flow of household savings into all commercial banks fell from roughly eight percent to under two percent of after-tax personal income. And though *Sberbank* did not entirely escape this trend, overwhelmingly the decline was due to an increased aversion to private banks.

Although the reversal of the 1992–1994 trend was precipitated by the collapse of non-bank credit organizations, it may have been exacerbated by the shaky performance of licensed banks a year later. For banks, 1994 had been a profitable year (Warner, 1998). But the insolvency of several large players on the overnight interbank market in Moscow produced a severe system-wide liquidity crisis in August 1995. The CBR provided temporary re-financing to some, but several major private banks were allowed to go bankrupt. These events, no doubt, only re-enforced a growing lack of confidence in banks. In 1996, private banks attracted less than 1% of households' after-tax personal income. In the second half of 1997, households actually reduced their stock of savings in private banks.

Table 2 provides further evidence for early optimism preceding a growing display of distrust toward private banks. A respected Russian polling agency asked a random sampling of citizens in at least four different Russian cities the following question: "If you have (or had) money savings, in what way would you prefer to keep them in the present situation?" Respondents were allowed to choose one or more answers from the following options: *Sberbank*; private commercial banks; government bonds; company shares; cash rubles; hard currency; in goods (e.g. jewels, antiques, paintings, etc.); or "other."⁵ The responses follow the same patterns as the aggregate data on actual household savings behavior. Through 1994, the public displayed a growing willingness to entrust their savings to private banks. In fact, between June 1993 and February 1994, the increase in the stated desire to save in private banks outpaced all of the other savings options.

But after 1994, the trend reversed. From over 20% of the population in February 1994, the percentage of those expressing a willingness to save in private banks dropped to just over two percent in June 1998. Only the bonds on which the Russian government would default later that summer were held in lower esteem. The increased aversion to private banks corresponded to an increase in the stated desire to save in foreign currency at home and relatively consistent responses to the desirability of saving in *Sberbank*.

Since official data on household savings before January 1, 1998 did not include foreign currency deposits, direct comparisons before and after that date are problematic.⁶ Nevertheless, we see in Fig. 2 that since 1998 private

Table 2. Survey Data on Intended Investment Decisions:
 “If you have (or had) money savings, in what way would you prefer
 to keep them in the present situation?”

	Respondents	<i>Sberbank</i>	Private Banks	Gov’t Bonds	Stocks	Cash	Foreign Currency	Goods
6/93	1628	42.4%	14.1%	4.9%	16.8%	14.8%	36.7%	26.9%
10/93	1392	36.7%	19.1%	4.5%	16.3%	13.7%	33.9%	31.7%
2/94	1628	35.3%	20.7%	3.3%	11.7%	12.0%	42.9%	29.7%
1/95	2363	35.2%	14.7%	3.1%	7.0%	13.3%	47.6%	17.6%
3/95	1584	37.5%	9.9%	1.2%	6.2%	7.5%	56.2%	20.0%
4/95	2030	37.0%	8.9%	3.0%	3.1%	11.8%	46.5%	15.4%
3/96	2001	46.5%	7.7%	3.0%	2.7%	21.9%	45.1%	19.6%
3/97	1933	39.4%	3.2%	4.1%	3.8%	19.1%	56.2%	15.3%
6/98	1902	31.2%	2.1%	0.6%	3.5%	26.0%	61.1%	17.1%
9/98	1991	29.8%	2.0%	0.6%	3.3%	24.9%	58.4%	16.4%
9/99	2002	29.9%	1.7%	1.4%	4.3%	18.8%	63.8%	15.3%

Note: Respondents were able to respond in multiple categories. Respondents who did not answer the question directly were excluded from this reporting of the results. The dates reflect the time when the surveys were conducted, not necessarily the publication date of the journal.

Source: Compiled from *Monitoring of Public Opinion*, various issues, 1992–2000.

