The Past, Present, and Future of Economics: A Celebration of the 125-Year Anniversary of the *JPE* and of Chicago Economics

Introduction

John List

Chairperson, Department of Economics

Harald Uhlig

Head Editor, Journal of Political Economy

The *Journal of Political Economy* is celebrating its 125th anniversary this year. For that occasion, we decided to do something special for the *JPE* and for Chicago economics. We invited our senior colleagues at the department and several at Booth to contribute to this collection of essays. We asked them to contribute around 5 pages of final printed pages plus references, providing their own and possibly unique perspective on the various fields that we cover.

There was not much in terms of instructions. On purpose, this special section is intended as a kaleidoscope, as a colorful assembly of views and perspectives, with the authors each bringing their own perspective and personality to bear. Each was given a topic according to his or her specialty as a starting point, though quite a few chose to deviate from that, and that was welcome. Some chose to collaborate, whereas others did not. While not intended to be as encompassing as, say, a handbook chapter, we asked our colleagues that it would be good to point to a few key papers published in the *JPE* as a way of celebrating the influence of this journal in their field. It was suggested that we assemble the 200 most-cited papers published in the *JPE* as a guide and divvy them up across the contributors, and so we did (with all the appropriate caveats). Some chose to fea-

[*Journal of Political Economy*, 2017, vol. 125, no. 6]
 0022-3808/2017/12506-0037$10.00

ture these centrally, some picked a few, and others ignored that list altogether. Some have chosen to provide an overview, some have chosen to give their perspective on the field or added a wish list of their own, some have zoomed in on one contribution in particular, and one of the authors actually added a new theorem, generalizing a previous key contribution. There was a large canvas to draw on, and we are pleased that our colleagues have explored all its corners, in their own and unique way, in the spirit of the multifaceted mosaic that our faculty represents within the subject of economics.

We asked that their contribution be about what the field has accomplished or about where the field might or should be going in the future. It is probably the nature of the beast that all chose a largely backward-looking perspective, providing an overview of how the field has developed over time and how the *JPE* helped this process along by publishing some of the key ideas and key contributions. But hop on board and start reading! Because of the intended density of the summary, it then actually becomes quite a forward-looking piece. The speed at which the ideas fly by makes it hard to stop that train of thought at the present time and frontier: one immediately yearns to proceed and to do and learn more, once this flow has stopped. This then is the ultimate goal: to encourage new thinking and excellent, thought-provoking, and paradigm-shifting research, which will pave the way for a future and better understanding of economic phenomena and to have the *JPE* continue to participate in publishing those key contributions. When we were PhD students or young researchers, we wish we had had access to a dense overview of this sort. If you belong to that group, or if you are simply young at heart, and seek to read a research-stimulating piece, then this is our present to you.

There is no good and certainly no linear way of ordering these contributions. Printing technology still demands that the pieces be presented in some order: so here is ours, with all due apology. We have chosen to open with the most general, most bird's-eye, and most long-run of perspectives to then gradually "zoom in" toward the deep foundations and the essays discussing assignment problems as well as experimental approaches. We can hear the outcry already: wait, isn't economics all about assigning goods to agents? We share that concern. The reader is thus encouraged to browse, pick, read, and reorder to her or his heart's content.

Indeed, what could possibly be more general than general equilibrium theory? Hugo Sonnenschein reflects on the role of "Chicago and the Origins of Modern General Equilibrium" in his essay, juxtaposing the Arrow-Debreu Walrasian analysis to the excess demand function approach and discussing the role of preference orderings. Ufuk Akcigit writes about "Economic Growth," its crucial impact on the welfare of nations, and the crucial role of innovations for understanding the growth process itself. David Galenson complements this with the "Economic History" perspective, ex-

amining in particular the role of human creativity for innovations. He argues that conceptual innovators are young, while experimental innovators are, shall we say, experienced. Long-run growth requires the assignment of property rights and effective institutions for contract enforcement, issues intimately tied to the development of political institutions and political decision making, as Roger Myerson discusses in his essay on "Political Economics in the *Journal of Political Economy*." A crucial aggregate dimension of political choice is "Aggregative Fiscal Policy," as summarized by Nancy Stokey, highlighting in particular the role of government debt financing.

Zooming in a bit more on the shorter-run aggregate fluctuations and the international trade dimension is the topic in "Business Cycles and International Trade" by Harald Uhlig, emphasizing the role of rational expectations in particular. Greg Kaplan points out that the richer aspects of "Inequality, Heterogeneity, and Consumption" are "today . . . front and center in macroeconomics," building up to the analysis of general equilibrium models with heterogeneous households and aggregate shocks. These aggregate shocks expose households to risks, which need proper analysis with "Time-Series Econometrics in Macroeconomics and Finance," as Lars Hansen highlights, emphasizing in particular the role of consumption, permanent income, and asset pricing. The asset pricing aspects are then further elaborated on in the essay on "Asset Pricing: Models and Empirical Evidence" by George Constantinides, highlighting the role of preferences, and the piece by Eugene Fama on "Finance at the University of Chicago," narrating the transition to modern finance and the Chicago-led research on market efficiency, factor models, and the economics of organizations. Richard Thaler offers a different approach to thinking about financial markets, savings, and consumption in his discussion of "Behavioral Economics." He contrasts it with "price theory à la Chicago School led by the intellectual giants Gary Becker, Milton Friedman, and George Stigler," which indeed plays a central role in many of the other contributions here. He then questions whether the stock market can add or subtract, why there are noise traders, and why optional but then mandatory savings plans can help to increase savings.

The issue of "Corporate Finance" receives a more detailed treatment by Robert Vishny and Luigi Zingales in their contribution, organizing it around the Modigliani-Miller irrelevance proposition, agency costs, and the market for corporate control. Particularly important for all these financial dimensions is the banking and monetary system. Thus, Douglas Diamond, Anil Kashyap, and Raghuram Rajan write about "Banking and the Evolving Objectives of Bank Regulation," emphasizing the importance of liquidity provision and aggregate liquidity shortages for thinking about financial regulation. Liquidity is the topic at the heart of "Monetary Economics," which Fernando Alvarez summarizes, organized around the three

traditional functions of money. Robert Lucas reflects on his "Memories of Friedman and Patinkin," their perspectives on quantity theory as well as its juxtaposition or lack thereof to the Keynesian paradigm and its liquidity trap unemployment equilibrium.

This is naturally followed by a more in-depth examination of "Labor Markets" by Robert Shimer, taking a mostly macroeconomic perspective on the Phillips curve, search, labor supply elasticities, sectoral shifts, as well as contractual issues and their frictions. The more microeconomic perspective is the topic of "Chicago Labor Economics" by James Heckman, emphasizing the rigorous interplay between data and theory. He formulates the guiding principles of Chicago tradition that "theory is used to interpret data. Data are used to test theory. Understanding the mechanisms . . . is essential." This interplay between data and theory is likewise the key for Stephane Bonhomme and Azeem Shaikh, arguing to "Keep the ECON in Econometrics," an intended pun on Leamer's admonition to take the con out of econometrics. Showcasing that interaction, they discuss structural econometric models of the labor market, the marriage market, and partial identification. Derek Neal returns to the finer theoretical details of the labor market for understanding "Life Cycle Wage Dynamics and Labor Mobility," examining issues of life cycle investment in general human capital, life cycle wage growth, and the search for good matches. This naturally leads to the examination of "The Human Capital Approach to Intergenerational Mobility" by Magne Mogstad, giving central importance to two papers by Becker and Tomes, the first of which was published in the *JPE*. Human capital also plays a central role in Robert Topel's essay on "Health Economics," taking the perspective that health can be viewed as human capital. He notes that "net of medical expenditures, the value of increased life expectancy between 1970 and 2000 in the United States" equaled "a flow of about $2 trillion per year," thus emphasizing the quantitative importance of this topic. Providing a different perspective on the issues arising in labor market contracts or contractual issues more generally is the aim of "Agency Issues" by Canice Prendergast. He starts from a theoretical perspective, relating pay to performance, points out that "most people do not get paid this way," and then moves to discuss other motivations such as career concerns. Effectively echoing the guidelines by James Heckman, Stephane Bonhomme, and Azeem Shaikh, he bemoans the "paucity of empirical work" to complement these theoretical insights. Long-term contracts often are solutions in environments, featuring asymmetric information and associated incentives. "Information Economics" is the theme of Emir Kamenica's essay. He points out that additional issues such as information acquisition and communication arise and that information design is poised to complement mechanism design, when thinking about arrangements between agents.

Turning from labor markets to goods markets, Michael Greenstone traces "The Continuing Impact of Sherwin Rosen's 'Hedonic Prices and Implicit Markets: Product Differentiation in Pure Competition,'" emphasizing the power of the paradigm as well as the substantial empirical challenges in utilizing it, met with increasing, but as-of-yet incomplete, success by the "credibility revolution" of exploiting quasi-experimental variation, starting in the late 1990s. Price systems ultimately serve the goal of assigning goods to individuals. As Phil Reny shows in his piece on "Assignment Problems," these can be quite thorny, in particular if no monetary transfers are allowed or if individual preferences over goods exhibit complementarities. For that, he discusses an important result by Eric Budish, published in the *JPE* in 2011, and then extends it to allow for arbitrary preferences and both divisible and indivisible goods, providing a new theorem and its proof. A particularly important mechanism of assigning goods to buyers is auctions. Ali Hortaçsu therefore summarizes the literature on "Auctions in the *Journal of Political Economy*, 1894–2017." He highlights the literature on the auctioning of government land, the wireless spectrum, Treasury securities, and electricity, pointing to the issues arising from complementarities and potential collusion. A decidedly less desirable procedure for assigning goods results from criminal activity. Thus, the underlying economic motives as well as methods for deterrence are the focus in "The Economics of Crime," described by Steven Levitt. He emphasizes how that literature has taken an empirical turn in recent years, exploiting natural experiments, and that many questions remain open. A common theme in this as well as many of the other essays is that data are sorely needed to examine and test existing theories or to develop these theories further, but that it may be hard to find such data in the first place. Conducting experiments, in particular experiments in the field, to generate such data should then prove particularly compelling, as John List argues in his essay "Experimental Economics in the *Journal of Political Economy*." He discusses how experiments have already proven useful to examine market institutions and individual choice, and how these experiments moved from the lab to the field and a natural setting, deepening our understanding of the underlying phenomena.

The *Journal of Political Economy* and the University of Chicago are proud to have been active participants in these exciting lines of inquiries over the years. The essays make clear that much has been accomplished but that much remains to be done. It is our hope and desire that the *JPE* will remain at the forefront of these new and exciting developments and that authors will continue to consider the journal as their outlet of choice for their path-breaking contributions. In that spirit, we are looking forward to the next 125 years.

Chicago and the Origins of Modern General Equilibrium

Hugo F. Sonnenschein

University of Chicago

I.

In the early 1950s, with substantial time together at the University of Chicago, Kenneth Arrow and Gerard Debreu (1954) provided a first mathematically rigorous proof of the existence of equilibrium in a general Walrasian model.[1] Before that time, aside from some specialized modeling, the equality between the number of equations and unknowns in the Walrasian model was taken to be the basis for believing that every Walrasian system contains at least one equilibrium.[2]

With the result of Arrow and Debreu (1954), we learn that the Walrasian model is fully able to determine the prices of all commodities as a function of tastes, technologies, and the distribution of wealth. Their findings made rigorous the interdependencies that are essential to the determination of prices. We have, for the first time, an adequate general equilibrium theory of value. Their result is a cornerstone of the foundation of multimarket price theory. It demonstrates that in every well-specified Walrasian economy there is always at least one set of relative prices that balance supply and demand in all markets.

The Arrow-Debreu existence theorem is noteworthy in other important dimensions. First, it provides a first modern statement of the Walrasian model. It also promoted a standard in economics for precision of formulation and mathematical rigor. In addition, the approach can be seen to explicitly unite the issue of the existence of equilibrium for situations of imperfect competition, as in the Cournot model, with the issue of the existence of Walrasian equilibrium and perfect competition. Finally, with the benefit of hindsight, the approach has had a remarkable ability to anticipate and facilitate several important advances in the existence theory.

The perspective on the Debreu and Arrow-Debreu contributions presented here was developed in joint work with Wayne Shafer. A central point of this note is communicated in Duffie and Sonnenschein (1989). Philip Liang provided most helpful research assistance.

[1] The article "a" is carefully chosen. Lionel McKenzie (1954) presented a general proof of the existence theorem at precisely the same time.

[2] Wald's contributions (1935, 1936a, 1936b) do more than count equations and unknowns. They are a milestone in the development of a satisfactory general existence theorem but have features that make his analysis quite specialized by contemporary standards.

This celebratory issue of the *JPE* is an ideal place to remind us of the extent to which the origins of modern general equilibrium theory were at the University of Chicago.

I will argue further that the Arrow-Debreu approach has many advantages when compared to the more customary excess demand approach as found in Debreu (1959) and Arrow and Hahn (1971).

II.

From the point of view of this presentation, it is the work of Debreu (1952) that is most central to the Arrow-Debreu existence theorem. In terms of the mathematics, the essence of Arrow-Debreu is an application of Debreu (1952), who did this work as a member of the Cowles Commission at the University of Chicago. However, there is apparently no good record of the origin of the idea for Debreu's study. It is clear that the existence problem was much in the air at Cowles in Chicago in the early 1950s and that the Kakutani fixed-point theorem (1941) was well understood and available to play a role. Arrow and Debreu were in close communication, and each had made definitive contributions to the modern framing of the Walrasian model that are essential to establishing a general equilibrium existence theorem.[3]

III.

Most textbook treatments of modern general equilibrium theory follow Debreu (1959) and approach the existence of equilibrium via the construction of excess demand functions. The Arrow-Debreu approach is fundamentally different and comes with a number of advantages. As I have already suggested, Debreu (1952) provides a mathematical treatment that applies to both the problem of equilibrium existence in the Walrasian model and the problem of existence in Cournot markets in which agents have the power to influence prices. Furthermore, Debreu's theorem pro-

[3] For a particularly important example, see Arrow (1951a), where the treatment of the second welfare theorem reveals boundary issues that must be dealt with in equilibrium existence. Arrow's first appointment as assistant professor was at the University of Chicago, but he was a "time-splitter," notably with responsibilities at Rand and Columbia, where he was a graduate student. In 1949, he left Chicago to accept a position as acting assistant professor at Stanford. Arrow (1951a) was reprinted as a Cowles Commission paper, and it is reasonable to assume that the Chicago environment, and in particular the collaboration with Debreu, were of substantial importance. But a mystery remains: who saw Debreu (1952) as a key to the Arrow-Debreu existence theorem? Yes, it is a natural extension of Nash (1950), but who had the idea of framing it and using it for Walrasian existence? I am unable to find the answer to this question in the writings of Arrow and Debreu.

vides an entry into the Cournot theory that includes the case in which there are externalities in production. As we will see, it also applies to Walrasian economies with significant externalities. It is surely a benefit of the excess demand function approach that the excess demand functions of each agent are independent of the behavior of other agents both in and out of equilibrium. In this situation one simply sums equations to determine the aggregate excess demand whose zeros define equilibrium prices. However, the Arrow-Debreu approach, which directly confronts the fact that the prices faced by agents depend on the choices of other agents, has advantages in terms of the range of behaviors and models that can be covered by an equilibrium existence theorem.

In the formal presentation that follows, I will demonstrate that the Arrow-Debreu approach does not accommodate only externalities and price-dependent preferences. I will also show how, with a slight adjustment, it goes a substantial way toward accommodating preferences that are not orders, as in Mas-Colell (1974). I assume that the reader has some background with the modern literature and is comfortable with the standard textbook approaches to the equilibrium existence theorem via excess demand functions.[4] To simplify matters I emphasize the case of pure exchange. Debreu (1952), with a slight emendation, is a starting point for the formal commentary.

Debreu defined an abstract economy or generalized n-person game Γ by, for each agent, a choice set X_i, a constraint correspondence $A_i : \prod X_j \to X_i$, and a utility function $U_i : \prod X_j \to \mathbb{R}$. One slightly extends this formulation by making the utility of each agent i state dependent where the state includes the position of all agents including one's self. Let $X = \prod X_j$. The utility function of i is $U_i : X_i \times X \to \mathbb{R}$.

One interprets the ith agent's objective as that of choosing for each $x \in X$ a $z_i \in A_i(x)$ that maximizes $U_i(z_i, x)$ subject to $z_i \in A_i(x)$. The vector of actions $x = (x_1, x_2, \ldots, x_n)$ is an equilibrium for Γ if each x_i maximizes $U_i(\cdot, x)$ subject to $z_i \in A_i(x)$ for each i. This is Debreu's extension of Nash (1950) for noncooperative n-person games, but here we have the emendation that preferences vary with one's own position. Debreu writes, "the [constraint set] is actually independent of the ith component of $[x]$, but . . . we find it more convenient to [include the ith component]" (1982, 702). The assumption that U_i depends on the state, including one's own allocation, and then again on one's choice, is similar in spirit, although more substantive. It captures the idea that perspective influences preferences over choices and one's perspective includes one's own position in a state. This idea is particularly congenial to a behavioral economics interpretation.

[4] Debreu (1959, 1982) and Mas-Colell, Whinston, and Green (1995) are particularly useful starting points.

If each X_i is a compact, convex, and nonempty subset of $\mathbb{R}^n$, if each A_i is continuous, nonempty, and convex-valued, and if each U_i is continuous on X and quasi-concave in its first variable, then the above emendation is without mathematical consequence and Debreu (1952) shows that an equilibrium exists. The proof applies the fixed-point theorem of Kakutani (1941) to a best-response correspondence, which is upper hemicontinuous by the Berge theorem (1959).

I turn now to the manner in which the Debreu theorem is applied in the Arrow-Debreu theorem on the existence of equilibrium. An economy E is specified by $n - 1$ consumers, each with consumption sets $Y_i \in \mathbb{R}^l$ and with initial endowments $\omega_i \in \mathbb{R}^l_+$. Let $Y = \prod Y_j$ and $\Delta = \{p \in \mathbb{R}^l_+ : \Sigma p_i = 1\}$ be the set of normalized prices. Each consumer has a real-valued utility function defined on $Y_i \times Y \times \Delta$, so that the utility of a choice depends on one's perspective, which is conditioned on the consumption state and prices. A Walrasian free-disposal equilibrium is a point (y, p) such that $\Sigma y_i \leq \Sigma \omega_i$, $p \cdot y_i = p \cdot \omega_i$ for each i, and y_i maximizes state-dependent utility on the set of points $y_i \in Y$ that satisfy $p \cdot y_i \leq p \cdot \omega_i$.

To prove the existence of equilibrium, Arrow and Debreu associate with E a generalized n-person game in the following manner. The first $n - 1$ agents correspond to the consumers above with choice sets Y_i and the nth agent is a fictitious "market player" who chooses from the price simplex Δ.[5] Thus, the state variable is a pair (y, p). The constraint correspondences for the first $n - 1$ agents are the usual budget correspondences, which depend on p and initial endowments. The constraint correspondence for the nth agent is a constant and in each state allows him to pick any price vector in Δ. The utility function for the first $n - 1$ agents is as specified in the definition of the economy E. Given the state (y, p), the market player chooses $q \in \Delta$ to maximize $q(\Sigma y_i - \Sigma \omega_i)$. Provided that each Y_i is compact and convex, each ω_i is interior to Y_i, and each U_i is continuous and quasi-concave in its first argument, this generalized game will satisfy the sufficient conditions mentioned above for an equilibrium. It is straightforward to verify that this equilibrium is a Walrasian equilibrium for E provided that each consumption is not a local (unconstrained) maximum in its first argument.

One notes that the preceding approach explicitly allows for consumer preferences that depend on the consumption choices of other consumers. It similarly allows for externalities in economies with production, including the case in which the choices of producers affect the utility of consumers. The manner in which government actions, such as optimal taxation, are incorporated into an existence theorem requires a bit more

[5] The "market player" is not to be thought of as an actual agent. It is added in the same spirit as one adds a fictitious player to adjust prices in Walrasian tatonnement.

care, but again, the right way to do this is through the approach of the Debreu theorem on abstract economies.[6]

IV.

I now demonstrate that the Arrow-Debreu approach, minimally extended, sheds light on a beautiful contribution of Mas-Colell (1974), which generalized in a fundamental manner the demand side of equilibrium theory. Rather than defining preferences as a binary relation, usually transitive and complete, Mas-Colell demonstrated that what is really essential to the equilibrium existence theory is the assumption that each consumer knows what he prefers to each possible consumption, with the obvious requirement that he does not prefer any bundle to itself! To understand why this development was very surprising, consider the case of finite-choice spaces with no room for convexity and various other topological requirements that are essential in virtually all approaches to equilibrium theory. For constraint sets with three elements x, y, and z, with x preferred to y, y preferred to z, and z preferred to x, it is not coherent for any choice to be made. Similarly, with a two-element constraint set composed of x and y, with x in the set of bundles preferred to y and y in the set of bundles preferred to x, there can be no coherent choice. But there turns out to be "magic" in the continuity and convexity that are essential to general equilibrium existence, and Mas-Colell employed these to redo the equilibrium existence theorems so that "better than sets" rather than transitive and complete preferences are all that is required for the existence of equilibrium prices.[7]

So how does the Arrow-Debreu method, minimally extended, apply to the existence of equilibrium in these behavioral worlds without ordered preferences? When utility functions are state-dependent and depend on the consumption state of an agent, the preferred sets of each agent play much the same role as the preference correspondences of Mas-Colell: the

[6] There is some important literature that is designed to accommodate such government activities (see, e.g., Sontheimer 1971; Shoven 1974); however, with the benefit of hindsight these efforts demonstrate the benefits of approaches that are not excess demand based. See Gale and Mas-Colell (1975) and Shafer and Sonnenschein (1975, 1976).

[7] In the mid-1960s, I circulated a manuscript that proved that equilibrium existence could be established in the Debreu (1959) model without the requirement of preference transitivity, and this manuscript was widely circulated and discussed with Arrow, Debreu, and McKenzie. I am fortunate that the lengthy delay in the publication of the volume *Preferences, Utility, and Demand* (Chipman et al. 1971) did not have any bad effects on my career. It should also be pointed out that the Mas-Colell advance, which has one replace preference relations with preferred set correspondences, is not only a deep substantive and conceptual breakthrough but unlike my analysis also requires one to abandon the excess demand approach to equilibrium existence. To understand this point, see the appendix of Mas-Colell (1974).

points on the boundaries of these sets, which can be thought of as "behavioral indifference sets," may cross, and it is possible for y to be preferred to x when the base choice is x and x to be preferred to y when the base choice is y. For similar reasons transitivity fails.[8] But again, this emendation of Arrow and Debreu is easy to put forth with the benefit of hindsight, and in particular with Mas-Colell's approach to the representation of preferences firmly in hand.[9]

V.

I have written about one of those moments when the giants who walked the halls of my university were particularly productive. It is the story of Chicago's role in the origins of general equilibrium. The particular moment that I chose, and the subsequent applications that I cited, closely follow my particular interests. But general equilibrium is a very large area, even when one pays particular attention to the existence literature. The work of the early 1950s has been followed by fundamental contributions by Aumann, Scarf, and a long list of others. Chicago and the *JPE* have played their part in these subsequent developments, and I want to cite four papers from the *JPE* in general equilibrium, welfare economics, and consumer choice applicable to general equilibrium that have influenced my own thinking and teaching.

Kenneth J. Arrow, "A Difficulty in the Concept of Social Welfare" (1950). The timing suggests a substantial Cowles-Chicago influence, and this predates the publication of his monumental *Social Choice and Individual Values* (1951b).

Kelvin Lancaster, "A New Approach to Consumer Theory" (1966). This is among the most-cited *JPE* contributions, and perhaps the most cited in general equilibrium broadly interpreted.

David Cass and Menahem Yaari, "A Re-examination of the Pure Consumption Loans Model" (1966). This is a basic work for opening a discus-

[8] This point is made in Duffie and Sonnenschein (1989).

[9] One should note that Mas-Colell's continuity assumption on the preference correspondence is an open graph condition. This is less demanding than the assumption that the state-dependent utility functions U_i are continuous. Shafer and Sonnenschein (1975), motivated by Mas-Colell's work, showed that Debreu (1952) can be strengthened to allow for preference correspondences with open graphs as opposed to continuous utility functions, but the proof is more than emendation of Debreu's arguments. With their result in hand, one can use the Arrow-Debreu method to prove equilibrium existence theorems of the Mas-Colell type. This further supports the argument that Debreu (1952) and Arrow and Debreu (1954) did more than provide rigorous foundations for the theory of value. It also provides a mathematical approach to the existence theorem that is adequate for dealing with a variety of important generalizations. And again, much of this work was done at the University of Chicago.

sion of the general equilibrium welfare theorems in models with infinities.

David Cass and Karl Shell, "Do Sunspots Matter?" (1983). The paper is a milestone in the general equilibrium approach to asset theory.

I conclude with some remarks that contrast the excess demand approach to equilibrium existence with the approach via abstract economies as pioneered by Debreu (1952) and Arrow and Debreu (1954). Because of my work on "anything goes," sometimes referred to as Sonnenschein, Mantel, and Debreu, I have something at stake here, and one might conjecture that I am a fan of the excess demand approach.[10]

This is not at all my view. I am pressed to think of any significant result on equilibrium existence that cannot be achieved as well via the abstract economy approach. The approach via excess demand has the advantage that it sometimes separates the behavior of agents in a descriptively convenient manner: in the absence of externalities, one's excess demand function does not depend on the actions of other agents. But from the point of view of expanding applications, this is also sometimes the weakness of the approach. There are benefits in directly confronting the fact that the prices faced by agents depend on the choices of other agents. Furthermore, perhaps the continued absence of a foundation for using tatonnement to define laws of motion for Walrasian economies renders the excess demand construct less interesting? I regard Scarf (1960, 1973) to be milestones in helping economists to think through important issues of how one might compute the equilibrium prices of an economy. Also, I do not question the relevance of computing Walrasian equilibria; however, it is less than clear that, even for this purpose, excess demand functions are the way to go.

Finally, I do not believe that the "anything goes theorem" suggests the end of economic theory. Rather, it tells us that the empirically relevant restrictions that are generated by theory are likely to depend on specific assumptions regarding the form of preferences, technology, and the distribution of income.

I close by observing that Debreu (1982) follows closely his 1952 approach and the method of Arrow-Debreu. For me, Arrow and Hahn (1971) gives too little appreciation to the Arrow-Debreu method. But this is not the place to quibble with my teachers. In summary, the Arrow-Debreu approach, with Debreu (1952) replaced by the Mas-Colell-inspired strengthening presented in Shafer and Sonnenschein (1975), is a good way to go.

[10] For a summary of "anything goes," see Shafer and Sonnenschein (1982).

References

Arrow, K. J. 1950. "A Difficulty in the Concept of Social Welfare." *J.P.E.* 58 (4): 328–46.

———. 1951a. "An Extension of the Basic Theorems of Classical Welfare Economics." In *Second Berkeley Symposium on Mathematical Statistics and Probability*, 507–32. Berkeley: Univ. California Press.

———. 1951b. *Social Choice and Individual Values*. New York: Wiley.

Arrow, K. J., and G. Debreu. 1954. "Existence of an Equilibrium for a Competitive Economy." *Econometrica* 22:265–90.

Arrow, K. J., and F. H. Hahn. 1971. *General Competitive Analysis*. San Francisco: Holden-Day.

Berge, C. 1959. *Espaces topologiques*. Stuttgart: Macmillan.

Cass, D., and K. Shell. 1983. "Do Sunspots Matter?" *J.P.E.* 91 (2): 193–227.

Cass, D., and M. E. Yaari. 1966. "A Re-examination of the Pure Consumption Loans Model." *J.P.E.* 74 (4): 353–67.

Chipman, J. S., et al., eds. 1971. *Preferences, Utility, and Demand: A Minnesota Symposium*. New York: Harcourt Brace Jovanovich.

Debreu, G. 1952. "A Social Equilibrium Existence Theorem." *Proc. Nat. Acad. Sci.* 38 (10): 866–93.

———. 1959. *Theory of Value: An Axiomatic Analysis of Economic Equilibrium*. Cowles Foundation Monograph no. 17. New Haven, CT: Yale Univ. Press.

———. 1982. "Existence of Competitive Equilibrium." In *Handbook of Mathematical Economics*, vol. 2, edited by K. J. Arrow and M. Intriligator, 697–743. Amsterdam: Elsevier.

Duffie, D., and H. F. Sonnenschein. 1989. "Arrow and General Equilibrium Theory." *J. Econ. Literature* 27 (2): 565–98.

Gale, D., and A. Mas-Colell. 1975. "An Equilibrium Existence Theorem for a General Model without Ordered Preferences." *J. Math. Econ.* 2 (1): 9–15.

Kakutani, S. 1941. "A Generalization of Brouwer's Fixed Point Theorem." *Duke Math. J.* 8 (3): 457–59.

Lancaster, K. J. 1966. "A New Approach to Consumer Theory." *J.P.E.* 74 (2): 132–57.

Mas-Colell, A. 1974. "An Equilibrium Existence Theorem without Complete or Transitive Preferences." *J. Math. Econ.* 1 (3): 237–46.

Mas-Colell, A., M. D. Whinston, and J. R. Green. 1995. *Microeconomic Theory*. Vol. 1. New York: Oxford Univ. Press.

McKenzie, L. W. 1954. "On Equilibrium in Graham's Model of World Trade and Other Competitive Systems." *Econometrica* 22:147–61.

Nash, J. F. 1950. "Equilibrium Points in *n*-Person Games." *Proc. Nat. Acad. Sci. USA* 36 (1): 48–49.

Scarf, H. 1960. "Some Examples of Global Instability of the Competitive Equilibrium." *Internat. Econ. Rev.* 1 (3): 157–72.

———. 1973. *The Computation of Economic Equilibria*. New Haven, CT: Yale Univ. Press.

Shafer, W. J., and H. F. Sonnenschein. 1975. "Equilibrium in Abstract Economies without Ordered Preferences." *J. Math. Econ.* 2 (3): 345–48.

———. 1976. "Equilibrium with Externalities, Commodity Taxation, and Lump Sum Transfers." *Internat. Econ. Rev.* 17:601–11.

———. 1982. "Market Demand and Excess Demand Functions." In *Handbook of Mathematical Economics*, vol. 2, edited by K. J. Arrow and M. Intriligator, 671–93. Amsterdam: Elsevier.

Shoven, J. B. 1974. "On the Computation of Competitive Equilibrium on International Markets with Tariffs." *J. Internat. Econ.* 4 (4): 341–54.
Sontheimer, K. C. 1971. "The Existence of International Trade Equilibrium with Trade Tax-Subsidy Distortions." *Econometrica* 39:1015–35.
Wald, A. 1935. "Über die eindeutige positive Lösbarkeit der neuen Produktionsgleichungen." *Ergebnisse eines mathematischen Kolloquiums* (6): 12–20.
———. 1936a. "Über die Produktionsgleichungen der ökonomischen Wertlehre." *Ergebnisse eines mathematischen Kolloquiums* (7): 1–6.
———. 1936b. "Über einige Gleichungssysteme der mathematischen Ökonomie." *Zeitschrift für Nationalökonomie* 7:637–70. Translated as "On Some Systems of Equations of Mathematical Economics." *Econometrica* 19 (1951): 368–403.

Economic Growth: The Past, the Present, and the Future

Ufuk Akcigit

University of Chicago, National Bureau of Economic Research, and Centre for Economic Policy Research

> Is there some action a government of India could take that would lead the Indian economy to grow like Indonesia's or Egypt's? If so, what, exactly? If not, what is it about the "nature of India" that makes it so? The consequences for human welfare involved in questions like these are simply staggering: Once one starts to think about them, it is hard to think about anything else. (Lucas 1988, 5)

Introduction

These words by the Nobel laureate Chicago economist Robert Lucas Jr. summarize why so many great scholars found it hard to "think about anything else" and spent their careers trying to understand the process of economic growth. Economies are complex systems resulting from the actions of many actors. This complexity makes it challenging, but also infinitely interesting, to understand the determinants of economic growth. What are the roles of human capital, fertility, ideas, basic science, and public policy for growth? These are just some of the important questions that

I would like to thank my colleagues and current and former students for very valuable feedback and discussions. The paper reflects solely my own views.

were posed by many highly influential studies featured in the issues of the *Journal of Political Economy* over the years. Indeed, this journal has been the platform to diffuse many of the brilliant ideas and start important debates in the field of economic growth. In this short paper, my goal is to revisit some of those seminal papers, briefly describe some of the more recent contributions, and end with some thoughts about the future direction of the field. The reader should note in advance that the list of work covered here is by no means exhaustive and mostly targets work that has been featured in issues of the *JPE*.

The Past

In this section, I discuss some of the past seminal contributions to the field of economic growth. The papers will be grouped under the following subtopics: (i) capital accumulation, (ii) innovation, (iii) technology adoption, and (iv) human capital and fertility.

Physical Capital Accumulation, Spillovers, and Growth

For a long time, the workhorse to study economic growth was the neoclassical growth model (Ramsey 1928; Cass 1965; Koopmans 1965). The main difference from the classic Solow-Swan model (Solow 1956; Swan 1956) is that the neoclassical model explicitly endogenizes the savings rate through a utility-maximizing household. One of the key components of both models is the production function

$$Y_t = AK_t^{\alpha} L_t^{1-\alpha}, \tag{1}$$

where Y_t is aggregate output at time t, K_t is the capital stock, L_t is the labor, A is the level of productivity, and $\alpha \in (0, 1)$.[1] A major implication of the neoclassical and the Solow-Swan models is that capital accumulation can serve as a source of short-run economic growth but cannot be a driver of long-run growth because of decreasing returns through $\alpha < 1$. These models predict "convergence in per capita income," whereby poorer countries grow faster until they catch up with richer countries.

The *JPE* featured many influential empirical and theoretical works on economic growth. On the empirical side, Barro and Sala-i-Martin (1992) tested the convergence result of neoclassical theory. In this work, Barro and Sala-i-Martin studied the growth rates of 48 US states between 1840 and 1963 as a function of their initial per capita income and found significant empirical evidence for convergence at the state level. Barro and Sala-i-

[1] This is the particular specification used to facilitate the discussion in this paper. This production function can be written in a more general constant returns form.

Martin also analyzed the convergence hypothesis in the cross-country data. This time their results were more mixed. The authors identified convergence only after controlling for school enrollment rate and the ratio of government consumption to GDP. They thus provide some evidence for "conditional convergence" at the country level.

A production function with decreasing returns was the main reason why endogenous growth did not occur in the neoclassical growth model. In one of the seminal *JPE* papers, Romer (1986) overcame the problem of decreasing returns and generated endogenous long-run growth by introducing spillovers that led to increasing returns in the production function. In this model, growth resulted from the combination of capital accumulation and the associated spillovers. More specifically, Romer assumed productivity to be a linear function of the capital stock:

$$A_t = \gamma K_t.$$

In this model, just like the neoclassical model, markets are perfectly competitive and knowledge creation is simply a by-product of capital accumulation. Therefore, one can also interpret productivity formation as coming from learning by doing. Stokey (1988) was another great *JPE* paper that showed how economywide learning by doing could lead to sustained long-run growth. It was not until Romer's later work, which I will describe below, that agents in the economy had explicit incentives to create new ideas. While Romer (1986) introduced productivity spillovers through capital accumulation, Lucas (1988) introduced similar spillovers through human capital externalities. In their interesting *JPE* paper, Glaeser et al. (1992) empirically study the existence of knowledge spillovers within and across industries.