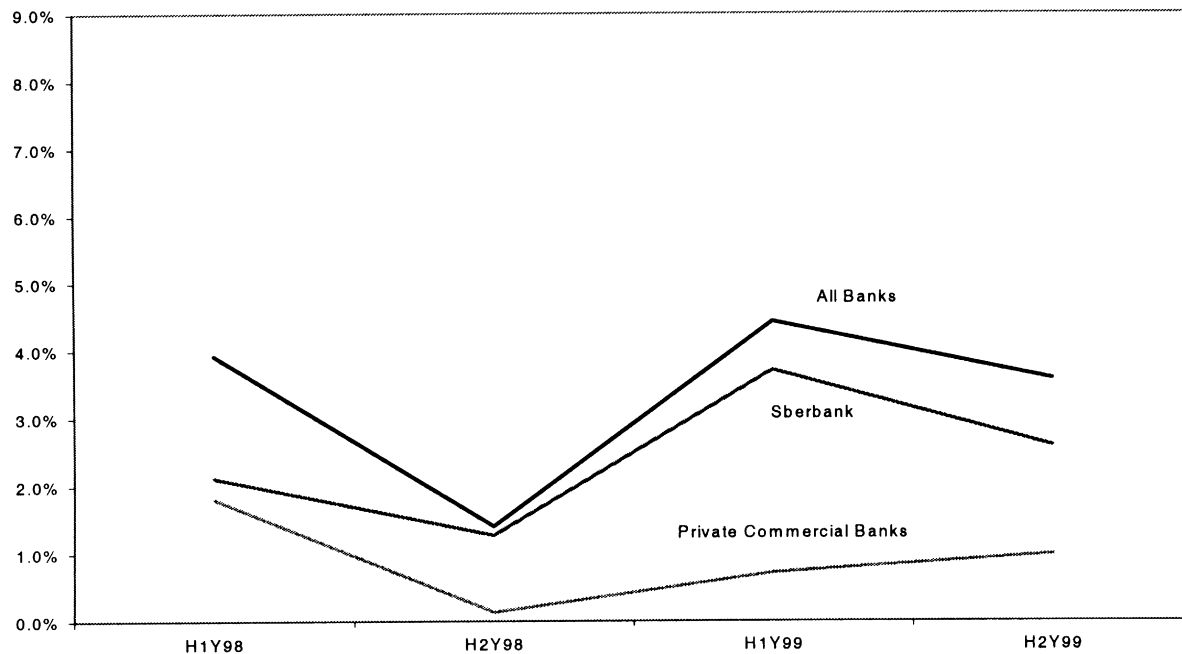


Fig. 2. Percentage of After-Tax Personal Income Held in Bank Deposits, 1998–1999
(Includes Foreign Currency Deposits).

Source: *Bulletin of Banking Statistics*, Central Bank of Russia (various issues).

commercial banks have continued to have trouble in attracting new inflows of deposits. The figure shows that, even though one might expect that households' capacity to save had increased because of the robust economic growth in Russia in 1999, and even though foreign currency deposits were now included in the deposit data, in net terms only about one percent of new income was deposited into private banks.⁷ The questionnaire data from this period confirm the extent of the distrust. Only 1.7% of those polled expressed a willingness to deposit monies in private banks.

In no small part, the perpetuation of banks' image problem was brought about by the macroeconomic crisis of 1998. In August, the Russian government devalued the ruble and defaulted on its bond obligations. Because of their unbalanced exposure to hard currency liabilities and ruble-denominated assets, including government securities, a number of banks were driven into insolvency. Many of the largest were unable to meet their obligations to depositors, reinforcing strongly-held beliefs that commercial banks were not to be trusted.

FIRM STRATEGY AND INSTITUTIONAL DISTRUST

In this section, we re-orient our focus from a macro-level discussion of market trends and depositor attitudes to the micro-institutional factors that underpinned them. Specifically, we examine why individual firms either would not or could not develop "process-based" mechanisms of trust to overcome the broader lack of confidence in the market. We argue that the conditions under which the market initially developed reduced any single bank's ability to benefit from, and thus willingness to engage in, trust-producing behavior. The population of depositors, in turn, learned to become more skeptical of private banks. Subsequent entrants into the deposit market failed to reverse this learning process. Instead, distrust bred a continuation of untrustworthy behavior in a self-reinforcing process of market decline.

Russian households' lack of experience in interacting with a private financial sector left them poorly prepared to cope with the new market in two ways. For one, their appreciation for financial risks, generally, was minimal. For another, they had no experience in discriminating among specific options in a competitive financial market. The distinction between a legitimate commercial bank and a fly-by-night pyramid scheme was not meaningful. With the exception of *Sberbank*, all financial organizations effectively "looked" the same to households.

Given the unformed nature of their beliefs about the financial sector, households' initial interactions with financial organizations would likely do much to shape their perception of the industry as a whole. That is, we might have expected there to be significant reputational externalities associated with

the behaviors of individual firms.⁸ The creation of institutional understandings of what “banks” were and how they were supposed to behave involved a process of collective learning on the part of inexperienced investors. During the initial phase of market formation, one financial institution’s good behavior (e.g. honoring deposit obligations in the promised manner) would positively influence the broader perception of the entire financial sector; bad behavior, conversely, would reflect negatively on all. Thus, we saw that when pyramid schemes collapsed, the reputation of financial organizations, collectively, suffered.

The presence of these externalities most likely discouraged behaviors that would have produced process-based trust. The logic here is straightforward. When transaction costs are non-trivial and externalities are present, self-interested actors engage in too much of the behavior generating negative externalities and not enough of the behavior producing positive externalities. For private banks in Russia’s nascent deposit market, the existence of reputational externalities weakened the incentives to behave well. The costs, after all, of pursuing a risky investment strategy that might lead to a failure to honor deposit contracts would not be fully internalized. Rather, they would be spread across the industry as a whole in the form of a damaged reputation. Likewise, the benefits of pursuing strategies that would increase the potential of meeting promised obligations would accrue to the collective.

The massive, un-regulated entry of both licensed and un-licensed financial institutions into the household savings markets likely exacerbated the effects of reputational externalities. If only a single bank had operated in this new market, then its actions would have affected only its own reputation. But with hundreds of banks in the market for a limited stock of savings, the costs of opportunism and the benefits of trust production, from the perspective of a single firm, were minimal.

The unstable political and economic situation in Russia during these years exacerbated these disincentives for building trust. Hyper-inflation and political struggles between the executive and legislative branches left financial organizations uncertain about the types of opportunities that would be available in the future, thus decreasing their effective time horizons and diminishing the expected returns from any type of long-term investment, including trust production. The prospects for developing long-term relationships with households were not great in the highly volatile environment that characterized the first years after market liberalization. It should thus not be surprising that firms might choose behaviors at odds with building the basis for long-term interaction. Firm-level actions aimed at slowly and incrementally building the trust of households were unlikely to be fully rewarded.

But what of strategies that might have a more immediate impact? Could a bank that was serious about developing a long-term presence in the retail market for deposits credibly and quickly signal its intention to potential clients? Could trust be won without an extended relationship? The mere claim to be a trustworthy bank would not, by itself, be meaningful since it would be a relatively costless signal to send. The better firms must find alternative means to distinguish themselves. It has been suggested that advertising expenditures could in certain contexts serve this function (Nelson, 1974). In markets for “experience goods,” clients of high quality providers should be more loyal and, consequently, more valuable to attract.⁹ Higher quality providers should thus be willing to incur greater up-front costs than other firms to attract new clients. Heavy expenditures on advertising (even advertising that is not particularly informative) could signal this willingness. A correlation between product quality and advertising expenditure might thus be expected in markets for experience goods.

The advertising expenditures of Russian banks and the credibility of their promises to depositors, however, do not exhibit this relationship. Avdasheva and Yakovlev (2000) provide evidence for a negative correlation between Russian banks’ advertising expenditures and their market longevity in 1994. The greater the amount of advertising that a bank carried out, the more likely it was to disappear from the market.¹⁰ Indeed, in the third quarter of 1994, over a third of all advertising costs was incurred by banks that were no longer in operation a year later. Depositors thus learned to discount advertising as a signal of credibility. In fact, Avdasheva and Yakovlev provide survey evidence showing that households attached little weight to advertising when choosing where to make deposits. Savers reported giving greater attention to interest rates, location and the advice of friends. Advertising was reported to be the least important factor.

The banks that engaged in heavy advertising and then failed to honor their liabilities produced a negative externality. In addition to the reputational externality noted above, their actions de-valued a potential signaling mechanism that could have served, at least in part, as a mechanism for firms to produce process-based trust. In her discussion of trust production, Zucker (1986, p. 65) observes that signals “cannot be solely symbolic – over time receivers will discount the signal unless it remains correlated with the underlying quality or characteristic it represents.” In Russia, investors learned to distrust not only commercial banks but also potential market signals that could have enabled them to distinguish between trustworthy and untrustworthy competitors.

One interpretation of the initial period of institutional learning in the Russian deposit market was that it represented a harsh, but necessary, lesson for

inexperienced investors about the need for diligence in market exchange. Indeed, one Russian banker noted that “The best lesson for the country as a whole seems to be the MMM fiasco . . . The scandal actually taught the Russian people the lessons of capitalism, perhaps for the first time.” (Yuri Lvov, advertisement for Bank St. Petersburg, *EuroMoney*, September 1994, 137, cited in Johnson, 2000, p. 112). Avdasheva and Yakovlev (2000, p. 180) are similarly upbeat about the impact of the initial market experience. They present evidence of a positive correlation between a bank’s advertising expenditures in 1995 and its longevity over the subsequent three-year period, arguing that the initial discrediting of advertising as a signal of quality laid the groundwork for its subsequent resuscitation:

Since the second half of 1995 the valuation of advertising by clients of private banks can be expressed as follows: ‘I know that most people do not trust advertising campaigns and advertising itself can hardly attract a large amount of money. If any private bank invests money in advertising, this type of investment cannot bring a quick return. This means that the bank is going to stay in the market for a long period.’

In other words, since it had previously proven to be un-correlated with short-term profitability, advertising would now only make sense for firms that planned to be in the market for a long time.

Unfortunately, Avdasheva and Yakovlev’s data end in early 1998 so they are unable to assess the impact of the financial crisis that hit in August. This significantly distorts their conclusions. As we will see, the outcomes for depositors following the August 1998 crash only reinforced the general distrust of commercial banks. Far from providing a valuable learning experience or reviving a potential signaling mechanism, the initial growth of depositor distrust contributed to some of the very behaviors that led to widespread bank failures in the aftermath of the 1998 crisis.

DEPOSIT BANKS AND THE AUGUST 1998 FINANCIAL CRISIS

In the mid-1990s, a number of new banks entered the deposit market to try to chip away at *Sberbank*’s still dominant position. In January 1998, the stock of deposits in private banks amounted to approximately 25% of a 160 billion ruble market; six private banks held over one billion rubles each (see Table 3). The largest of these was SBS-Agro. It had grown rapidly after 1994 when, as the Stolichny Bank of Savings (SBS), it had only sixty thousand private account holders. By 1996, SBS had five hundred thousand private depositors in sixty-seven branches (SBS Annual Report, 1996). That year, it also won a tender

Table 3. Top Banks in Deposit Market, Pre-August 1998 Crash.

Banks with over 1 billion rubles of deposits (Jan. 1998)	Total Deposits (Millions of Rubles)	Advertising (Rank Order – July 1998)	Reliability Rating (Jan. 1998)
Sberbank	123,869.3	5	A3 (top rating)
SBS-Agro*	6,701.2	2	A1
Inkombank	5,572.8	4	A2
Most-Bank	2,174.4	13	A1
Menatap	1,305.1	6	A1
Promstroibank Rossiya	1,203.8	30	A3
Rossiiskii Kredit	1,125.6	1	A2
Total Deposits in Commercial Banks	40,097.2		
Total Deposit in Top Six Commercial Banks	18,082.9		

* We combine the deposits of Stolichny Bank (SBS) and Agroprombank in this total.

Sources: Advertising is based on expenditures in July 1998, based on Gallup-International Figures (1998). Rating data is from Rating Information Center, www.rating.ru. Deposit data is from Interfax Rating Company.

to rehabilitate *Agroprombank*, the former state agricultural bank that had recently gone bankrupt. The combined organization, SBS-Agro, controlled over 1300 branches and offices around the country.