On the theory side, Rebelo (1991) is one of the well-known *JPE* papers that generated endogenous long-run growth by eliminating the decreasing returns from the neoclassical production function (1). This paper considered the so-called "AK model," in which the production technology does not feature labor and is linear in capital with $\alpha = 1$:

$$Y_t = AK_t.$$

In this model, the linear structure prevents capital accumulation from running into decreasing returns and can generate long-run growth. This tractable framework allowed Rebelo to also study the impact of public policy on economic growth, which was not possible in the Ramsey-Cass-Koopmans or Solow-Swan models since they did not generate long-run endogenous growth.

It is now widely accepted in the literature that the world economy has experienced persistent technological progress over the past 200 years and

R&D and innovation have played a central role in the advancement of the world technology frontier during this period (Acemoglu 2008). Unfortunately, endogenous growth models based on the neoclassical framework were insufficient and had little to say about this aspect of economic growth. Therefore, the early 1990s saw the rise of innovation-based growth models, as I explain next.

Innovation-Based Growth

New technologies emerge as a result of costly R&D efforts by individuals and companies. The new technologies eventually introduce into the market a new product variety as in Romer (1990) or a better version of an existing product or technology that makes the earlier version obsolete through Schumpeterian creative destruction as in Aghion and Howitt (1992) and Grossman and Helpman (1991). An entrepreneur's or firm's incentive to undertake these costly R&D efforts is to gain market power. These explicit efforts and market incentives were missing in the endogenous growth models based on a neoclassical framework.

The new ingredient in the innovation-based endogenous growth models is the production function for ideas. The number of new ideas—the change in productivity $\dot{A}$—is assumed to be a function of the existing knowledge stock A_t and the number of researchers R who spent time producing them:

$$\dot{A}_t = \delta A_t R, \tag{2}$$

where $\delta > 0$ captures the research productivity. In these models, agents face an occupational choice. Individuals can work either as production workers (L) and earn the production wage (w_t) or as research workers (R) who produce new ideas and receive the return to their innovation (V_t). The key equation is the free-entry condition into research, which determines the allocation of the work force to the production and research sectors:

$$w_t = V_t \delta A_t.$$

This equilibrium split of the workforce determines the current level of production through L and the rate of growth of knowledge (and hence per capita income) through R.

Romer's model views each innovation as the introduction of a new product variety that becomes a permanent part of the economy and the inventor of which becomes its permanent producer. Thus, the model abstracts from competition, firm exit, and firm turnover. Schumpeterian models, on the other hand, prioritized the industrial organization aspect of economic growth. Through the notion of creative destruction, Schumpeterian

models introduced firm entry-exit and competition into the endogenous growth literature (e.g., Aghion and Howitt 1992; Grossman and Helpman 1991). This feature has been essential to map these models to the firm- and inventor-level microdata and estimate them, as I explain below. Moreover, these features enable Schumpeterian models to generate richer policy implications. Since the Romer-style product variety models feature uninternalized spillovers through (2), they generate major underinvestment in R&D and call for research subsidies. Schumpeterian models, on the other hand, brought in a competitive force through the so-called "business stealing" effect whereby new entrants try to replace incumbents through new innovations. If the spillover associated with each innovation is not big enough, the social return from it would be less than the private return, which, in equilibrium, could make firms overinvest in R&D. Therefore, Schumpeterian models call for empirical estimates of the spillovers and business stealing externality to inform optimal policy.

In the innovation-based growth models, the R&D production function in (2) has been taken as a reduced-form representation, in which some inputs (either human capital or R&D dollars) turn into innovations. However, a few papers have studied this production function in detail and tried to understand the steps that lead to innovation. One of the key findings of this literature has been that there are at least two steps in creating a practical innovation. First, universities, public research labs, and occasionally private firms invest in theory-oriented and abstract "basic research" to produce fundamental, essential, first-stage background knowledge in the form of theories and equations. In the second step, profit-seeking innovators and companies invest in more familiar, data-oriented or end product–oriented "applied research" to produce practical and patentable findings. For instance, Wallace Carothers's basic research led to the invention of the famous Carothers equation that formed the basis for the applied research by DuPont that resulted in the invention of "nylon." What are the incentives to do basic research? What are the roles of government and universities in innovation and economic growth? These were the key questions that Richard Nelson investigated in his highly influential *JPE* article (Nelson 1959). He questioned the sources of spillovers from basic research and discussed the reasons why some firms would invest more in basic research that is more uncertain, with less clear goals or less close ties to specific practical problems or to the creation of a specific object. Nelson's answer to this question was that successful scientific advances through basic research often have many practical applications that are not predictable ex ante. These scientific findings form the key inputs for many subsequent applied research projects, which can then lead to practical and patentable findings. Because of this nature of basic research, firms are unlikely to utilize all the economic value through patents. Further, a research finding

that has applications in one field or sector might have applications in many others. He noted that "it is for this reason that firms which support research toward the basic science end of the spectrum are firms that have fingers in many pies" (302). Even though this is a fundamental problem for innovation policy, the split between basic and applied research and the explicit role of the universities in the growth process are still understudied.[2]

Relatedly, in his well-known *JPE* article, Chad Jones (1995) questioned the R&D production function (2). The basis for Jones's critique was that one of the key implications of the endogenous growth models that used (2) is that increased population size is associated with an increased steady-state growth rate. This was due to the fact that a larger population raises the return to innovation through the so-called "market size effect" and also increases the supply of potential researchers to do R&D. Jones argued that these predictions were empirically not plausible since larger countries do not necessarily grow faster, and even though the United States increased its R&D effort over the years, its growth rate did not go up. He proposed a "semi-endogenous" growth model by modifying the R&D production function of (2) as follows:

$$\dot{A} = \delta A^{\phi} R^{\lambda},$$

where $\phi, \lambda \in (0, 1)$. This modification removed the impact of the population level and led to the following long-run growth rate: $g = \lambda n/(1 - \phi)$, where n is simply the rate of population growth. A strong prediction of Jones's specification is that long-run growth is affected only by the exogenous population growth rate n and is invariant to any government policy. This paper started the "scale effect" debate in the literature.

During the 1990s, the *JPE* was the stage for the "scale effect" debate. Soon after Jones's influential paper, the *JPE* published two other interesting articles by Alwyn Young (1998) and Peter Howitt (1999), who proposed "product proliferation" as a remedy to the "scale effect" problem. Howitt's solution, which builds on Young's proliferation idea, is to propose a mechanism whereby, as the number of product varieties in an economy grows, the effectiveness of research effort on each variety decreases as the population gets spread more thinly over a larger number of varieties. This makes the expected growth rate in each variety independent of the overall population level while preserving the role of policy for economic growth.

[2] A notable exception is the recent work by Akcigit, Hanley, and Serrano-Velarde (2016), which shows that even after controlling for firm size, firms that operate in more industries are more likely to invest in basic research. This finding provides empirical support for Nelson's "fingers in many pies" hypothesis of basic research.

Technology Adoption and Growth

While countries at the world technology frontier grow through innovations, nonfrontier countries grow mostly through imitation or technology adoption. Why do some countries adopt new technologies while others do not? In their seminal *JPE* paper, Murphy, Shleifer, and Vishny (1989) study this question. They theoretically formalize the famous "big-push" hypothesis of industrialization of Rosenstein-Rodan (1943). In their analysis, there is an underlying market size effect that works through the complementarities between different sectors in the economy. When certain firms or industries invest in new technologies, these investments also raise demand in other industries. The investing firm receives only a fraction of their contribution to the profits in the overall economy. Therefore, each individual firm might not find it profitable to adopt a new technology, whereas a coordinated investment by all firms could justify such a technology adoption. As a result, in those economies where actions can be coordinated through public policies, firms might be willing to invest in new technologies and can generate a big push that leads to industrialization.

In another well-known *JPE* paper, Stephen Parente and Edward Prescott (1994) studied the macroeconomic implications of barriers to technology adoption. One of the major shortcomings of the neoclassical growth model is its inability to generate empirically plausible income differences across countries. Parente and Prescott fixed this problem by extending the neoclassical growth model to incorporate firm investment for technology adoption from the world frontier. Country-specific barriers to technology adoption, due to weak property rights or other institutional aspects, could hinder the flow of technologies, leading to the observed technology and income gaps across countries.

Relatedly, in another important *JPE* paper, Katz and Shapiro (1986) studied the importance of network externalities in technology adoption. This paper showed that technology adoption could depend on whether a technology is "sponsored," that is, whether an entity owns the right to that technology and therefore has the incentives to promote it. An important aspect in this analysis is the dynamic consideration of technology adoption. For instance, this paper highlights the concept of "penetration pricing," a strategy of offering low prices today to build up a network and influence the expectations of the customer about the future size of the network. In this case, it is possible that an inferior technology might get adopted because it is sponsored by an entity that strategically affects the size of the network.

Empirical studies on technology adoption have been relatively rare. The influential *JPE* article by Foster and Rosenzweig (1995) studied the "Green Revolution" in India, during which farmers in some Indian regions received an opportunity to adopt new technologies. The paper finds that in-

formational frictions and the lack of knowledge about how to use the new seeds were major impediments to technology adoption. It also identifies significant and quantitatively important learning spillovers; that is, farmers learn from neighbors who already use the new seeds.

Human Capital, Fertility, and Growth

A large literature has studied the link between fertility and economic growth. The Malthusian model (Malthus 1798) is a well-known framework according to which population grows faster once per capita income rises. However, this prediction is at odds with the empirical evidence that fertility has gone down as income grew around the world, especially in Western economies. Neoclassical growth models circumvented this contradiction by focusing on endogenous capital accumulation rather than on the fertility choice. Nobel laureate and Chicago economist Gary Becker (1960) formulated the quantity-quality trade-off that parents face when deciding on the number of children and the investment in each child's human capital. Subsequently, Becker and Barro (1988) and Barro and Becker (1989) introduced fertility choice into growth models. The seminal *JPE* paper by Becker, Murphy, and Tamura (1990) is the first paper that introduced fertility choice into an "endogenous growth" framework. The model implies that the rate of return on human capital relative to those on children is high when human capital in the society is abundant. Therefore, the model predicts large families and little investment in each child when the society has limited human capital.

The Present

The early models of endogenous growth set the stage for a fruitful literature to come, and the *JPE* was their leading outlet. What was missing in the early literature was firm- or individual-level heterogeneity that would provide a better connection between growth theory and micro-level data. Providing solid micro-level foundations would lead to stronger macro-growth models that generate more accurate positive predictions and relevant normative implications.

A major step in that direction was another important *JPE* paper by Klette and Kortum (2004), which built a novel Schumpeterian growth model. The earlier endogenous growth models predict that innovations come only from new entrants, new entrants and young firms are the largest firms in the economy, and exit rates would be uncorrelated with firm age or size. While these predictions are at odds with the firm-level data, the Klette-Kortum model fixed these problems by defining a firm as a collection of production units. In this framework, firms could grow by introducing better-

quality versions of the products of other firms and thus add those products to their portfolio. Klette and Kortum could thus map the benchmark Schumpeterian models to a more realistic firm dynamics setting. In a more recent *JPE* article, Akcigit and Kerr (forthcoming) extended the Klette-Kortum framework by allowing firms to improve not only other firms' products through "external innovations" but also their own products through "internal innovations." Innovation and firm size heterogeneities allow the model to generate a close fit to the firm- and innovation-level data from the US Census Bureau and the US Patent Office. The estimated model shows that small firms are expending disproportionately more effort to do external innovations, and the spillovers associated with external innovations are significantly larger. These findings suggest that R&D policies that aim to correct for underinvestment in R&D should take into account the differential spillovers generated by different-sized firms in the economy. As it is exemplified by these papers, a close dialogue between endogenous growth theory with heterogeneity and microdata allows researchers to quantify certain mechanisms and study the impact of counterfactual industrial policies.

Heterogeneity across individuals has also been the key feature of two recent *JPE* papers. Jaimovich and Rebelo (2017) showed that taxation has a nonlinear impact on growth once talent heterogeneity is taken into account. Jones and Kim (forthcoming) showed that focusing on differential innovation efforts by entrants and incumbents can be important in understanding the rise in top income inequality.

Heterogeneity in ideas has been the essential component of the surging endogenous growth models that incorporate the importance of human interactions for human capital accumulation and growth (e.g., Alvarez, Buera, and Lucas 2008; Lucas 2009). The *JPE* recently published two papers, Lucas and Moll (2014) and Perla and Tonetti (2014), that show that growth can also occur when less knowledgeable agents in an economy interact with more knowledgeable agents and imitate them to improve their own productivities.

The *JPE* continues to promote frontier research on various additional aspects of economic growth. Trade and innovation (e.g., Atkeson and Burstein 2010), growth miracles (e.g., Young 2012), value of life (e.g., Jones 2016), environment (e.g., Acemoglu et al. 2016; Aghion et al. 2016), and institutions (e.g., Acemoglu, Robinson, and Verdier 2017) are just a few of the exciting facets of economic growth featured in recent issues of the *JPE*.

The Future

The future of the field of economic growth is more exciting than ever. Computers are becoming more powerful. Many countries are making their

firm- and individual-level microdata sets available to researchers. In addition, thanks to optical character recognition techniques, more and more large-scale historical records are being digitized to be used in economic research.[3] My recommendation to young researchers is to take notice of the rapid change in the field of economic growth and invest not only in their theoretical skills but also in their computational and empirical knowledge. I also recommend reading the seminal papers—some of which I described here. They are full of stimulating ideas, many of which could not be investigated empirically because of a lack of data or developed quantitatively because of a lack of computational power. Now could be the time to revisit those ideas.

The University of Chicago's *JPE* has been the host of many seminal contributions, and it will, without any doubt, continue to promote the frontier research in the field. Have we answered the questions that Robert Lucas posed in the opening quote? Certainly not entirely. Yet the field has made significant progress toward that goal over the years. This is a call for young researchers to come and think harder about growth-related issues. The field has so many interesting open questions that they will surely, once they start, have a hard time "thinking about anything else."

References

Acemoglu, D. 2008. *Introduction to Modern Economic Growth*. Princeton, NJ: Princeton Univ. Press.

Acemoglu, D., U. Akcigit, D. Hanley, and W. Kerr. 2016. "Transition to Clean Technology." *J.P.E.* 124 (1): 52–104.

Acemoglu, D., J. Robinson, and T. Verdier. 2017. "Asymmetric Growth and Institutions in an Interdependent World." *J.P.E.* 125 (5): 1245–1305.

Aghion, P., A. Dechezleprêtre, D. Hémous, R. Martin, and J. Van Reenen. 2016. "Carbon Taxes, Path Dependency, and Directed Technical Change: Evidence from the Auto Industry." *J.P.E.* 124 (1): 1–51.

Aghion, P., and P. Howitt. 1992. "A Model of Growth through Creative Destruction." *Econometrica* 60 (2): 323–51.

Akcigit, U., J. Grigsby, and T. Nicholas. 2017. "The Rise of American Ingenuity: Innovation and Inventors of the Golden Age." Working Paper no. 23047, NBER, Cambridge, MA.

Akcigit, U., D. Hanley, and N. Serrano-Velarde. 2016. "Back to Basics: Basic Research Spillovers, Innovation Policy and Growth." Discussion Paper no. 11707, Centre Econ. Policy Res., London.

Akcigit, U., and W. R. Kerr. Forthcoming. "Growth through Heterogeneous Innovations." *J.P.E.*

Alvarez, F. E., F. J. Buera, and R. E. Lucas Jr. 2008. "Models of Idea Flows." Working Paper no. 14135, NBER, Cambridge, MA.

[3] For instance, the digitized historical patent files matched to census records by Akcigit, Grigsby, and Nicholas (2017) could provide a unique opportunity for researchers to shed light on Golden Age innovators, inventions, and economic growth more broadly.

Atkeson, A., and A. T. Burstein. 2010. "Innovation, Firm Dynamics, and International Trade." *J.P.E.* 118 (3): 433–84.
Barro, R. J., and G. S. Becker. 1989. "Fertility Choice in a Model of Economic Growth." *Econometrica* 57 (2): 481–501.
Barro, R. J., and X. Sala-i-Martin. 1992. "Convergence." *J.P.E.* 100 (2): 223–51.
Becker, G. S. 1960. "An Economic Analysis of Fertility." In *Demographic and Economic Change in Developed Countries*, 209–40. New York: Columbia Univ. Press (for NBER).
Becker, G. S., and R. J. Barro. 1988. "A Reformulation of the Economic Theory of Fertility." *Q.J.E.* 103 (1): 1–25.
Becker, G. S., K. M. Murphy, and R. Tamura. 1990. "Human Capital, Fertility, and Economic Growth." *J.P.E.* 98, no. 5, pt. 2 (October): S12–S37.
Cass, D. 1965. "Optimum Growth in an Aggregative Model of Capital Accumulation." *Rev. Econ. Studies* 32 (3): 233–40.
Foster, A. D., and M. R. Rosenzweig. 1995. "Learning by Doing and Learning from Others: Human Capital and Technical Change in Agriculture." *J.P.E.* 103 (6): 1176–1209.
Glaeser, E., H. D. Kallal, J. A. Scheinkman, and A. Shleifer. 1992. "Growth in Cities." *J.P.E.* 100 (6): 1126–52.
Grossman, G., and E. Helpman. 1991. *Innovation and Growth in the World Economy*. Cambridge, MA: MIT Press.
Howitt, P. 1999. "Steady Endogenous Growth with Population and R&D Inputs Growing." *J.P.E.* 107 (4): 715–30.
Jaimovich, N., and S. Rebelo. 2017. "Nonlinear Effects of Taxation on Growth." *J.P.E.* 125 (1): 265–91.
Jones, C. I. 1995. "R&D-Based Models of Economic Growth." *J.P.E.* 103 (4): 759–84.
———. 2016. "Life and Growth." *J.P.E.* 124 (2): 539–78.
Jones, C. I., and J. Kim. Forthcoming. "A Schumpeterian Model of Top Income Inequality." *J.P.E.*
Katz, M. L., and C. Shapiro. 1986. "Technology Adoption in the Presence of Network Externalities." *J.P.E.* 94 (4): 822–41.
Klette, T. J., and S. Kortum. 2004. "Innovating Firms and Aggregate Innovation." *J.P.E.* 112 (5): 986–1018.
Koopmans, T. C. 1965. "On the Concept of Optimal Economic Growth." In *The Econometric Approach to Development Planning*, 225–300. Amsterdam: North-Holland (for Pontificia Acad. Sci.).
Lucas, R. E., Jr. 1988. "On the Mechanics of Economic Development." *J. Monetary Econ.* 22 (1): 3–42.
———. 2009. "Ideas and Growth." *Economica* 76 (301): 1–19.
Lucas, R. E., Jr., and B. Moll. 2014. "Knowledge Growth and the Allocation of Time." *J.P.E.* 122 (1): 1–51.
Malthus, T. R. 1798. *An Essay on the Principle of Population*. London: Lawbook Exchange.
Murphy, K. M., A. Shleifer, and R. W. Vishny. 1989. "Industrialization and the Big Push." *J.P.E.* 97 (5): 1003–26.
Nelson, R. R. 1959. "The Simple Economics of Basic Scientific Research." *J.P.E.* 67 (3): 297–306.
Parente, S. L., and E. C. Prescott. 1994. "Barriers to Technology Adoption and Development." *J.P.E.* 102 (2): 298–321.
Perla, J., and C. Tonetti. 2014. "Equilibrium Imitation and Growth." *J.P.E.* 122 (1): 52–76.
Ramsey, F. P. 1928. "A Mathematical Theory of Saving." *Econ. J.* 38 (152): 543–59.

Rebelo, S. 1991. "Long-Run Policy Analysis and Long-Run Growth." *J.P.E.* 99 (3): 500–521.
Romer, P. M. 1986. "Increasing Returns and Long-Run Growth." *J.P.E.* 94 (5): 1002–37.
———. 1990. "Endogenous Technological Change." *J.P.E.* 98, no. 5, pt. 2 (October): S71–S102.
Rosenstein-Rodan, P. N. 1943. "Problems of Industrialisation of Eastern and South-Eastern Europe." *Econ. J.* 53 (210/211): 202–11.
Solow, R. M. 1956. "A Contribution to the Theory of Economic Growth." *Q.J.E.* 70 (1): 65–94.
Stokey, N. L. 1988. "Learning by Doing and the Introduction of New Goods." *J.P.E.* 96 (4): 701–17.
Swan, T. W. 1956. "Economic Growth and Capital Accumulation." *Econ. Record* 32 (2): 334–61.
Young, A. 1998. "Growth without Scale Effects." *J.P.E.* 106 (1): 41–63.
———. 2012. "The African Growth Miracle." *J.P.E.* 120 (4): 696–739.

Economic History

David W. Galenson

University of Chicago and Universidad del CEMA

The field of economic history was largely transformed during the 1960s, when scholars trained in economics departments began to apply economic theory and econometrics to historical questions. The resulting quantitative research, often called the new economic history or cliometrics, has now produced significant revisions of some issues that had previously been studied primarily by historians, as well as novel results on long-run relationships of interest to economists. Much of this research has been based on computer analysis of micro-level data sets collected from historical archives.

The first major debate within the new field was initiated by a 1958 article by Alfred Conrad and John Meyer, "The Economics of Slavery in the Antebellum South." The authors' conclusion that slavery was profitable in the nineteenth century directly contradicted the contention of some historians that the Civil War was not necessary to eliminate slavery, because the institution would have disappeared even in the absence of legal intervention. Few issues in American history are as contentious as slavery, and this initial paper prompted dozens of empirical investigations of the de-

mography and economics of the slave economy. The most comprehensive of these was *Time on the Cross*, published by Robert Fogel and Stanley Engerman in 1974, which used micro-level evidence from the federal census of 1860 to establish that southern farms were more efficient than their northern counterparts and that there were significant economies of scale within southern agriculture. Further analysis of these scale effects led to identification of the most important source of slavery's economic efficiency (Fogel and Engerman 1977). From the finding that scale economies appeared above a threshold plantation size of 15 slaves—which typically implied the presence of 10 adult slaves, or enough to form one slave gang—Fogel and Engerman inferred that it was the gang system that made slavery particularly efficient. The economic advantage of slave plantations was greatest for crops that were physically suited to rapid and routinized cultivation and least for more delicate crops that were damaged by rapid handling: sugar, rice, and cotton could be produced efficiently by gangs, whereas tobacco and grain could not. Slave gangs were equivalent to assembly lines in the fields, as the labor of cultivation could be clearly divided among the members of a gang, and the pace of work could be controlled by a driver who could readily monitor each slave's work. This research not only provided a more precise understanding of the economics of slavery but also underscored the magnitude of the achievement of the abolitionists, whose moral crusade triumphed over a wealthy and powerful slave economy, not a weak and failing one.

Quantitative investigations have also produced a new understanding of the colonial economy and the origins of slavery in North America. Seventeenth-century English colonizers had knowledge only of an Old World in which land was scarce and expensive, and labor abundant and cheap, and they consequently anticipated that ownership of vast amounts of land would yield vast wealth. Instead, they were shocked to experience a veritable social revolution, as the reversal of factor scarcity produced radically higher economic and social mobility. Colonizers now had to recognize that the greatest economic problem in the New World was to recruit and control a supply of labor to exploit the abundant land (Galenson 1996). The planters' initial response was to bring to America the young men and women who worked on English farms. Indentured servitude allowed these workers to afford the high cost of transportation: prospective servants signed contracts in England promising to work for colonial planters in return for passage to the colonies and maintenance there during their terms. These terms were necessarily longer than the annual contracts customary in England, as servants were bound for at least 4 years and often 7 or 8. More productive servants could repay the fixed cost of passage more quickly, and econometric analysis of large collections of individual contracts reveals a strong inverse relationship between term length and expected individual productivity, as older servants and skilled craftsmen received shorter terms.

The contracts also show compensating differentials for less desirable destinations, as servants willing to travel to the West Indies, where mortality rates were higher and opportunities after servitude worse, received shorter terms than those bound for mainland colonies (Galenson 1981).

Over time rising English wages raised the cost of servants to colonial planters, and falling world sugar prices concurrently lowered the price of African slaves brought to the Americas. These trends together produced a dramatic shift in colonial labor forces: in the last quarter of the seventeenth century, for example, the price of servants relative to slaves recorded in Maryland probate inventories increased by more than 50 percent, and the inventories' holdings of bound labor fell from four servants per slave to nearly the reverse, more than three slaves per servant. The rise of black slavery in North America was thus not caused by any preference for slaves, but by the rising cost of white labor (Menard 1977).

The settlement of the nineteenth-century Midwest has been an active topic of historical research ever since Frederick Jackson Turner's celebrated 1896 speech on the role of the frontier in American history. Quantitative studies have confirmed Turner's contention that the frontier was a place of equality and opportunity for both the native-born and immigrants and have added to our knowledge of the mechanisms that caused this. Studies of persistence—done by tracing residents enumerated in a community in the manuscript schedules of successive decennial censuses—have shown that settlers did not typically settle for long: few midwestern counties or cities had 10-year persistence rates as high as 50 percent (Curti 1959; Thernstrom 1973). Stage migration to the frontier, with a series of short moves, was most common, so the farther a settler was from his place of birth, the older he tended to be (Bogue 1963). These investments in migration produced positive returns: one study found that Utah families who had changed residence in the preceding decade had, on average, lower wealth than families otherwise alike who had not migrated, but higher incomes; the foreign-born similarly had lower wealth, but higher incomes, than their native-born counterparts. The timing of migration was important: early arrival in a new frontier community had a substantial positive effect on a household's income and wealth. Although inequality was low in early frontier communities, concentration of both income and wealth appears to have increased steadily with duration of settlement (Pope 1989).

Within the past two decades, new theories of economic development have been based on historical evidence. Stanley Engerman and Kenneth Sokoloff (2012) observed that the economic history of European colonization in the Americas posed the puzzle that the wealthiest areas of settlement in the colonial era, in the Caribbean and Latin America, fared worse in the long run than the North American regions that were initially marginal economically. They argued that the key difference lay in the regions' initial factor endowments, which led to major differences in the structure

of their economies and consequently in inequality. The ability of the southern regions to produce staple crops on large plantations led to great economic inequality. The wealthy elites that dominated these economies created political institutions that limited land ownership and education, and these deficiencies subsequently hampered modern economic growth. In contrast, the predominance of family farms in the northern regions led to much lower inequality and, consequently, to the development of political institutions that fostered policies toward education, land ownership, and immigration that proved much more favorable to economic growth in the modern era. Daron Acemoglu and James Robinson (2012) have based an even wider-ranging theory of development on extensive analysis of historical evidence, arguing that political institutions are the key determinant of the long-run economic success or failure of nations. In their scheme, *extractive* institutions allow narrow elites to capture a society's resources for their own benefit, whereas *inclusive* institutions spread economic benefits more widely. In the long run, extractive institutions deter innovation and economic growth, whereas inclusive institutions foster education and technology, and consequently growth.

There is widespread agreement among economists who study economic growth that technological change is its most important long-run source. And economists have recognized that talented individuals account for most innovations (Arrow 1962). In view of this, Simon Kuznets (1962) declared that "we need far more empirical study than we have had so far of the universe of inventors" (32). Historical studies have now responded to this appeal, with new findings about innovators.

A major revision concerns the life cycle of human creativity. A long-standing belief has been that creativity is predominantly a prerogative of youth. Yet this belief is wrong. One recent study of the careers of nearly 3,000 physicists concluded that the timing of a scientist's most important publications was randomly distributed within the scientist's body of work (Sinatra et al. 2016). And analysis of the careers of innovators in a wide variety of activities has shown why this is so. In nearly every intellectual activity, there are two distinctly different types of creativity, each of which is associated with a very different pattern of discovery over a career. *Conceptual* innovators make sudden breakthroughs by formulating new ideas, creating syntheses of concepts that had not previously been considered to be related. The most radical of these new ideas usually occur early in innovators' careers, when they are least constrained by acquired habits of thought. In contrast, *experimental* innovators work inductively and gradually. Their greatest contributions generally arrive after long periods of research, when they have accumulated great knowledge of their subject. Among important conceptual innovators in the modern era, Albert Einstein, Herman Melville, Pablo Picasso, T. S. Eliot, Orson Welles, Andy War-

hol, and Bob Dylan all made their landmark contributions between the ages of 24 and 36, whereas the great experimental innovators Charles Darwin, Mark Twain, Paul Cézanne, Frank Lloyd Wright, Robert Frost, Irving Berlin, and Alfred Hitchcock made their greatest contributions between the ages of 48 and 76 (Galenson 2006).

The field of economic history is extremely broad in its coverage of both time and space, and scores of additional areas of research could be added to those summarized briefly above. Yet these serve as significant examples of the gains that have resulted from the systematic application of economic theory and econometrics to large bodies of historical evidence, both in the precision of our knowledge of the past and in the confidence we can attach to generalizations about economic change.

References

Acemoglu, Daron, and James Robinson. 2012. *Why Nations Fail: The Origins of Power, Prosperity, and Poverty*. New York: Crown Bus.

Arrow, Kenneth. 1962. "Economic Welfare and the Allocation of Resources for Innovation." In *The Rate and Direction of Inventive Activity: Economic and Social Factors*, edited by Richard Nelson, 609–25. Chicago: Univ. Chicago Press.

Bogue, Allan. 1963. *From Prairie to Cornbelt: Farming on the Illinois and Iowa Prairies in the Nineteenth Century*. Chicago: Univ. Chicago Press.

Conrad, Alfred, and John Meyer. 1958. "The Economics of Slavery in the Antebellum South." *J.P.E.* 66 (2): 95–130.

Curti, Merle. 1959. *The Making of an American Community: A Case Study of Democracy in a Frontier County*. Stanford, CA: Stanford Univ. Press.

Engerman, Stanley, and Kenneth Sokoloff. 2012. *Economic Development in the Americas since 1500: Endowments and Institutions*. Cambridge: Cambridge Univ. Press.

Fogel, Robert, and Stanley Engerman. 1974. *Time on the Cross: The Economics of American Negro Slavery*. 2 vols. Boston: Little, Brown.

———. 1977. "Explaining the Relative Efficiency of Slave Agriculture in the Antebellum South." *A.E.R.* 67 (3): 275–96.

Galenson, David. 1981. "The Market Evaluation of Human Capital: The Case of Indentured Servitude." *J.P.E.* 89 (3): 446–67.

———. 1996. "The Settlement and Growth of the Colonies: Population, Labor, and Economic Development." In *The Cambridge Economic History of the United States*, vol. 1, edited by Stanley Engerman and Robert Gallman, 135–207. Cambridge: Cambridge Univ. Press.

———. 2006. *Old Masters and Young Geniuses: The Two Life Cycles of Artistic Creativity*. Princeton, NJ: Princeton Univ. Press.

Kuznets, Simon. 1962. "Inventive Activity: Problems of Definition and Measurement." In *The Rate and Direction of Inventive Activity: Economic and Social Factors*, edited by Richard Nelson, 19–51. Chicago: Univ. Chicago Press.

Menard, Russell. 1977. "From Servants to Slaves: The Transformation of the Chesapeake Labor System." *Southern Studies* 16 (4): 355–90.

Pope, Clayne. 1989. "Households on the American Frontier: The Distribution of Income and Wealth in Utah, 1850–1900." In *Markets in History*, edited by David Galenson, 148–89. Cambridge: Cambridge Univ. Press.

Sinatra, Roberta, et al. 2016. "Quantifying the Evolution of Individual Scientific Impact." *Science* 354 (6312): aaf5239-1–8.
Thernstrom, Stephan. 1973. *The Other Bostonians: Poverty and Progress in the American Metropolis, 1880–1970.* Cambridge, MA: Harvard Univ. Press.

Political Economics in the *Journal of Political Economy*: Six Landmark Papers

Roger Myerson

University of Chicago

When the *Journal of Political Economy* was founded, the term "political economy" was commonly used to denote the academic field that we now call "economics," where most scholars focus more on the study of markets than on the study of politics and government. But the *JPE* has published important papers that apply the theoretical and empirical methodologies of modern economics to the study of politics and political institutions, a subfield which may be called "political economics."

Duncan Black's 1948 *JPE* paper "On the Rationale of Group Decision-Making" was written with the explicit goal of providing a basis for the development of a pure science of politics. Black analyzed the fundamental problems of public decision making by considering the problems of voting in a committee that must choose among some given set of alternatives. He found that, if the alternatives are points on a line and each voter prefers alternatives that are closer to his or her ideal point, then the median of the voters' ideal points can be identified as the choice that is preferred by a majority over any other alternative. In other cases in which this one-dimensional structure is lacking, however, Black recognized that the committee's decision could depend on the order in which the alternatives are considered. That is, in the general case, the outcome of social decision making can depend on details about who gets to set the agenda.

These fundamental problems of social choice theory were revisited in Kenneth Arrow's 1950 *JPE* paper on "A Difficulty in the Concept of Social Welfare," which proved an early version of Arrow's famous possibility theorem. In this paper, Arrow listed some basic consistency conditions for how a social preference ordering should depend on individual preferences, and he showed that if there are three or more alternatives for so-

ciety to rank, then collective decision making cannot satisfy these consistency conditions unless one individual is a dictator.

For an introductory example to illustrate the potential problems for consistent social decision making, Arrow considered the question of how to choose among three alternatives {*A*, *B*, *C*} when there are three voters: one voter who ranks *A* best and *C* worst, one voter who ranks *B* best and *A* worst, and one voter who ranks *C* best and *B* worst. In this simple society, a majority of the voters (two out of three) prefer *A* over *B*, but another majority of the voters prefer *B* over *C*, and yet another majority of the voters prefer *C* over *A*. Duncan Black's 1948 paper also discussed this paradoxical example, which is now known as the *Condorcet cycle*, although Condorcet's original version was rather more complicated. A general social choice rule that treats alternatives symmetrically (neutrality) and treats voters symmetrically (anonymity) could not identify a single chosen alternative for this example.

In 1970, Amartya Sen published a *JPE* paper on "The Impossibility of a Paretian Liberal," which extended Arrow's impossibility results by showing a potential incompatibility between liberalism and Pareto's criterion for social efficiency. Sen illustrated his impossibility theorem with a simple paradoxical example that involves two individuals and a social choice about which of them should read a certain pornographic book. A basic axiom of liberalism might stipulate that, when society faces a question of whether a particular book should be read by a particular individual *i* or by nobody, the social choice should be determined according to *i*'s own preferences. In Sen's example, we suppose that at most one individual can read the book. Individual 1, who is a prude, would most prefer that nobody should read the book; but his second choice would be that he (individual 1) should read it, and he would consider individual 2 reading the book as the worst alternative. Individual 2 would prefer reading the book herself rather than having nobody read it, but 2's first preference would be that individual 1 (the prude) should read the book. Then liberalism would require that society should accept 1's ranking that nobody reading the book is better than individual 1 reading it, and liberalism would also require that society should accept 2's ranking that individual 2 reading the book is better than nobody reading it. But individual 2 reading the book is Pareto-dominated by individual 1 reading the book, as that is the one comparison that both individuals agree about. Thus, liberalism can lead to a Pareto-dominated social choice.