SBS-Agro and the other leading private deposit banks received among the highest ratings given by the Rating Information Center (RIC) – a Moscow-based financial information provider whose banking data had, at one point, been considered to be more reliable than the CBR's.¹¹ RIC's monthly ratings placed banks on a "reliability" continuum from extreme reliability (A3) to satisfactory stability (B1), with banks judged to be below the B1 standard not appearing in the rankings at all. The largest six commercial banks, in terms of household deposits, all received "A" grades for reliability at the end of 1997.

Despite their apparent health, the biggest players on the commercial deposit market proved to be disproportionately vulnerable to the effects of the 1998 financial crisis. Table 4 compares the fate of the top six deposit banks with the other commercial banks that received an "A" rating from RIC at the end of 1997.¹² By December 1998, all six of the top commercial deposit banks had fallen off RIC's rating list; these banks, that is, had either ceased to operate or did not merit even a "B" rating. On the other hand, of the sixteen other commercial banks that had received an "A" rating in the end of 1997, nine received a

“B” rating or above.¹³ This comparison shows the August 1998 financial crisis did not lead to the collapse of all the largest commercial banks in Russia. Some of Russia’s largest banks were able to survive with their reputations in tact. In contrast, all six of the leading deposit banks were unable to survive.

Given the general opacity of the Russian financial sector, we cannot know the precise intentions and motives of specific banks. We can, however, offer informed hypotheses as to why deposit banks were disproportionately affected by the August crash. One explanation is that attracting deposits required absorbing costs and assuming risks that turned out to be greater than initially anticipated. Winning depositors’ trust required large up-front investments in the form of heavy advertising expenditures (Avdasheva & Yakovlev, 2000). However, given the experiences from earlier in the decade, the ability of any given amount of advertising to increase deposits had likely been diminished. The negative correlation between advertising and institutional longevity in 1994–1995 undermined advertising’s value as a signal of longevity. In short, as depositors became increasingly skeptical of advertising, the advertising costs that needed to be incurred to attract a fixed amount of deposits rose.

The rising cost of attracting new customers into an increasingly discredited market may have pushed the banks intent on building their deposit base into assuming increasingly high levels of risk on the asset side of their balance sheets. The only way for many of these banks to generate the resources needed to attract new customers, given their waning trust, was to adopt high-risk investment strategies. Low-risk, low-return strategies would not generate the income needed to reverse the cycle of distrust. Indeed, relative to their liabilities, Russia’s leading deposit banks became disproportionately exposed to ruble denominated assets, particularly Russian government securities. Russian debt carried a significant risk premium in light of the government’s ongoing budgetary problems, but so long as the ruble remained strong and the government made good on its obligations, the deposit banks and their clients would

Table 4. Comparison of Rating Changes, Pre-Crash and Post-Crash.

December 1997 RIC List: All “A” Rated Commercial Banks (22 in Total)	December 1998 RIC List: Previously “A” Rated Banks Still on List	December 1998 RIC List: Previously “A” Rated Banks Removed from List
(6) Deposit Banks (From Table 3)	0	6
(16) Other Commercial Banks	9	7

Source: Rating Information Center, www.rating.ru.

remain solvent. But when the government devalued its currency and defaulted on its bonds in August 1998, banks wedded to this strategy were hit extremely hard. An independent audit of the pre-crash leaders in the deposit market revealed the following post-crash capital-to-assets ratios: Inkombank, -4.5; SBS-Agro, -2; Rossiiskii Kredit -1.5; and Menatap, -1.0 (*Russian Economic Trends*, 1999, Issue 1). In the immediate aftermath of the crash, these banks simply did not have the resources to meet their obligations to depositors.¹⁴

The impact of their risky investment strategies on household depositors was compounded by banks' grossly opportunistic behavior in the aftermath of the August crash. Recognizing the damage done to their reputation, banks were more apt to view their ongoing interactions with depositors as an endgame. With the prospects of re-building or even maintaining their deposit base in doubt, a "take the money and run" strategy became increasingly attractive. In mature market economies, failed banks enter a liquidation process in which external regulators shield a bank's good assets from departing managers. In Russia, however, the Central Bank's supervision of insolvent banks was both slow and opaque (*Moscow Times*, August 17, 1999, p. VIII). Many of the "failed" deposit banks, now less concerned about their reputations, exploited the weak regulatory environment, quickly transferring their remaining good assets into new legal entities (so-called "bridge" or "shadow" banks) with essentially the same set of owners.

New "bridge" banks held none of the liabilities of their "daughter" banks, leaving obligations to depositors in the hands of banks that just had many of their best-performing assets stripped from their control. For instance, soon after the August crash, Alexander Smolensky, the founder and head of SBS-Agro, helped establish the First Mutual Credit Society, a bank that absorbed many of the assets of his old institution but none of its liabilities. Smolensky claimed that the "rebirth" of his bank was completely legal even though many of SBS-Agro's depositors received only a small fraction of what was owed them (*Moscow Times*, October 10, 2000). Despite receiving more than \$200 million in stabilization credits from the CBR, SBS-Agro claimed to be unable to meet its obligations, offering many of its depositors settlement packages composed of office furniture and package tours (*Moscow Times*, August 17, 1999). Smolensky even mocked his former clients for their naivete, saying that his "hair stood on end" when he saw that some Russians had deposited more than \$1 million in his bank; they "must be idiots," he observed (*Moscow Times*, October 10, 2000, p. 6).¹⁵