These impossibility theorems of social choice theory warn us that, when a group must choose among three or more alternatives, we cannot generally expect to predict what their collective choice will be (or should be) only on the basis of the members' individual preferences over these alternatives. The outcome must also depend on the process by which the group makes collective decisions. That is, public policies must depend

not only on the economic fundamentals that determine individuals' preferences over these policies but also on the structure of the political institutions in which public policy decisions are actually made. Every society has constitutional rules that delegate substantial social decision-making powers to a smaller group of individuals, including leaders who can set the agenda for social decision making (with powers to determine which alternatives will be considered and in what order) and legislative representatives who are empowered to vote on policies for their constituents who elected them.

Thus, since 1980, the literature of political economics has increasingly focused on questions about how the quality of social choices will depend on the structure of political institutions. We want to understand how the structure of political institutions can affect the conduct of political leaders and the performance of government in public policy making.

For example, a 1981 *JPE* paper by Barry Weingast, Kenneth Shepsle, and Christopher Johnsen examined reasons why elected local representatives in a legislature may have incentives to choose inefficiently high levels of public spending. Each locally elected representative cares primarily about the benefits of public spending in his own district, while the costs of public spending are spread among all districts, and so systematic biases may be introduced by the basic structure of representative democracy. Even on a public policy question that has the kind of simple one-dimensional structure that Duncan Black assumed, the selected policy might be determined not by the preferences of the median voter in the nation but by the median among all voters who are at the median of their respective districts.

The 1988 *JPE* paper by Barry Weingast and William Marshall considers the problem of creating a market for Pareto-improving transactions in a legislature that has responsibility for determining multidimensional public policies that it can revise by a majority vote at any time. A simplistic analogy with economic markets might suggest a market for dynamic vote trading, where one representative might sell his vote on some policy dimensions in exchange for others' votes on a dimension that he cares more about. But even if such exchanges were legal and enforceable, efficient vote trades for dynamic policy making would generally require that promises of future votes should be dependent on future political conditions that may be too complex to be stipulated in practical contracts.

Weingast and Marshall argue that a legislature can solve this problem by allocating agenda control over different policy dimensions to different legislative committees and then letting legislators bid for seats in their most-preferred committees. Thus, a committee system allows that an individual legislator can have some confidence of his ability to pre-

vent adverse changes of policy on a dimension that is particularly important to him by preventing such changes from coming to a vote in the legislature. On the other hand, the fact that committee-approved legislation cannot be enacted without ratification by a majority of the legislature effectively constrains the committee's power to change policy except in political conditions in which many other legislators would be receptive to such changes. Thus, rules that vest legislative agenda control in a strong committee system, as in the US Congress, can be understood as a practical institutional adaptation to the difficulties of trading votes in a dynamic policy-making process. Weingast and Marshall also observe that, in countries with strongly disciplined political parties, transactions for Pareto improvement in dynamic legislative policy making may instead be facilitated by party leaders, who serve as trusted intermediaries for exchanges of promises of political support.

Weingast and Marshall's 1988 *JPE* paper is exemplary of research in political economics papers that applies concepts of economic analysis for better understanding of political institutions. This strand in the literature treats political institutions as analytically similar to economic markets, in that both can be understood as systems of competitive interactions among rational agents. But the literature in political economics also includes many papers in which the focus is on how political institutions affect the performance of economic markets themselves. The 2005 *JPE* paper by Daron Acemoglu and Simon Johnson is an outstanding example of this part of the literature.

Economic markets depend on political and legal institutions for the enforcement of property rights and contracts. Acemoglu and Johnson suggest a way to distinguish these two effects empirically in nations that were formerly European colonies. Former European colonies have generally adopted different legal mechanisms for contract enforcement according to whichever country in Europe was the colonizing power. But the quality of property rights enforcement in different colonies has tended to depend more on the degree to which Europeans could expect to make long-term settlements in the colony, which in turn depended on mortality rates for potential European settlers and population density around 1500. Thus, Acemoglu and Johnson (2005, 971) argue that "the way in which countries were colonized, but not who colonized them, is a robust determinant of property rights institutions, whereas who colonized, but not the details of colonization strategy, shapes contracting institutions."

With these instruments, Acemoglu and Johnson find that the historical development of property rights institutions has had a significant effect on long-run economic growth, investment, and financial development. In contrast, they find that institutions for contract enforcement appear to have mattered mainly for the relative importance of different

forms of financial intermediation (debt or equity). To understand these results theoretically, Acemoglu and Johnson observe that any adverse effects of their government's weakness in contract enforcement could be mitigated by citizens organizing social networks with their own private systems of contract enforcement. But ordinary citizens would have no such remedy if institutions for property rights failed to protect private investments from expropriation by those who control the government, and so millions of people would lose any incentive to invest in improving their economic situation. This basic recognition, that failures of economic development can be caused by dysfunctional political institutions, is a fundamental motivation for ongoing research in political economics, applying economic analysis to the comparative study of political institutions.

References

Acemoglu, Daron, and Simon Johnson. 2005. "Unbundling Institutions." *J.P.E.* 113 (5): 949–95.

Arrow, Kenneth. 1950. "A Difficulty in the Concept of Social Welfare." *J.P.E.* 58 (4): 328–46.

Black, Duncan. 1948. "On the Rationale of Group Decision-Making." *J.P.E.* 56 (1): 23–34.

Sen, Amartya. 1970. "The Impossibility of a Paretian Liberal." *J.P.E.* 78 (1): 152–57.

Weingast, Barry R., and William J. Marshall. 1988. "The Industrial Organization of Congress; Or, Why Legislatures, like Firms, Are Not Organized as Markets." *J.P.E.* 96 (1): 132–63.

Weingast, Barry R., Kenneth A. Shepsle, and Christopher Johnsen. 1981. "The Political Economy of Benefits and Costs: A Neoclassical Approach to Distributive Politics." *J.P.E.* 89 (4): 642–64.

Aggregative Fiscal Policy

Nancy L. Stokey

University of Chicago

I. Introduction

There is some truth in the old adage that in economics the questions never change, only the answers. This note looks at three questions in ag-

gregative fiscal policy that have been the subject of recent *JPE* contributions and also have long prior histories: the use of debt finance, the role of commitments about future policy, and the size of government.

II. Debt Finance

What are the effects of financing government spending with debt rather than contemporaneous taxation?

David Ricardo (1817) raised the issue, discussing both ordinary expenses of the state and unusual expenses like wars. He asserted that borrowing by the government "is a system which tends to make us less thrifty—to blind us to our real situation." As an example, Ricardo considered an individual who must pay £100 for the expense of a war, saying that "he would endeavor, on being at once called upon for his portion, to save speedily the £100 from his income. By the system of loans, he is called on to pay only the interest of this £100, or £5 per annum, and considers that he does enough by saving this £5 from his expenditure, and then deludes himself with the belief that he is as rich as before" ([1817] 2004, chap. 18, 163).

In "Are Government Bonds Net Wealth?" (1974), Robert Barro studied the effect of debt finance on consumption and saving in a setting in which it is later generations who will repay the debt. To this end he uses a variant of the Samuelson (1958)–Diamond (1965) overlapping generations model. Each individual works when young and uses his earnings for consumption and asset accumulation. When old he may receive a (nonnegative) inheritance and uses his assets to finance consumption and, perhaps, a bequest to his heir.

Generations are altruistically linked through preferences: each individual values the utility of his heir. Thus, one can readily construct "dynastic preferences" over the consumptions of all the members and a "dynastic budget constraint" augmented by an additional set of constraints representing the fact that bequests must be nonnegative.

Barro asks about the effect of a one-time government transfer to the old, financed with bonds. In every subsequent period the government uses a lump-sum tax to finance the interest on the debt and rolls over the principal. It is easy to show that the bond injection leaves the dynastic budget constraint unchanged. Consequently, if, with no transfer, debt, or taxes, all of the bequests would have been strictly positive, then they will remain so: each bequest will be increased by the size of the next generation's tax liability. The dynasty's consumption plan will be unchanged, and the bond-financed transfer has no effect.

As Buchanan (1976) pointed out, Ricardo had noticed this argument about the capitalization of future tax obligations. However, as Buchanan

suggested and O'Driscoll (1977) argued more forcefully, Ricardo did not accept the argument as a description of behavior. O'Driscoll quotes Ricardo as saying "but the people who pay the taxes never so estimate them, and therefore do not manage their private affairs accordingly" (208). Evidently Ricardo himself was not a Ricardian.

In his subsequent paper "On the Determination of the Public Debt" (1979), Barro noted that even if one accepts the hypothesis about Ricardian equivalence, it holds only if taxes are lump-sum. If revenue must be raised with distorting taxes, a second consideration appears.

Since the deadweight loss from a distorting tax is convex in the tax rate, given a fixed revenue requirement, the total distortion is reduced by levying lower taxes on many commodities instead of a high tax on a single commodity. Thus, if government expenditure is uneven over time, the total distortion is lower if the tax rate remains approximately constant and the government issues debt and/or acquires assets, borrowing and lending as required, to keep the tax rate smooth. Financing extraordinary expenditures like wars is an obvious example, but the principle applies more broadly.

Of course, this idea also has an earlier antecedent, in Frank Ramsey's (1927) classic study. Ramsey looked at a static problem, where the issue is taxing various commodities at possibly different rates. But his idea applies as well to a dynamic setting in which the commodities are dated consumption goods.

III. The Role of Commitment

Can government policies that seem attractive in the short run be detrimental in the long run? Can policy ever be improved by taking options away from a benevolent government?

Soon after the adoption of the American Constitution, the new Treasury Secretary, Alexander Hamilton, argued in his *First Report on Public Credit* (1790) for assumption and repayment by the federal government of all outstanding debt that the states had issued during the Revolutionary War, asserting that full repayment would establish the nation's reputation with credit markets and allow it to borrow in the future at affordable interest rates. Thomas Jefferson and James Madison took the opposite side of the debate, arguing that much of the debt had been bought up by speculators at less than face value and that those creditors should be paid less. In the end Hamilton prevailed, and all of the debt was repaid.

In "Rules Rather than Discretion: The Inconsistency of Optimal Plans" (1977), Finn Kydland and Edward Prescott identified a fundamental limitation in using the tools of dynamic optimization to formulate govern-

ment policy. The key issue is that "current decisions of economic agents depend in part upon their *expectations* of future policy actions" (474; emphasis added).

Consequently, a government at one date may announce/suggest a policy for a later date, because of the beneficial effect on current private-sector behavior. But the (benevolent) policy maker at the later date—if it can—may choose to ignore the suggestion and do something else instead. Of course, if agents in the private sector anticipate this change, it undermines the beneficial effect the announcement was intended to have.

What Kydland and Prescott call a rule is a limitation, through a constitutional provision or legislation or some other means, on what a policy maker can do. By contrast, a regime of discretion allows the policy maker wide latitude in deciding how to respond. In many situations a rule may be strictly preferred—by all—to a regime with discretion.

As an example, they describe the problem faced by a central bank. Under a regime with discretion, it may be tempted to announce a low inflation target and then exceed it, with the goal of reducing unemployment. But rational agents will anticipate the additional inflation, so unemployment is unaltered and inflation is higher than desired. A rule would tie the central bank's hands.

Similarly, in the short run, capital income seems to be an excellent target for taxation, since it is supplied inelastically. But in the longer run, such a tax discourages investment. A government could tax capital income heavily in the short run and at the same time promise very low taxes in the future. But is the promise credible?

And finally, defaulting on public debt, either explicitly or implicitly—through inflation—reduces the need to levy distorting taxes, a pure benefit. But if potential lenders anticipate a risk of default, the cost of subsequent borrowing rises. Perhaps Jefferson had a change of heart about Hamilton's position when, during his presidency, debt was issued to finance the Louisiana Purchase, and Madison during his presidency, when it was used to finance the War of 1812.

IV. The Size of Government

What determines the size of government?

Alexis de Tocqueville (1835) argued that it depends on which of three classes is in control of making the laws. If it is the rich, "probably it will be little enough concerned about economizing on public funds, because a tax that strikes a considerable fortune takes away only from the surplus and produces little sensible effect." If the middle classes make the laws, "they will not be prodigal with taxes because there is nothing so disastrous as a large tax striking a small fortune." Finally, if those with "little or no property" make the laws, public costs will be higher, since "all the

money that is expended in the interest of society seems able only to profit them without ever harming them," and they are capable of finding "the means of assessing the tax in a manner that strikes only the rich and profits only the poor" (2000, vol. 1, pt. 2, chap. 5, 200).

Since the "different categories can be more or less numerous," in a democracy the group sizes and the extent of the franchise determine which group is decisive.

In "A Rational Theory of the Size of Government" (1981), Allan Meltzer and Scott Richard expanded on this idea, developing a model in which the only role of the government is to redistribute income, and its only instrument is a flat-rate tax on earnings, used to finance lump-sum transfers. The government's budget is balanced, and the size of government—the share of tax revenue in total income—is determined by voting.

Individuals are identical in terms of their preferences over consumption and leisure but heterogeneous in terms of labor productivity. Given the tax system in place, each individual chooses how much time to spend working.

As voters, individuals realize that a higher tax rate, up to a point, finances a bigger transfer, although they rationally forecast the disincentive effects of taxation: there is no "fiscal illusion." Hence preferences over the tax rate are monotone in the individual's own productivity, with all those below some threshold preferring to implement the revenue-maximizing rate and doing no work.

The political equilibrium is determined by the median voter, so it depends on the distribution of productivity across individuals and the extent of the franchise. For a fixed distribution of income, expanding voting rights to groups with lower incomes increases the demand for redistribution. Once suffrage is universal, the growth of government depends only on changes in the distribution of income. If economic growth increases income inequality, then the model predicts that growth also increases the demand for redistributive taxation.

V. Conclusion

Are the formal models of modern economics only an embellishment of older ideas? They are more than that, for two reasons. First, they clarify exactly what is being asserted, providing a more solid base from which further theoretical arguments can proceed. In addition, they provide a guide for empirical work, suggesting what kinds of data should be gathered and how they should be used to examine the competing hypotheses. The old adage is only partly correct: the questions get sharper and clearer, even if entirely satisfactory answers remain elusive. Unsurprisingly, research on these three questions has continued: all of these papers are among the most highly cited in the *JPE* over its 125 years.

References

Barro, Robert J. 1974. "Are Government Bonds Net Wealth?" *J.P.E.* 82 (November/December): 1095–1117.
———. 1979. "On the Determination of the Public Debt." *J.P.E.* 87, no. 5, pt. 1 (October): 940–71.
Buchanan, James M. 1976. "Barro on the Ricardian Equivalence Theorem." *J.P.E.* 84 (April): 337–42.
Diamond, Peter A. 1965. "National Debt in a Neoclassical Growth Model." *A.E.R.* 55 (December): 1126–50.
Hamilton, Alexander. 1790. *First Report on Public Credit.* Washington, DC: Treasury Dept.
Kydland, Finn E., and Edward C. Prescott. 1977. "Rules Rather than Discretion: The Inconsistency of Optimal Plans." *J.P.E.* 85 (June): 473–92.
Meltzer, Allan H., and Scott F. Richard. 1981. "A Rational Theory of the Size of Government." *J.P.E.* 89 (October): 914–27.
O'Driscoll, Gerald P., Jr. 1977. "The Ricardian Nonequivalence Theorem." *J.P.E.* 85 (February): 207–10.
Ramsey, Frank P. 1927. "A Contribution to the Theory of Taxation." *Econ. J.* 37:47–61.
Ricardo, David. [1817] 2004. *The Principles of Political Economy and Taxation.* Reprint. Mineola, NY: Dover.
Samuelson, Paul A. 1958. "An Exact Consumption-Loan Model of Interest With or Without the Social Contrivance of Money." *J.P.E.* 66 (December): 467–82.
Tocqueville, Alexis de. [1835] 2000. *Democracy in America.* Translated and edited by Harvey C. Mansfield and Delba Winthrop. Chicago: Univ. Chicago Press.

Business Cycles and International Trade

Harald Uhlig

University of Chicago

My essay will examine two rather separate topics, though there is a bit of a connection. One concerns business cycles. The other concerns international trade and exchange rates. With all due apologies and very few exceptions, I shall focus on the most highly cited papers published in the *Journal of Political Economy.*

Business Cycles

The 1970s and early 1980s saw a revolution in our thinking about macroeconomics generally and business cycles specifically. Central to this de-

velopment was a revolutionary paradigm shift in how the expectations of agents regarding the future should be taken into account in their current choices, notably for consumption, mostly insisting that these expectations should be rational and agree with the formulation of probability theory. Before that revolution, it was customary to assume a consumption function, according to which aggregate consumption rises in somewhat less than a proportional manner and according to the marginal propensity to consume, when current aggregate income rises, regardless of concerns about developments of these incomes in the future. That older paradigm is still remarkably alive in quite a number of undergraduate textbooks on macroeconomics and policy discussions, but it has been entirely upended by the rational expectations revolution as far as the scientific analysis and thinking about business cycles and other economic phenomena are concerned: only in special cases then does the old paradigm still work. From the beginning and in particular in the work by Sargent, the hypothesis of rational expectations was connected to the issue of how rational expectations or other forms of expectations can be learned (see Uhlig 2012a).

Perhaps the most seminal contribution in this (then) new thinking about consumption is the paper by Lucas (1978). He examined the optimization problem of an agent with time-separable preferences, who can freely trade assets with stochastic returns R_{t+1} in period t on one unit of resources invested at t. He derived what is now typically referred to as the Lucas asset pricing equation:

$$1 = E_t[M_{t+1} R_{t+1}], \quad \text{where } M_{t+1} = \beta \frac{u'(c_{t+1})}{u'(c_t)},$$

where $u'(c_t)$ is the period t felicity for an agent, consuming c_t; β is the discount factor; E_t denotes the conditional expectation, given all available information at time t; and M_{t+1} is the resulting stochastic discount factor. The first part of the equation also holds for far more general preference formulations and has given rise to a substantial literature on asset pricing, as other essays in this issue discuss.

Here I shall focus on the macroeconomic implications and (mostly) keep to the separable formulation. Assuming $R_{t+1} = 1 + r$ to be a constant and safe return, it then follows that detrended marginal utility $[\beta(1 + r)]^t u'(c_t)$ is a random walk. If, moreover, utility is quadratic, then consumption likewise detrended is itself a random walk with drift. These are the celebrated results in Hall (1978), who writes that therefore "consumption is unrelated to *any* economic variable that is observed in earlier periods. In particular, lagged income should have no explanatory power with respect to consumption" (972). Hall proceeds to test and to then confirm these permanent-income predictions of the theory, while Sar-

gent (1978) instead obtains a rather decisive rejection. Flavin (1981) reconciles these two apparently conflicting findings. She rejects the joint rational expectations–permanent income hypothesis and finds that consumption exhibits excess sensitivity to current income.

For that exercise, it is ultimately crucial to estimate the revision in permanent income and the persistent reaction of income due to current news. Cochrane (1988) estimates the permanent reaction to be fairly small. The literature on persistence, unit roots, and cointegration since then has grown to impressive size.

Hall (1988) assumes that $u(c) = c^{1-1/\sigma}$ so that σ is the intertemporal elasticity of substitution: a popular specification in much of macroeconomics. Exploiting time variation in R_{t+1}, he calculates various estimates of σ and generally finds them to be small, near zero, or even negative. The macroeconomic literature since then has tended to assume σ to be between 0.5 and 1, and sometimes as low as 0.2, as well as allowed for extensions such as habit formation and borrowing-constrained or hand-to-mouth consumers. That literature furthermore typically assumes the log of total factor productivity (TFP) to exhibit short-run fluctuations around a time trend or to be a random walk with drift. These are then ingredients for building more substantial business cycle models.

The revolution in thinking about business cycles was to view them as equilibrium phenomena, where agents optimally react to shocks and policy changes, utilizing rational expectations. The program was laid out in Lucas (1975), though that paper did not yet feature preference-based optimizing behavior of agents. The program was completed in particular in the seminal contribution of Kydland and Prescott (1982), extending the stochastic neoclassical growth theory and giving rise to real business cycle theory. The contribution by Long and Plosser (1983) allows for a rich industry structure: a theme that recently has received considerable renewed attention in the production-network-based analysis of macroeconomic fluctuations. Real business cycle theory postulates that aggregate fluctuations are driven by exogenous fluctuations in TFP rather than, say, exogenous fluctuations in "aggregate demand" (which now would need to be derived from exogenous fluctuations in preference parameters) or policy. Prices and wages are assumed to be flexible and markets are assumed to clear. Many now dismiss such flexibility out of hand as unrealistic, and the literature has since moved to typically imposing a range of other frictions. Then again, one may argue that "reality" provides a better rationale in its favor than may be apparent at first (see Uhlig 2012b).

The real business cycle paradigm has since been extended and critically examined in a variety of ways. Backus, Kehoe, and Kydland (1992) extend the paradigm to the international realm, providing a connection

between the two sections here. Empirically, Hamilton (1983) argues that oil prices provide a substantial source of aggregate fluctuations, with Mork (1989) arguing that the effect is asymmetric and much stronger for oil price increases than oil price decreases. Basu and Fernald (1997) is an important paper, examining the intricacies of measuring the exogenous component of TFP and the challenges in utilizing it as a driving force. More recent versions of business cycle theories enrich them with a considerably larger set of shocks and frictions. In particular, the assumption of sticky prices is appealing to many and has become a standard ingredient of most of the business cycle analysis in the recent decade or so. One important example is the framework by Christiano, Eichenbaum, and Evans (2005), which, together with the related Smets and Wouters (2003) model, has become the blueprint for many workhorse models used in central banks around the world for policy analysis. In the wake of the financial crisis of 2008, these models have recently become extended by paying greater attention to financial intermediation and the role of the financial frictions more generally. At its core, all these models still feature a real business cycle engine, albeit modified and extended in considerable ways.

International Trade and Exchange Rates

Balassa (1964) together with Samuelson (1964) is the classic source for the well-known Balassa-Samuelson effect, that the purchasing power parity or consumer price level is higher in richer countries.

Dornbusch (1976) develops his classic exchange rate overshooting result for exchange rates. He assumes perfect foresight: the companion to rational expectations, if there are no further stochastic disturbances in the future. He considers a monetary expansion in a model of perfect capital mobility and slow adjustments of goods markets. He demonstrates that the initial and immediate depreciation of the exchange rate is then followed by a gradual appreciation of the exchange rate, to compensate for the ensuing inflation differential. Lothian and Taylor (1996) use unit root econometric methods, freshly developed in the decade prior to the publication of their paper, and demonstrate that the dollar-sterling and the franc-sterling real exchange rates are stationary, an important issue for the construction of international trade models.

The study of international trade and exchange rates has undergone profound paradigm shifts in the last few decades. It has incorporated the macroeconomic paradigm shift toward rational expectations and general equilibrium analysis described in the first section. The new trade theory furthermore views trade as arising from imperfect competition between possibly multinational firms, each producing its own variety.

Helpman (1984) provides a simple theory of international trade with multinational corporations, building on then-recent advances in analyz-

ing vertical integration and international trade in differentiated products. For production of a specific variety, he distinguishes between a general-purpose input possibly produced elsewhere, such as management, distribution and product-specific R&D, and local labor. Multinational corporations with entrepreneurial centers and subsidiaries together with their location decision then arise endogenously, explaining the simultaneous existence of intersectoral trade, intraindustry trade, and intrafirm trade. More recently, Antràs and Helpman (2004) examine the issue of global sourcing and the choice of organizational form for firms in international trade and relate sectoral productivity dispersion and headquarter intensity to the degree of integration and input imports.

Backus et al. (1992) have extended the real business analysis described in the previous section to a two-country setting. While they do not feature firm heterogeneity or sticky prices, they emphasize in particular the role of the capital stock and capital investment. More recent trade models often abstract from physical capital accumulation, though it may remain fruitful to include such forces as well.

Obstfeld and Rogoff (1995) critically reexamine the Dornbusch overshooting result as well as a number of other classic predictions in a new two-country model. Their model marries global macroeconomic dynamics to a supply framework based on monopolistic competition and sticky nominal prices, thereby providing novel insights into the dynamics of exchange rates and current accounts. It has become a benchmark and workhorse model in this field of inquiry. The latest generation of international trade models builds on the seminal contributions of Eaton and Kortum (2002) and Melitz (2003), focusing on matters such as firm entry and exit as well as trade costs, which the *JPE* unfortunately missed out on publishing: at least Samuel Kortum was on the faculty at the University of Chicago for a number of years. The field has been moving forward quickly in recent years, and these developments will be exciting to watch or to participate in.

References

Antràs, Pol, and Elhanan Helpman. 2004. "Global Sourcing." *J.P.E.* 112 (3): 552–80.

Backus, David K., Patrick J. Kehoe, and Finn E. Kydland. 1992. "International Real Business Cycles." *J.P.E.* 100 (August): 745–75.

Balassa, Bela. 1964. "The Purchasing-Power Parity Doctrine: A Reappraisal." *J.P.E.* 72 (December): 584–96.

Basu, Susanto, and John G. Fernald. 1997. "Returns to Scale in U.S. Production: Estimates and Implications." *J.P.E.* 105 (April): 249–83.

Christiano, Lawrence J., Martin Eichenbaum, and Charles L. Evans. 2005. "Nominal Rigidities and the Dynamic Effects of a Shock to Monetary Policy." *J.P.E.* 113 (February): 1–45.

Cochrane, John H. 1988. "How Big Is the Random Walk in GNP?" *J.P.E.* 96 (October): 893–920.
Dornbusch, Rudiger. 1976. "Expectations and Exchange Rate Dynamics." *J.P.E.* 84 (December): 1161–76.
Eaton, Jonathan, and Samuel Kortum. 2002. "Technology, Geography, and Trade." *Econometrica* 70 (5): 1741–79.
Flavin, Marjorie A. 1981. "The Adjustment of Consumption to Changing Expectations about Future Income." *J.P.E.* 89 (October): 974–1009.
Hall, Robert E. 1978. "Stochastic Implications of the Life Cycle–Permanent Income Hypothesis: Theory and Evidence." *J.P.E.* 86 (December): 971–87.
———. 1988. "Intertemporal Substitution in Consumption." *J.P.E.* 96 (April): 339–57.
Hamilton, James D. 1983. "Oil and the Macroeconomy since World War II." *J.P.E.* 91 (April): 228–48.
Helpman, Elhanan. 1984. "A Simple Theory of International Trade with Multinational Corporations." *J.P.E.* 92 (June): 451–71.
Kydland, Finn E., and Edward C. Prescott. 1982. "Time to Build and Aggregate Fluctuations." *Econometrica* 50:1345–70.
Long, John B., and Charles I. Plosser. 1983. "Real Business Cycles." *J.P.E.* 91 (February): 39–69.
Lothian, James R., and Mark P. Taylor. 1996. "Real Exchange Rate Behavior: The Recent Float from the Perspective of the Past Two Centuries." *J.P.E.* 104 (June): 488–509.
Lucas, Robert E., Jr. 1975. "An Equilibrium Model of the Business Cycle." *J.P.E.* 83 (December): 1113–44.
———. 1978. "Asset Prices in an Exchange Economy." *Econometrica* 46 (6): 1429–45.
Melitz, Marc J. 2003. "The Impact of Trade on Intra-industry Reallocations and Aggregate Industry Productivity." *Econometrica* 71:1695–1725.
Mork, Knut Anton. 1989. "Oil and the Macroeconomy When Prices Go Up and Down: An Extension of Hamilton's Results." *J.P.E.* 97 (June): 740–44.
Obstfeld, Maurice, and Kenneth Rogoff. 1995. "Exchange Rate Dynamics Redux." *J.P.E.* 103 (June): 624–60.
Samuelson, Paul A. 1964. "Theoretical Notes on Trade Problems." *Rev. Econ. and Statis.* 46 (2): 145–54.
Sargent, Thomas J. 1978. "Rational Expectations, Econometric Exogeneity, and Consumption." *J.P.E.* 86 (August): 673–700.
Smets, Frank, and Raf Wouters. 2003. "An Estimated Dynamic Stochastic General Equilibrium Model of the Euro Area." *J. European Econ. Assoc.* 1 (5): 1123–75.
Uhlig, Harald. 2012a. "Agents as Empirical Macroeconomists: Thomas J. Sargent's Contribution to Economics." *Scandinavian J. Econ.* 114 (December): 1055–81.
———. 2012b. "Economics and Reality." *J. Macroeconomics* 34:29–41.

Inequality, Heterogeneity, and Consumption in the *Journal of Political Economy*

Greg Kaplan

University of Chicago

Today, inequality and heterogeneity are front and center in macroeconomics. Most macroeconomists agree that the distribution of household-level variables, in particular consumption and wealth, matters for the dynamics of macroeconomic aggregates and that macroeconomic shocks affect the distribution of consumption and wealth across households. However, it was not always this way. Getting to this point has been a long road, along which papers published in the *JPE* have been essential guideposts.

I will discuss six influential papers from the *JPE* that have helped shape the way in which inequality is studied by economists today. I have organized my discussion along three strands that have each contributed to the introduction of heterogeneity into macroeconomic models: the microeconomics of consumption behavior, the use of structural models of precautionary savings for insights about policy, and finally, the development of general equilibrium models with heterogeneous households and aggregate shocks. Within each of the three strands, I have chosen two defining *JPE* papers that roughly correspond to the trajectory that this intellectual journey has followed.

First I will examine two empirical papers—Zeldes (1989) and Attanasio and Weber (1995)—that provided new insights into the microeconomics of consumption behavior. Both are empirical analyses that are closely guided by theory. These and other related empirical papers paved the way for a class of structural models of consumption—precautionary savings models—that later became the workhorse models of heterogeneous agent macroeconomics. I will next highlight two early examples of how structural models of consumption can be used to explore the ways in which economic policies shape the distribution of household outcomes. Hubbard, Skinner, and Zeldes (1995) and Rosenzweig and Wolpin (1993) illustrate two alternative ways in which structural models of consumption can be disciplined by data and turned into quantitative laboratories; the first calibrated their model of the United States using mostly external sources, while the second estimated their model of rural India using maximum likelihood. In both models, precautionary motives drive household responses to changing government policies. Finally, I will discuss two early

heterogeneous agent models with aggregate shocks: Imrohoroglu (1989), one of the earliest partial equilibrium models, and Krusell and Smith (1998), a now-seminal general equilibrium model. These took previously stand-alone precautionary savings models and cast them in an equilibrium framework exposed to aggregate shocks, thus opening the door for fully fledged macroeconomic models of inequality.

Zeldes (1989) was one of the first papers to provide convincing evidence using micro panel data that liquidity constraints are indeed important for household-level consumption. The importance of his contribution is reflected in the fact that today, virtually every paper being written about consumption, at either the individual or aggregate level, somewhere addresses the implications of binding liquidity constraints. The paper is notable because rather than simply rejecting the implications of consumption models that abstract from liquidity constraints, as had been done by much of the preexisting literature, he carefully derives testable implications of a model that includes liquidity constraints. Zeldes suggests three novel tests for the importance of these constraints. The first test is beautiful for its simplicity. Because the model with liquidity constraints predicts that consumption growth should be sensitive to income for low-wealth households but not for high-wealth households, Zeldes splits his sample between these two groups of households and measures the sensitivity of consumption growth to income growth for each. Using data on food consumption from the Panel Study of Income Dynamics (PSID), he confirms the predictions of the model: that low-wealth households have a much lower sensitivity of consumption growth to income than high-wealth households. Refined versions of this test for the presence of liquidity constraints are still the go-to approaches in empirical analyses of consumption behavior today. Many of these recent studies (and there are many) essentially repeat Zeldes's analysis using better-identified income shocks and much larger and higher-quality data sets, reaching the same conclusion.

Zeldes (1989) also derives two additional tests implied by the consumption model with liquidity constraints: (i) that Lagrange multipliers on liquidity constraints should be positive for constrained households and (ii) that the multipliers should be negatively related to current income. He estimates Lagrange multipliers for low-wealth households by using the residuals from their Euler equations when the remaining parameters are estimated in the sample of high-wealth households. He finds that the estimated multipliers are indeed positive and are negatively related to income, as predicted. One leaves the paper with the sense that any model of household consumption should treat liquidity constraints seriously. The structural literature that followed did.

Another beautifully executed empirical analysis of consumption behavior that is also carefully motivated by theory—this time without liquid-

ity constraints—is Attanasio and Weber (1995). Using the Consumer Expenditure Survey (CEX), the authors pioneered the approach of working with synthetic cohorts (i.e., data moments for groups of households from the same birth cohort with similar demographic characteristics) to overcome the limitation that the CEX has only a short panel component, unlike the PSID. This enabled them to exploit the comprehensive consumption data in the CEX, rather than the PSID, which at the time contained data on only food consumption.

Attanasio and Weber (1995) make three points. First, they show that looking only at food consumption can be misleading because preferences are nonseparable between food consumption and other consumption categories. Partly for this reason, the vast majority of the structural consumption literature that has followed favors comprehensive measures of consumption or explicitly models this nonseparability. Second, they illustrate the pitfalls of using aggregate data to test models of heterogeneous households. By aggregating microdata in exactly the way prescribed by theory, they show that there can be large differences between the dynamics of the log of mean consumption (the focus of representative agent models and what is measured in aggregate data) and the dynamics of the mean of log consumption (the focus of heterogeneous agent models and that can be constructed only with household-level data). Third, they show that it is easy to spuriously reject frictionless life cycle models of consumption if one ignores predictable changes in either household composition or labor supply of individual household members. They explain how the hump-shaped age profiles for family size and female labor supply would lead to a hump-shaped age profile for consumption—an observation that had frequently been cited as evidence against the frictionless model.

All three points lead one to rethink the numerous previous studies that had seemingly shown deviations from frictionless consumption models, including Zeldes (1989). Attanasio and Weber (1995) never actually claimed that liquidity constraints and precautionary motives were not important, only that the then-existing tests were much more fragile than one might have thought. Reading both of these papers today, one is struck by the careful connection between empirics and theory. Both papers carefully explain their null and alternative hypotheses, build up their estimating equations from precisely specified models, and go to great lengths to spell out the assumptions required to go from their model to their regression equations. These are classic qualities of empirical analyses in the *JPE*.

Attanasio and Weber (1995) essentially highlight the limitations of examining theories of consumption through the lens of only a small subset of a model's predictions. By exposing these limitations, they drove later papers to use a larger set of predictions that are obtained by explicitly computing consumption and savings decisions under alternative parameteri-

zations. These structural papers that followed moved beyond testing models to using models to quantify the effects of public policies on household consumption and savings behavior. Two *JPE*-published papers represent some of the best early examples of how to utilize a quantitative structural model of consumption for effective policy analysis.