Nearly every major deposit bank created a "bridge" bank. *Rossiiskii Kredit* created *Impeksbank* to continue its operations. *Menatap St. Petersburg* took over *Menatap's* forty-six Moscow branches and acquired several of its sister bank's

regional offices as well. *Promstroibank St. Petersburg* similarly took over the operations of *Promstroibank Russia*. *Inkombank* was the only major deposit bank that went bankrupt without transferring a large portion of its best-performing assets to a new organization with many of the same owners.¹⁶

In practice, the exact combination of factors that gave rise to a particular bank's unwillingness or inability to honor its deposit obligations is difficult to discern. Whether the mix of reasons was weighted more toward the macro-level shock, risky investment strategies, or asset stripping, the August events and their aftermath only validated depositors' feelings of distrust toward commercial banks. Household ruble deposits in commercial banks dropped by sixteen billion (47.6%) in the last half of 1998, while household deposits in foreign exchange decreased even more dramatically by \$2.3 billion (58%) (CBR 1998 Annual Report).

The money that did remain in the commercial banking sector migrated toward banks in which state bodies still played a significant role. As shown in Table 5, the state held significant ownership rights in four of the ten leading deposit banks in January 1999.¹⁷ These four experienced the largest absolute increases in deposits during the first half of 1999. Although only *Sberbank* offers deposits that are explicitly insured, the participation of state entities seemed to be increasingly interpreted by depositors as an implicit signal of an institution's stability.

Table 5. Top Operating Banks in the Household Deposit Market, Post-Crash.

Top Operating Banks in Deposit Market	Change in Ruble Deposits from January 1st–July 1st 1999 (1,000s of Rubles)	Bank's Rating (and changes) from 12/31/98 to 15/1/99
<i>Sberbank</i>	37,369,912	A2
<i>Gazprombank</i>	638,883	A1
International Moscow Bank	2,113	A1
Alfa	512,052	B3 to B2–
<i>Vneshtorgbank</i>	1,668,557	A1
Vozroshdeniya	–209,915	B2– to B2
<i>Bank of Moscow</i>	854,536	B3
Avtobank	–205,468	B3 to B2
Evvrofinance	282,201	B3 to B3–
Petrokommerts	123,748	B1 to B1–
Moscow Industrial Bank	–61,733	B3– to B2

Note: Italics designate banks with strong government ownership.

Source: Rating Information Center, www.rating.ru.

The correlation here between state participation and deposit mobilization stands in stark contrast to the absence of any clear relationship between deposit mobilization and the reliability grades of RIC. Presumably there are many factors that influence the choice of a household to deposit its savings with a given bank. The probability that the bank will be able to honor its obligation to depositors when they wish to withdraw funds is only one factor, albeit an influential one. The geographical proximity of a branch, the quality and range of services offered, as well, of course, as the rate of interest paid on deposits all play a role. We, therefore, should not expect to see all deposits being placed in the bank widely recognized to be the safest haven for deposits. Nor might we even expect to observe a particularly strong positive correlation between a bank's reliability ranking and the percentage of aggregate deposits attracted. We can, however, get at the answer of a rating's impact by examining whether or not a bank's ability to attract funds changes when its rating changes. Assuming that the other factors affecting deposits remain fixed, a bank could be expected to have more luck attracting funds when the measurement of its reliability improves. And conversely, a withdrawal of deposits after a downgrading in its reliability would not be a surprise. We might further hypothesize that such effects would be exaggerated during a period in which the stability of the banking system as a whole is in question. But, surprisingly, Table 4 shows that the change in deposits at these banks does not correlate at all with the reassessment of their reliability by the rating agency.

The seeming ineffectiveness of RIC represents one more piece of evidence of the failure to produce trust in the Russian retail banking market. As Table 3 pointed out, virtually all of the top deposit-banks had received an "A" rating before the crash. Yet, most of these banks proved to be either unable or unwilling to fulfill their obligations to depositors in the latter half of 1998. Coarse group-level differentiating mechanisms – whether a bank was government-owned or privately-owned – proved more useful to depositors than more refined rating information on individual banks.

DISCUSSION AND CONCLUSION

When Russian financial markets were liberalized in 1992, hundreds of new private financial institutions offered up their services to households that possessed both the means and the desire to save. Both a well-defined supply of and demand for savings deposits clearly existed. However, a well-functioning market that linked the two was not yet in place, and the degree to which Russian households would entrust their savings to the new commercial banks was uncertain. Despite the potential gains to trade in the Russian deposit market,

distrust toward market actors and institutions led to a vicious circle of market decline.

The macro-economic consequences of these developments should not be under-estimated. Government estimates place the amount of cash that Russians have stashed at home, outside the formal financial system, in the tens of billions of dollars. Despite the re-structuring demands of former state-owned enterprises and the capital needs of start-ups, Russia's banks have done little lending to the private sector (EBRD, 1998). Although this is in part due to difficulties in assessing creditworthiness and enforcing loan contracts, the inability of Russia's banks to mobilize deposits has played a large role as well.

What explains Russian banks' inability to reverse the cycle of distrust? Diego Gambetta's (1988, p. 234) explanation of the self-reinforcing nature of distrust parallels our own:

Deep distrust is very difficult to invalidate through experience, for either it prevents people from engaging in the appropriate kind of social experiment or, worse, it leads to behavior which bolsters the validity of distrust itself . . . Once distrust has set in, it soon becomes impossible to know if it was ever in fact justified, for it has the capacity to be self-fulfilling, to generate a reality consistent with itself. It then becomes individually "rational" to behave accordingly, even for those previously prepared to act on more optimistic expectations.