Hubbard et al. (1995) is a classic example of the power of a calibrated structural model. The authors observe that many households with low lifetime incomes accumulate little or no wealth over their lifetimes. Even just before retirement, when life cycle models of precautionary savings predict that households should hold substantial wealth, many such households are essentially hand to mouth. According to life cycle consumption theory, having low lifetime income, even in the presence of liquidity constraints, is no excuse for not saving for retirement: households should smooth consumption, albeit at a low level. While liquidity constraints might explain why households do not borrow, it does not explain why they do not save. Hubbard et al. suggest a possible reason for the lack of saving: asset-based means-tested public insurance programs, which they model as a consumption floor, reduce the incentives for households to save. The presence of a consumption floor not only reduces households' exposure to consumption fluctuations—lowering their incentive to save for precautionary reasons—but also implies an effective tax rate of 100 percent on assets in the states of the world where the consumption floor binds.

The authors use a simple two-period model as an elegant theoretical proof of concept. But the calibrated life cycle model, which is the meat of the paper, provides two additional benefits. First, it acts as a quantitative proof of concept, which, in my opinion, is one of the most valuable benefits of quantitative structural analysis. It is one thing to show that asset-based means-tested public insurance programs can distort savings decisions; it is another to show that in empirically plausible settings these distortions are large enough to have an economically important effect on observed savings. To do so, the authors choose parameter values that they argue reflect US data, of which the most important are the stochastic processes for earnings risk and medical expense risk, the level of the consumption floor, and the degree of risk aversion. They then simulate their model economy and show that it generates the aforementioned patterns of life cycle wealth accumulation by lifetime income. Second, the calibrated model can be used to evaluate the implications of alternative versions of means-testing public programs for wealth accumulation, which the authors show can be substantial.

Another example of how a quantitative structural model of precautionary savings can be used to evaluate public policies is Rosenzweig and Wolpin (1993). The authors consider the consumption-savings problem of farmers in India whose only mechanism for smoothing consumption is the accumulation of bullocks. Bullocks are also an input used in agricul-

tural production, making this paper one of the first examples of a structural model in which households save in a productive asset in the face of idiosyncratic risk. The authors confront their model with panel data from the International Crops Research Institute for the Semi-arid Tropics. They construct the likelihood over sequences of farmers' assets and profits and use a two-stage maximum likelihood procedure to estimate preference parameters, prices of bullocks and other inputs, and production parameters. Even today, Rosenzweig and Wolpin's paper remains one of the few examples of maximum likelihood estimation of a precautionary savings model of consumption using microdata. Their parameter estimates imply underinvestment in bullocks on the part of farmers, as a result of borrowing constraints and the inability of farmers to accumulate precautionary savings in a financial asset. Through a series of counterfactual experiments, the authors evaluate the relative merits of alternative interventions. They find that the provision of actuarially fair weather insurance would have little effect on farmer welfare, whereas access to assured income streams would have a large effect on welfare. These are quantitative conclusions that can be obtained only with a suitably parameterized model.

There are important senses in which the models of neither Hubbard et al. (1995) nor Rosenzweig and Wolpin (1993) are "macroeconomic." First, in neither model is the return on savings determined as an equilibrium outcome. Second, in neither paper do the authors explore how aggregate disturbances affect the economy. I will finish by discussing two papers in the *JPE* that contributed to the transition toward developing realistic models of consumption that are macroeconomic in this sense.

Imrohoroglu (1989) was a significant early paper that recognized the potential importance of precautionary motives in the presence of aggregate shocks. The paper was motivated by Lucas's (1987) famous costs of business cycles calculation. He had shown that in representative agent economies the welfare costs of business cycles are small both because fluctuations in aggregate income are themselves small and because these fluctuations have only a second-order effect on welfare. It was natural to conjecture that in heterogeneous agent economies with incomplete markets this quantitative conclusion might be overturned, both because fluctuations in individual income can be substantial and because the presence of liquidity constraints means that for some households these fluctuations have a first-order effect on welfare.

Imrohoroglu (1989) set out to evaluate this conjecture. She examines a consumption-savings model with liquidity constraints in which households face unemployment risk that varies stochastically with macroeconomic conditions. It is interesting to note how our understanding (and expectations) of what it means for a macroeconomic model to be labeled as "general equilibrium" has evolved. Despite describing her environment as general equilibrium, most macroeconomists today would describe

her model as a partial equilibrium environment because all prices—interest rates, wages, job destruction rates, and job finding rates—are exogenous. She finds that when aggregate shocks change the extent of unemployment risk faced by households, the welfare cost of business cycles can be four to five times larger than in a corresponding representative agent economy.

Perhaps the most influential macroeconomic model with heterogeneous agents and incomplete markets is that of Krusell and Smith (1998). They study an infinite-horizon consumption-savings problem in which ex ante identical households are subject to idiosyncratic unemployment risk. As in the other models I have discussed, households can self-insure this risk through a single risk-free asset. Krusell and Smith's innovation was to embed this precautionary savings problem in a stochastic version of the neoclassical growth model. As in Aiyagari (1994), they interpret the savings instrument as capital that is used by a representative firm as input to a constant returns to scale production function. The interest rate earned by households is thus determined in equilibrium as the marginal product of capital. However, they differ from Aiyagari in that they allow for the possibility that the production function is disturbed by exogenous stochastic productivity shocks.

Krusell and Smith (1998) wanted to understand how the equilibrium business cycle dynamics of macroeconomic variables in this heterogeneous agent economy compare to those in a corresponding representative agent economy—an important open question at the time. If the macroeconomic dynamics of the two economies were not too different, it would provide some justification for the common practice of studying macroeconomics through the lens of a single representative agent. Answering this question, however, required solving their model, which raised substantial challenges. Even before Krusell and Smith's study, it was well understood that the relevant state variable in this type of economy is an infinite-dimensional object—the endogenous cross-sectional distribution of households' employment states and holdings of capital.

The best word to describe Krusell and Smith's (1998) approach to this challenge is "chutzpah." Perhaps, they thought, all the information contained in the distribution of household wealth is overkill. What if we look for an equilibrium in a smaller space by summarizing the distribution with only a finite-dimensional set of moments? What if we use just one moment, the mean? Lo and behold, it worked, in a very precise sense. They showed that using only the mean of the distribution of capital holdings, households could forecast future interest rates extremely accurately, which are what matter for consumption decisions. Thus, Krusell and Smith could approximate the equilibrium with a much smaller and computationally feasible set of state variables.

Krusell and Smith (1998) labeled this finding "approximate aggregation." It arises because in precautionary savings models optimal savings

decisions are extremely close to linear, except for households with very little capital. But since the savings decisions of households with little capital matter little for the dynamics of aggregate capital, the dynamics of aggregate capital (and hence the interest rate) depends approximately on only the level of aggregate capital, not on the distribution of capital across households.

Using this computational strategy, Krusell and Smith (1998) simulate the dynamics of aggregate output, consumption, and investment in a plausibly calibrated version of their model. They find that the dynamics of these variables are virtually indistinguishable from the dynamics of a similarly calibrated representative agent economy. It is important to remember that their finding of indistinguishability between the aggregate dynamics of the heterogeneous agent and representative agent economies is conceptually different from their finding of approximate aggregation. It is relatively easy to construct economies in which approximate aggregation holds but in which aggregate dynamics look different in the corresponding heterogeneous agent and representative agent economies. For example, they show that when the model is modified to better match the empirical distribution of wealth (in part by exploiting the ideas in Hubbard et al. [1995]), the comovement of consumption and income looks very different from the corresponding representative agent economy.

The lasting influence of Krusell and Smith (1998) is remarkable. It has turned out that approximate aggregation is far more applicable than one might have thought and has been used in a number of other contexts in papers published in the *JPE*. For example, a variant of the Krusell and Smith algorithm was used by Khan and Thomas (2013) in the context of an economy with heterogeneous firms and by Favilukis, Ludvigson, and Van Nieuwerburgh (2017) in the context of a model with fluctuating aggregate house prices.

The *JPE* has played an essential role in fostering the growth of the study of macroeconomics with heterogeneity. I hope, and predict, that the journal will continue to play such a role in the future.

References

Aiyagari, S. Rao. 1994. "Uninsured Idiosyncratic Risk and Aggregate Savings." *Q.J.E.* 109 (3): 659–84.

Attanasio, Orazio P., and Guglielmo Weber. 1995. "Is Consumption Growth Consistent with Intertemporal Optimization? Evidence from the Consumer Expenditure Survey." *J.P.E.* 103 (6): 1121–57.

Favilukis, Jack, Sydney C. Ludvigson, and Stijn Van Nieuwerburgh. 2017. "The Macroeconomic Effects of Housing Wealth, Housing Finance, and Limited Risk Sharing in General Equilibrium." *J.P.E.* 125 (1): 140–223.

Hubbard, R. Glenn, Jonathan Skinner, and Stephen P. Zeldes. 1995. "Precautionary Saving and Social Insurance." *J.P.E.* 103 (2): 360–99.

Imrohoroglu, Ayse. 1989. "Cost of Business Cycles with Indivisibilities and Liquidity Constraints." *J.P.E.* 97 (6): 1364–83.
Khan, Aubhik, and Julia K. Thomas. 2013. "Credit Shocks and Aggregate Fluctuations in an Economy with Production Heterogeneity." *J.P.E.* 121 (6): 1055–1107.
Krusell, Per, and Anthony A. Smith Jr. 1998. "Income and Wealth Heterogeneity in the Macroeconomy." *J.P.E.* 106 (5): 867–96.
Lucas, Robert E., Jr. 1987. *Models of Business Cycles*. London: Blackwell.
Rosenzweig, Mark R., and Kenneth I. Wolpin. 1993. "Credit Market Constraints, Consumption Smoothing, and the Accumulation of Durable Production Assets in Low-Income Countries: Investments in Bullocks in India." *J.P.E.* 101 (2): 223–44.
Zeldes, Stephen P. 1989. "Consumption and Liquidity Constraints: An Empirical Investigation." *J.P.E.* 97 (2): 305–46.

Time-Series Econometrics in Macroeconomics and Finance

Lars Peter Hansen

University of Chicago

I. Introduction

Ninety years ago, Slutsky (1927) and Yule (1927) opened the door to the use of probability models in the analysis of economic time series. Their vision was to view economic time series as linear responses to current and past independent and identically distributed impulses or shocks. In distinct contributions, they showed how to generate approximate cycles with such models. Each had a unique background and perspective. Yule was an eminent statistician who, in the words of Stigler (1986), among his many contributions, managed "effectively to invent modern time series analysis" (361). Yule constructed and estimated what we call a second-order model and applied it to study the time-series behavior of sunspots. Slutsky wrote his paper in Russia in the 1920s motivated by the study of business cycles. Much later, his paper was published in *Econometrica*, but it was already on the radar screen of economists, such as Frisch. Indeed Frisch was keenly aware of both Slutsky (1927) and Yule (1927) and acknowledged both in

his seminal paper (1933) on the impulse and propagation problem. Building on insights from Slutsky and Yule, Frisch pioneered the use of impulse response functions in economic dynamics. His ambition was to provide explicit economic interpretations for how current-period shocks alter economic time series in current and future time periods.[1] The *Journal of Political Economy* provided an important platform for research that confronts Frisch's ambition in substantively interesting ways.

II. Rational Expectations Econometrics

A stumbling block for implementing Frisch's (1933) ambition was how to capture people's beliefs about the future. Investment and other decisions are in part based on people's views of the future. Constructing prudent economic policy depends in part on how private agents will respond in the future. Once economic decision makers are included in formal dynamic economic models, their expectations come into play and become an important ingredient to the model specification. Thus the time-series econometrics research agenda grounded in economics had to take a stand on how people inside economic models made forecasts.[2] The rational expectations approach pioneered by Muth (1961) and Lucas (1972a, 1972b) provided a coherent and model-consistent way to capture people's beliefs. It has been implemented in different ways in econometric practice. One way is to exploit the resulting rational expectations equilibrium by fully specifying the underlying economic model. The resulting model solution then determines the beliefs of the economic agents inside the economic model. Empirical evidence comes into play because econometricians face uncertainty about the underlying parameters of the rational expectations equilibrium and use data to infer their values. This vision is well articulated in Sargent's (1981) *JPE* treatise on interpreting economic time series. The restrictions are sometimes implemented with two-step shortcuts whereby parameters of processes for exogenous dynamics are estimated and plugged into econometrically derived relationships. Additional parameters are estimated in a second step. Other approaches start with partially specified models and then use historical time-series evidence to impose rational expectations without fully solving for the dynamic equilibrium. Econometric support for this approach was provided in my 1982 paper with an initial application in the *JPE* (Hansen and Hodrick 1980). This second paper

[1] Sims (1980) and others advanced this idea by developing tractable multivariate time-series methods and engaging in the identification of interpretable shocks in the multivariate setting.

[2] See Hansen (2014) for more discussions of modeling challenges for econometricians and economic agents inside the models that they build.

added a new perspective to the empirical link between forward and spot exchange rates.

III. Consumption and Permanent Income

Friedman's (1957) famed permanent income model has implications for both macroeconomic time series and microeconomic cross-sectional data. Its rational expectations counterpart is perhaps most simply depicted with a quadratic utility function, uncertain labor income, and a subjective rate of discount equal to the rate of return on assets. Insights have broader implications, but in this simplest setup, consumption is a martingale. This observation was featured by Hall (1978) in his well-known *JPE* paper on consumption and income dynamics. Permanent income theory in this guise illustrates how even transient implications for income can have permanent consequences for consumption while maintaining Friedman's basic insight that the permanent shocks to income are absorbed much more prominently into the consumption responses. The impact of the transient shocks is mitigated through savings behavior.

The Hall approach is a stark example of a partially specified model exploiting rational expectations. The martingale implications for consumption can be tested, as was done by Hall (1978), without having to specify correctly the income dynamics. Flavin (1983), also published in the *JPE*, completed the model specification and discussed the implied cross-equation restrictions of the type featured in Sargent (1981) to represent the excess sensitivity of consumption to transitory income. To derive the cross-equation restrictions as implied by a rational expectations equilibrium requires specifying the information about income used by consumers. For instance, information other than lagged income could be pertinent in predicting future income suggesting that the correct equilibrium may include other state variables. Testing the predictability of the first difference of consumption, however, does not require this complete specification. In this Hall-Flavin setup, the first difference of consumption reveals a news component in the information set of consumers (abstracting from measurement error). As Hansen, Roberds, and Sargent (1991) emphasized, this news component should be present-value neutral and offset by future income responses to this same shock. This gives a testable restriction on the corresponding impulse response function of income to the consumption news.

IV. Consumption and Asset Pricing

While the original Hall-Flavin research featured aggregate (and micro) implications with constant interest rates, the *JPE* published a variety of

papers that explored the empirical challenges that allowed for time variation in these rates. In addition, this literature built links between the macroeconomy and asset pricing with the aim of explaining empirical heterogeneity in the cross section of financial returns. As an outcome of this research, macroeconomists have featured the so-called equity premium, the observed gap between expected aggregate equity returns and Treasury bill returns, but the observed heterogeneity is much more pervasive. The Hall (1978) style reasoning turned out to be directly extendable to "Euler equation" representations of multiple assets, not just bonds and aggregate equity returns. Such representations support the equilibrium representation of asset prices using so-called stochastic discount factors that both discount the future and adjust for risk whereby the stochastic discount factors are explicitly linked to the macroeconomy through variables such as consumption. This stochastic discount factor approach provided a platform for empirical analysis. The conceptual underpinnings for this line of research were supported by theoretical derivations in Rubinstein (1976), Lucas (1978), and Breeden (1979).

The *JPE* published several important papers that explored empirical evidence related to this research. Hansen and Singleton (1983) used a linear time model to depict the implied linkages between consumption and returns. It featured restrictions across the predictable component of the time series and could accommodate a small cross section of returns. The linear time series approach had nice pedagogical value, making the overidentifying restrictions transparent, but it required a lognormal assumption without any scope for stochastic volatility. This linear time series approach was in contrast to the approach used by Hansen and Singleton (1982), who avoided the distributional assumption by studying alternative conditional moment restrictions. Both papers allowed for econometricians to understate the information used by economic agents. Moreover, both papers are among a collection of papers that document the empirical challenge posed by a representative or stand-in consumer model with time-separable power utility preferences as was commonly used in the macroeconomics literature. The power utility specification led to a stochastic discount factor that was a simple function of consumption growth. While many refer to this as the equity premium puzzle, it really is a more general phenomenon pertaining to the pricing of a heterogeneous cross section of returns.

In a later *JPE* paper, Hansen and Jagannathan (1991) provided a further characterization of the puzzle by stripping away the parametric structure of the stochastic discount factor. In the absence of arbitrage, there exist valid stochastic discount factors; however, they may possess different properties than what are implied by models with more parametric structure. Allowing for a much larger class of stochastic discount factors eliminated the possibility of fully identifying the stochastic discount factor

process from data and changed the econometric challenge to characterizing the set of potential stochastic discount factors that are consistent with empirical evidence. Specifically, Hansen and Jagannathan derived sharp bounds on the implied mean–standard deviation trade-off for stochastic discount factors that are consistent with the evidence from financial markets. Subsequent research extended and refined this analysis in a variety of ways. Empirical puzzles are well defined only relative to a family of models, and the bounds in the Hansen and Jagannathan paper and its extensions provided a more general way to pose puzzles with the aim of suggesting what is needed to construct models with better empirical underpinnings.