Gambetta emphasizes the role of beliefs in defining "rationality" in local settings. If individuals distrust actors of a certain type, then they will refrain from "engaging in the appropriate kind of social experiment" that might lead to a reconsideration of those beliefs. In Russia, growing distrust toward commercial banks caused potential depositors to avoid transacting with private commercial banks, the very type of "social experiment" that had the potential to allow individual banks to establish a reputation for reliability. In fact, as the market for privately-provided deposits sputtered, the incentive structure of private commercial banks changed. Short-term opportunism, a "take the money and run" strategy, became increasingly attractive. Distrust bred self-confirming actions.

Certainly, the Russian case reinforces the need for governmental supervision in nascent financial markets (Stiglitz, 1993). In order to avoid a cycle of distrust, Russia would have needed state regulatory bodies that were both willing and able to curb the opportunistic behaviors of early market entrants. Just as importantly, the Russian case highlights that the creation of institutional trust fundamentally rests on the beliefs and attitudes of those who most rely on institutions in everyday practice. Breaking the cycle of distrust will require not only regaining the confidence of depositors in financial organizations, but also developing confidence that state structures can credibly control the boundaries of legitimate competitive behavior. The history of the Russian market has

demonstrated again and again that both private and public organizations have had little success in defining the rules and norms of Russia's new markets.

Putnam's analysis of variation in the economic performance of Italian regions provides a comparative case for the Russian experience. He (1993, p. 177) suggests that the economic and political performance of an Italian region depends on whether it has created a high-trust or a low-trust social equilibrium:

Stocks of social capital, such as trust, norms and networks, tend to be self-reinforcing and cumulative. Virtuous circles result in social equilibrium with high-levels of cooperation, trust, reciprocity, civic engagement and collective well-being. These traits define the civic community. Conversely, the absence of these traits in the uncivic community is also self-reinforcing. Defection, distrust, shirking, exploitation, isolation, disorder and stagnation intensify one another in a suffocating miasma of vicious circles.

As applied to the study of arms-length exchange, the self-reinforcing nature of trust suggests two different equilibria. Successful markets have established a virtuous cycle of trust-production that provides both short and long-term benefits to high-levels of cooperation and trust. Yet, the Russian case suggests that the formation of the virtuous circle of initial market creation is in itself a significant challenge. A vicious cycle of distrust may just as easily lead to a low-equilibrium trap that brings growing benefits to defection and distrust.

Guiso, Sapienza and Zingales (2001) similarly demonstrate the importance of collective trust in influencing market development in Italy. Using similar measures as Putnam to measure differences in social capital across Italian regions, they find that in regions with higher levels of social capital, households invest more in stock, use checks more frequently and have greater access to institutionalized sources of credit. That is, trust correlates with a more important role for impersonal financial transactions. Similar findings are now becoming apparent in Russia. Commercial banks rely on personal ties and networks to develop and retain customers, as they have been unable to develop the impersonal markets in deposits, commercial lending or credit card activity that provide the foundation of banking activity in established western markets (Dinello, 1999; Guseva & Rona-Tas, 2001).

Our study, however, differs from the low-trust equilibrium traps identified in the research on Italy. The choice of financial instruments in Guiso's et al. (2001) study depends on a measure of regional social capital that remains relatively stable across long periods of time. The degree of social capital in a region is taken to be exogenous to the everyday activities of financial markets. Our examination of the formation of the Russian market, however, emphasizes the endogenous formation of institutional beliefs in initial periods of market creation. Cultural arguments that Russians were somehow disinclined to participate in financial markets seems to be at odds with actual events. In fact, Russians were

especially eager to enter the new markets in the early 1990s. Collective beliefs emerged from the negative experiences that these neophyte investors suffered at the hands of unscrupulous financial operators. Russian depositors did not act on culturally determined stereotypes of private banks; instead they created new stereotypes based on their lived experience with these new private actors.

Raiser et al. (2001) provides strong supportive evidence, recently collected through the World Values Survey, to suggest that Russians' distrust of private organizations is endogenous to the post-communist transition. Respondents were asked to rate their degree of confidence in over a dozen public and private sector institutions: the press, the legal system, the civil service, the enterprise sector, etc. As recently as 1990, Russians' measure of confidence in private companies, 2.37, was higher than the average for countries in Eastern Europe and the former Soviet Union, 2.18. In 1995, however, Russians' lack of confidence in companies stood out as the lowest of the twenty transition countries surveyed. Of course, during this five-year period, Russia went through a whole series of changes that would re-shape social institutions and attitudes. But to a large extent these changes were mirrored elsewhere in the former socialist bloc. As we can see in Table 6, popular confidence in the legal system, the civil service and the press changed in Russia much as it did in other transition economies.

But unlike other countries in the region that, on average, experienced a growing level of popular confidence in companies, Russian companies experienced a dramatic erosion of trust from 1990 to 1995. Most likely, the experience of Russian households with the nascent, unregulated financial sector had a large impact on these numbers. Interestingly, Albania, another country that had a particularly bad experience with the bankrupting of poorly controlled pyramid schemes, scored the second lowest in terms of the trust of companies in 1995. Although it is difficult to establish a direct causal link here, the

Table 6. Change in the Level of Confidence in Formal Institutions Between 1990–1995, World Value Surveys.

	Legal System	Civil Service	Press	Companies
Transition Countries	–0.03	–0.03	–0.13	+0.11
OECD Countries	+0.10	0.00	–0.06	+0.02
Russia	–0.03	0.02	–0.11	–0.52

Note: The numbers in the table represent the difference between the scores for 1995 and 1990 from the World Values Surveys on confidence in formal institutions. A positive number represents an increase in the level of expressed confidence toward an institutional form during this time period.

Source: Raiser et al. (2001).

loss of trust in private financial organizations most likely gave rise to an environment in which the level of skepticism toward all private enterprises (even rating agencies) rose.

An early lesson of economic liberalization in Russia is that market-supporting institutions cannot emerge solely through the competitive interaction of private firms. Private banks have proven unable to produce the trust necessary to develop and sustain the household deposit market. In fact, firm-level strategic behavior in this sector often has operated at cross-purposes to the creation of a sustainable system of arms-length exchange. An important question for future market development in Russia is whether new institutional arrangements can redirect the cycle of distrust to encourage greater rewards to cooperation and trust. To achieve the collective goal of successful market development, both private and public actors will need to cooperate to construct and sustain a framework of enforceable and reliable rules that will regain the trust lost in earlier periods of market formation.

NOTES

1. The added value of using the concept of “trust” in institutional analysis is to highlight the cognitive micro-foundations of successful institutional operations. The challenge to solving problems of opportunism and vulnerability in impersonal exchange extends beyond creating formal structures to monitor and enforce contracts. It also involves the creation of new beliefs and attitudes among the users of these formal structures that rules and norms will be applied in a consistent and unbiased manner. See Bigley and Pearce (1998), Kogut and Spicer (2002), and Shapiro (1987) for a description of the role of impersonal trust in market exchange.

2. The data in Fig. 1 were calculated by dividing the change in the stock of savings over a period of time (the net flow of savings) by after-tax household income over the same period of time.

3. For information about MMM, see *Izvestiya*, December 11, 1994, December 18, 1994, January 19, 1995; *Nezavisimaya Gazeta*, July 28, 1994 pp. 1–2; *Chicago Tribune*, August 16, 1994. For information about Tibet, see *Kommersant-Daily*, No.49, March 18, 1995 or *Kommersant*, No. 11, March 29, 1994, p.48.

4. There were no recorded bank failures in 1992, 19 in 1993 and 24 in the first half of 1994.

5. Responses of “hard to answer” were not included in the summary data presented in Table 2.

6. Somewhat inexplicably, these monies had previously been mixed in with the foreign currency settlement accounts of firms. We thank Mikhail Matovnikov for clarifying this point.

7. We show later in the paper that even the 1% investment figure in “private” banks is misleading. Many of the banks that attract new deposits outside of Sberbank are themselves government-owned or controlled. For instance, the Bank of Moscow,

which enters the deposit market during this time period, is controlled by the city of Moscow.

8. We use the concept of “reputational externalities” in a manner similar to Delacroix and Rao (1994).

9. “Experience goods” are products whose characteristics are not discernable immediately through inspection to potential customers. Rather, they are revealed over time through exposure and/or use. The veracity of a financial institution’s promise to honor its liabilities possesses this characteristic.

10. Avdasheva and Yakovlev (2000) only examine financial organizations that had formal licenses to operate as banks. If unlicensed financial companies like MMM were included in the analysis, then the proposition that advertising expenditures were a sterile signal of future behavior is even stronger. MMM incurred huge advertising expenditures to attract new investment into its pyramid scheme.

11. In the mid-1990s, the CBR published RIC’s data in its own in-house statistical abstracts. RIC’s rankings also appeared in prominent news and business periodicals with great frequency and regularity. In compiling its ratings, RIC relies in part on financial data assembled and released by the banks themselves. But given the relative ease of manipulating financial information in Russia, RIC emphasizes that it supplements these data with on-site interviews and frequent consultation with their own network of industry experts and insiders.

12. The other banks on the list were Alfabank, Avtobank, Conversbank, Imperial, International Company for Finances and Investment, International Industrial Bank, International Moscow Bank, Mezhscombank, Mosbiznesbank, Moscow Industrial Bank, Orgbank, Rosnevshtorgbank, Tokobank, Uneximbank, Vozroshdeniya. As we have already included SBS-Agro, we do not include Agromprombank separately.

13. The difference in the rating changes between the two groups (“A” rated deposit banks vs. “A” rated non-deposit banks) is statistically significant (chi-square = 5.712, $p = 0.017$).

14. Foreign investors seemed to make the same mistakes as domestic investors in trusting their money with the banks that were most likely to disappear following the August 1998 crash. The top three foreign debtors before the crash were SBS-Agro (\$1.196 billion in foreign debt), Inkombank (\$820 million) and Rossiiskii Kredit (\$648 million) (*Kommersant*, 1998, reported in Johnson, 2000, p. 210). Moreover, Russian banks defaulted on billions of dollars in forward currency contracts signed with foreign investors. Inkombank, for instance, reneged on an estimated \$1.88 billion in currency contracts following the devaluation of the ruble in 1998 (Johnson, 2000, p. 212).

15. Smolensky was no less kind to his foreign creditors. When asked about the foreigners who lost more than \$1 billion in SBS-Agro, he replied that they deserved only “dead donkey ears” (*Wall Street Journal*, October 2000, cited in Hoffman, 2002, p. 440).

16. Inkombank’s bankruptcy was also tarnished with scandal. The U.S. government alleged that Inkombank had been a central player in moving billions of dollars out of Russia secretly through Bank of New York accounts. See Johnson (2000, p. 132) and *USA Today*, September 24, 1999 for a description of the accusations of illegal activity made against Inkombank.

17. For instance, *Gazprom*, a natural gas monopoly with a high-degree of state ownership, controls *Gazprombank*; and the city of Moscow controls the Bank of Moscow.

ACKNOWLEDGMENTS

We would like to thank the University of California's Academic Senate, the A. Gary Anderson Graduate School of Business and Middlebury College for funding to support this research. We also thank Wendy Bailey, Bruce Kogut, Livia Markoczy and Mikhail Matovnikov for helping us develop the ideas present in this draft.

REFERENCES

- Abarbanell, J. S., & Meyendorff, A. (1997). Bank Privatization in Post-Communist Russia: Case of Zhilsotsbank. *Journal of Comparative Economics*, 25(1), 62–96.
- Avdasheva, S., & Yakovlev, A. (2000). Asymmetric Information and the Russian Individual Savings Market. *Post-Communist Economics*, 12(2), 165–185.
- Bazhenova O., Skvortsov, Y., & Katsman, Y. (1995). Nastupil god aktivnoi bor'by za den'gi naseleniia. *Kommersant-Daily*, No. 42 (March 9th).
- Bulletin of Banking Statistics*. Central Bank of the Russian Federation. Moscow: Russia.
- Bigley, G., & Pearce, J. (1998). Straining for Shared Meaning in Organizational Science: Problems of Trust and Distrust. *Academy of Management Review*, 23(3), 405–421.
- Delacroix, J., & Rao, H. (1994). Externalities and Ecological Theory: Unbundling Density Dependence. In: J. Baum & J. Singh (Eds), *Evolutionary Dynamics of Organizations*, New York: Oxford University Press.
- Dinello, N. (1999). The Russian F-Connection: Finance, Firms, Friends, Families and Favorites. *Problems of Post-Communism*, 46(1), 24–33.
- European Bank for Reconstruction and Development (EBRD). (1998). *Transition Report 1998: Financial Sector in Transition*. London: EBRD.
- Federal Commission for the Securities Market. (1996). *Training Manual for Perspective Unit Fund Managers*. Manuscript (In Russian).
- Gambetta, D. (1988). Can We Trust Trust?. In: D. Gambetta (Ed.), *Trust: Making and Breaking Cooperative Relations*. New York: Basil Blackwood Ltd.
- Garvy, George. (1977). *Money, Financial Flows, and Credit in the Soviet Union*. Cambridge, MA: Balinger Publishing.
- Guseva, A., & Rona-Tas, A. (2001). Uncertainty, Risk and Trust: Russian and American Credit Card Markets Compared. *American Sociological Review*, 66(5), 623–647.
- Guiso, L., Sapienza, P., & Zingales, L. (2001). The Role of Social Capital in Financial Development. *Working Paper No. 511*. The Center for Research in Security Prices.
- Hoffman, D. (2002). *The Oligarchs*. New York: Public Affairs.
- Johnson, J. (2000). *A Fistful of Rubles: The Rise and Fall of the Russian Banking System*. Ithaca: Cornell University Press.
- Klein, D. (2000). *Reputation: Studies in the Voluntary Elicitation of Good Conduct*. Ann Arbor: The University of Michigan Press.
- Kogut, B., & Spicer, A. (2002). Capital Market Development and Mass Privatization are Logical Contradictions: Lessons from the Czech Republic and Russia. *Industrial and Corporate Change*, 11(1), 1–37.

- Kogut, B., & Spicer, A. (2001). Ideas, Organizations and Politics in the Russian Transition. Unpublished manuscript.
- Milgrom, P., North, D., & Weingast, B. (1990). The Role of Institutions in the Revival of Trade: the Law Merchant, Private Judges, and the Champagne Fairs. *Economics and Politics*, 2, 1–23.
- Monitoring of Public Opinion*. Russian Centre for Public Opinion Research (VTsIOM). Moscow: Russia.
- Nelson, P. (1974). Advertising as Information. *Journal of Political Economy*, 84, 729–754.
- Newman, J. W.. (2000). Dun and Bradstreet: For the Promotion and Protection of Trade. In: D. Klein (Ed.), *Reputation: Studies in the Voluntary Elicitation of Good Conduct*. Ann Arbor: The University of Michigan Press.
- Putnam, R. (1993). *Making Democracy Work: Civic Traditions in Modern Italy*. Princeton, NJ: Princeton University Press.
- Raiser, M., Haerpfer, C., Nowotny, T., & Wallace, C. (2001). Social Capital in Transition: A First Look at the Evidence, *Working Paper No. 61*. European Bank for Reconstruction and Development (EBRD).
- Russian Economic Trends. *Russian European Centre for Economic Policy*. Moscow: Russia.
- Shapiro, S. (1987). The Social Control of Impersonal Trust. *American Journal of Sociology*, 93(3), 623–58.
- Stiglitz, J. (1993). The Role of the State in Financial Markets. *Proceedings of the World Bank Annual Conference on Development Economics*. Washington, D.C..
- Warner, A. (1998). The Emerging Russian Banking System. *Economics of Transition*, 6(2), 333–347.
- Zucker, L. (1986). Production of Trust: Institutional Sources of Economic Structure, 1840–1920. *Research in Organizational Behavior*, 8, 53–111.