In response in part to the empirical challenges, the *JPE* has published several innovative papers that explored different specifications of investor preferences. For instance, Constantinides (1990) built a fully specified model in which investors have preferences that display habit persistence. Investors' period utilities depend not only on current-period consumption but also on that consumption relative to a habit stock of past consumptions. In effect, the habit stock provides a reference point for current consumptions.[3] Campbell and Cochrane (1999) altered these preferences in two ways. They featured a model in which the habit stock contributes a socially determined reference point based on past social consumptions. In addition, the counterpart to the habit stock has a nonlinear evolution equation. The Campbell and Cochrane paper, in particular, featured a model in which the market compensation for the exposure to macroeconomic risk is larger in bad macroeconomic times than in good ones. They provided an endogenous mechanism for this variation. While stylized, their analysis was supported by some empirical evidence, much more so than its counterpart with a power utility function. Others have extended and refined this as an empirically relevant asset pricing model.

A different strand of empirical research explored an alternative specification of investor preferences based on a recursive utility formulation. Such preferences, by design, feature investor concerns about the intertemporal composition of risk. This research built on theoretical underpinnings provided in Kreps and Porteus (1978) and Epstein and Zin (1989) and was prominently represented in two important *JPE* papers: Epstein and Zin (1991) and Campbell (1996). The stochastic discount factor in recursive utility models depends on the next-period continuation value relative to a risk-adjusted counterpart. This continuation value, familiar from recursive methods in economic dynamics, encodes investor perceptions about the future consumption prospects. Thus recursive utility preferences bring in a forward-looking contribution into the valuation of even short-

[3] Becker and Murphy (1988), also published in the *JPE*, used a similar formulation in a microeconomic analysis of "rational addiction."

term returns. The empirically oriented Epstein and Zin and Campbell papers accommodated this forward-looking perspective in different ways. The Epstein and Zin (1991) research followed an econometric approach similar to that in Hansen and Singleton (1982) modified by using a clever measurement scheme. Under their parametric specification, the return on wealth reveals the relevant information about the continuation value contribution to the stochastic discount factor. Campbell (1996) used a time-series formulation with forward-looking restrictions of a type that is common in linear rational expectations models but applied to financial variables. Campbell cleverly avoided using consumption data and instead featured the time-series properties of the market return, including its predictability. One reason to avoid using aggregate consumption data, as in Campbell, is that only a limited segment of the population participates in security markets. There has been a variety of subsequent empirical work that has built on these initial empirical contributions and their insights. Many of the resulting papers have demonstrated that the forward-looking channel added by recursive utility could have an important impact in asset pricing. Bansal and Yaron (2004) is a prominent example. In a related *JPE* contribution, Hansen, Heaton, and Li (2008), like Campbell, used linear time-series methods and rational expectations restrictions. Specifically, Hansen et al. characterized and measured long-term risk components that are simultaneously in the macroeconomic time series and in the cash flows from broad-based portfolios of equities.

V. Imperfections in Financial Markets

Markets are not fully complete, and there are limits as to how much risk they can share. The presumed market structure also alters the predicted equilibrium pricing of financial securities. For instance, suppose that consumers/investors face idiosyncratic components to labor income risk that cannot be fully diversified in financial markets. Two prominent examples of papers that took this as a starting point are Krusell and Smith (1998) and Constantinides and Duffie (1996). Both were published in the *JPE*. Krusell and Smith featured dynamic models for which the impact of market incompleteness was relatively benign in the sense that simple averages could be used to summarize distributional impacts for representing the evolution of the macroeconomy. In contrast, Constantinides and Duffie featured models in which idiosyncratic shocks to labor income have permanent components. Moreover, they presumed that there are macroeconomic impacts on the distributions of these idiosyncratic shocks. In their model, the equilibrium stochastic discount factor inherits these macroeconomic impacts. Both papers have interesting benchmark economies, and their contributions have had a remarkable impact on subsequent

research. In a *JPE* paper related to Constantinides and Duffie (1996), Heaton and Lucas (1996) probed into the microeconomic evidence and explored the quantitative implications of market incompleteness for asset pricing.

One rationale for why financial markets cannot fully diversify labor income risks is that idiosyncratic shocks are private information. The Kocherlakota and Pistaferri (2009) *JPE* paper took this perspective and presumed that the observed cross-sectional allocations are Pareto optimal after taking account of the private information. They derived the corresponding asset pricing implications and contrasted them with the ones implied by the incomplete market formulation of Constantinides and Duffie (1996) and others. Different attributes of the cross-sectional distribution of shocks come into play for the private information economy. Kocherlakota and Pistaferri exposed some of the resulting measurement challenges for asset pricing.

VI. Conclusion

Journal of Political Economy publications have played a prominent role in the study of macroeconomics and finance using time-series methods. The research disseminated by this journal delivered on Frisch's (1933) and others' ambition to use economic dynamic models to interpret time-series evidence. The published research characterized empirical challenges and explored implications of new models designed to confront these challenges.

References

Bansal, Ravi, and Amir Yaron. 2004. "Risks for the Long Run: A Potential Resolution of Asset Pricing Puzzles." *J. Finance* 59 (4): 1481–1509.

Becker, Gary S., and Kevin M. Murphy. 1988. "A Theory of Rational Addiction." *J.P.E.* 96 (4): 675–700.

Breeden, Douglas T. 1979. "An Intertemporal Asset Pricing Model with Stochastic Consumption and Investment Opportunities." *J. Financial Econ.* 7 (3): 265–96.

Campbell, John Y. 1996. "Understanding Risk and Return." *J.P.E.* 104 (2): 298–345.

Campbell, John Y., and John H. Cochrane. 1999. "By Force of Habit: A Consumption-Based Explanation of Aggregate Stock Market Behavior." *J.P.E.* 107 (2): 205–51.

Constantinides, George M. 1990. "Habit Formation: A Resolution of the Equity Premium Puzzle." *J.P.E.* 98 (3): 519–43.

Constantinides, George M., and Darrell Duffie. 1996. "Asset Pricing with Heterogeneous Consumers." *J.P.E.* 104 (2): 219–40.

Epstein, Larry G., and Stanley E. Zin. 1989. "Substitution, Risk Aversion and the Temporal Behavior of Consumption and Asset Returns: A Theoretical Framework." *Econometrica* 57 (4): 937–69.

———. 1991. "Substitution, Risk Aversion, and the Temporal Behavior of Consumption and Asset Returns: An Empirical Analysis." *J.P.E.* 99 (2): 263–86.
Flavin, Marjorie A. 1983. "Excess Volatility in the Financial Markets: A Reassessment of the Empirical Evidence." *J.P.E.* 91 (6): 929–56.
Friedman, Milton. 1957. "The Permanent Income Hypothesis." In *A Theory of the Consumption Function*, 20–37. Princeton, NJ: Princeton Univ. Press.
Frisch, Ragnar. 1933. "Propagation Problems and Impulse Problems in Dynamic Economics." In *Economic Essays in Honour of Gustav Cassel*, 171–205. London: Allen & Unwin.
Hall, Robert E. 1978. "Stochastic Implications of the Life Cycle–Permanent Income Hypothesis: Theory and Evidence." *J.P.E.* 86 (6): 971–87.
Hansen, Lars Peter. 1982. "Large Sample Properties of Generalized Method of Moments Estimators." *Econometrica* 50 (4): 1029–54.
———. 2014. "Nobel Lecture: Uncertainty Outside and Inside Economic Models." *J.P.E.* 122 (5): 945–87.
Hansen, Lars Peter, John C. Heaton, and Nan Li. 2008. "Consumption Strikes Back? Measuring Long-Run Risk." *J.P.E.* 116 (2): 260–302.
Hansen, Lars Peter, and Robert J. Hodrick. 1980. "Forward Exchange Rates as Optimal Predictors of Future Spot Rates: An Econometric Analysis." *J.P.E.* 88 (5): 829–53.
Hansen, Lars Peter, and Ravi Jagannathan. 1991. "Implications of Security Market Data for Models of Dynamic Economies." *J.P.E.* 99 (2): 225–62.
Hansen, Lars Peter, William Roberds, and Thomas J. Sargent. 1991. "Observable Implications of Present-Value-Budget Balance." In *Rational Expectations Econometrics*. Boulder, CO: Westview.
Hansen, Lars Peter, and Kenneth J. Singleton. 1982. "Generalized Instrumental Variables Estimation of Nonlinear Rational Expectations Models." *Econometrica* 50:1269–86.
———. 1983. "Stochastic Consumption, Risk Aversion, and the Temporal Behavior of Asset Returns." *J.P.E.* 91 (2): 249–65.
Heaton, John, and Deborah J. Lucas. 1996. "Evaluating the Effects of Incomplete Markets on Risk Sharing and Asset Pricing." *J.P.E.* 104 (3): 443–87.
Kocherlakota, Narayana, and Luigi Pistaferri. 2009. "Asset Pricing Implications of Pareto Optimality with Private Information." *J.P.E.* 117 (3): 555–90.
Kreps, David M., and Evan L. Porteus. 1978. "Temporal Resolution of Uncertainty and Dynamic Choice." *Econometrica* 46 (1): 185–200.
Krusell, Per, and Anthony A. Smith. 1998. "Income and Wealth Heterogeneity in the Macroeconomy." *J.P.E.* 106 (5): 867–96.
Lucas, Robert E. 1972a. "Econometric Testing of the Natural Rate Hypothesis." In *The Econometrics of Price Determination*, edited by O. Eckstein, 50–59. Washington, DC: Board Governors, Fed. Reserve System.
———. 1972b. "Expectations and the Neutrality of Money." *J. Econ. Theory* 4 (2): 103–24.
———. 1978. "Asset Prices in an Exchange Economy." *Econometrica* 46 (6): 1429–45.
Muth, John H. 1961. "Rational Expectations and the Theory of Price Movements." *Econometrica* 29 (3): 315–35.
Rubinstein, Mark. 1976. "The Valuation of Uncertain Income Streams and the Pricing of Options." *Bell J. Econ.* 7:407–25.
Sargent, Thomas J. 1981. "Interpreting Economic Time Series." *J.P.E.* 89 (2): 213–48.
Sims, Christopher A. 1980. "Macroeconomics and Reality." *Econometrica* 48 (1): 1–48.

Slutsky, Eugen. 1927. "The Summation of Random Causes as the Source of Cyclic Processes." In *Problems of Economic Conditions*, vol. 3. Moscow: Conjuncture Inst. Reprinted in *Econometrica* 5 (April 1937): 105–46.

Stigler, Stephen M. 1986. *The History of Statistics: The Measurement of Uncertainty before 1900*. Cambridge, MA: Harvard Univ. Press.

Yule, George Udny. 1927. "On a Method of Investigating Periodicities in Disturbed Series, with Special Reference to Wolfer's Sunspot Numbers." *Philosophical Trans. Royal Soc. London. Ser. A, Containing Papers of a Mathematical or Physical Character* 226:267–98.

Asset Pricing: Models and Empirical Evidence

George M. Constantinides

University of Chicago

Hansen and Singleton (1982, 1983), Shiller (1982), Mehra and Prescott (1985), and Weil (1989) pose a major challenge in economics in the context of a Lucas (1978) exchange economy. Hansen and Singleton (1982) reject the Euler equations of per capita consumption at any level of relative risk aversion (RRA). Mehra and Prescott (1985) show that the average premium of the stock market over the risk-free rate cannot be rationalized in a calibrated standard economy and coin the "equity premium puzzle." Mehra and Prescott (1985) and Weil (1989) further show that the puzzle is a dual one: as the assumed RRA is increased to rationalize the equity premium, the implied risk-free rate becomes too high. More generally, the challenge is to simultaneously explain the moments of aggregate consumption and dividend growth, risk-free rate, market return, market price-dividend ratio, and the term structure of interest rates in the context of an economy with rational economic agents. The voluminous research effort to address this challenge continues to this day and includes explorations of preferences for early resolution of uncertainty, absence of complete consumption insurance, uncertainty about the state of the economy, habit persistence, macroeconomic crises resulting in a catastrophic drop in consumption, uncertainty about the economic model and its parameters, borrowing constraints, and deviations from rationality. In this essay I describe some of these explorations without providing an exhaustive review of the literature. My apologies to authors who are not being cited here.

Kreps and Porteus (1978) introduce a class of non–von Neumann–Morgenstern preferences that capture preference for early resolution of uncertainty. These preferences are sensitive to low-frequency but persistent innovations in the state variables, thereby addressing many asset pricing puzzles. Epstein and Zin (1989, 1991) adopt a conveniently homothetic form of these preferences (commonly referred to as "Epstein-Zin" preferences) and point out that these preferences disentangle the RRA coefficient and the elasticity of intertemporal substitution (EIS) in consumption. These preferences have the potential to resolve the dual equity premium and risk-free rate puzzles because the equity premium is driven by the RRA coefficient while the risk-free rate is driven by the EIS. Furthermore, they obtain in closed form the marginal rate of substitution (MRS), thereby making these preferences tractable. Hansen, Heaton, and Li (2008) elucidate properties of these preferences.

Bansal and Yaron (2004) are the first to adopt the Epstein-Zin preferences to address the asset pricing agenda. They build a model in which the mean and variance of consumption and dividend growth depend on a state variable with low frequency but persistent innovations and successfully address many targets of the asset pricing agenda. Constantinides and Ghosh (2011) and Beeler and Campbell (2012) point out that the model implies too high autocorrelation of the aggregate consumption growth rate and, therefore, excessive predictability of consumption growth by the price-dividend ratio. Despite these limitations, the coupling of preference for early resolution of uncertainty with low-frequency but persistent innovations in state variables is a significant breakthrough that will continue to influence research in the foreseeable future. Two models discussed next, one on uninsurable household income shocks and the other on learning about the state of the economy, owe an intellectual debt to Epstein-Zin preferences and Bansal and Yaron (2004).

Cochrane (1991), Attanasio and Davis (1996), Blundell, Pistaferri, and Preston (2008), and others provide empirical evidence that consumption insurance is incomplete: households face a substantial amount of uninsurable idiosyncratic labor income risk. Constantinides (1982) highlights the pivotal role of complete consumption insurance, showing that the equilibrium of such an economy with households that have heterogeneous endowments and von Neumann–Morgenstern preferences is isomorphic to the equilibrium of a homogeneous-household economy. Mankiw (1986) shows that, in a two-period economy with incomplete consumption insurance, the concentration of aggregate shocks among the population is an important determinant of the level of the equity premium. Constantinides and Duffie (1996) further show that, in the absence of complete consumption insurance, given the aggregate income and dividend processes, any given (arbitrage-free) price process can be supported in the equilibrium

of a heterogeneous-household economy with judiciously chosen persistent and countercyclical idiosyncratic income shocks.

Brav, Constantinides, and Geczy (2002) present empirical evidence that the equity and value premia are consistent with a stochastic discount factor (SDF) obtained as the average of individual households' MRS with low and economically plausible values of the RRA coefficient. Since these premia are not explained with an SDF obtained as the per capita MRS with low values of the RRA coefficient, the evidence supports the incomplete consumption insurance hypothesis. They further show that the countercyclical skewness of the idiosyncratic income shocks plays a key role in driving prices, a property confirmed by Guvenen, Ozkan, and Song (2014) using a very large data set from the US Social Security Administration.

Being framed in terms of economies in which households are endowed with power utility, neither of these papers allows the RRA coefficient and the EIS to be disentangled, a step that appears to be important in addressing the level and time-series properties of the risk-free rate, price-dividend ratio, and market return. By introducing recursive preferences, Constantinides and Ghosh (2017) provide empirical evidence that negatively skewed, persistent, and countercyclical household consumption shocks explain the moments of aggregate consumption and dividend growth, risk-free rate, aggregate market return, and market price-dividend ratio and explain the cross section of size-sorted, book-to-market-equity-sorted, and industry-sorted portfolio returns.

The second model that owes an intellectual debt to Epstein-Zin preferences and Bansal and Yaron (2004) is a model of uncertainty about the state of the economy. An overload of worldwide macroeconomic, business, and political news inundates investors. Little is known as to how investors cope with this vast amount of information and, in particular, which subset of information they pay attention to. In the macro-finance literature, researchers typically model investors as focusing on the histories of a limited number of macroeconomic variables, typically consumption and GDP growth, and applying a filter to rationally extract relevant information about the economy. These models fare poorly in explaining several features of stock market data, including the high average level of the equity premium, the low level of the risk-free rate, the high variability of the price-dividend ratio, and the low predictability of consumption growth by the price-dividend ratio.

Ghosh and Constantinides (2017) establish that two broad categories of publicly available macroeconomic information are the most highly correlated with the marketwide price-dividend ratio. The first category consists of price levels, including the Consumer Price Index for All Urban Consumers (CPI-U) and the Producer Price Index. The second category consists of labor market variables, including average hourly earn-

ings, average hours of production, and numbers of employees in private nonfarm payrolls in different sectors. These are the two classes of macro variables that, according to FactSet, Bloomberg users pay the most attention to. On the other hand, contrary to the implications of learning models in which investors are assumed to learn from the consumption and GDP histories alone, the price-dividend ratio has negligible correlation with the contemporaneous consumption and GDP growth or a weighted average of current and lagged consumption and GDP growth rates.

Motivated by the above evidence, Ghosh and Constantinides (2017) model investors as learning about the latent state of the economy from either the CPI or earnings per hour histories. The model provides a good fit to the sample moments of consumption and dividend growth, market return, marketwide price-dividend ratio, and risk-free rate. In contrast, an alternative nested model in which the investors learn from the consumption history alone fails along a number of dimensions: it implies essentially zero volatility of the price-dividend ratio, thereby failing to explain the excess volatility puzzle; it fails to generate the high persistence in the marketwide price-dividend ratio—one of the most robust features observed in the data; and the estimated consumption growth in the second regime is −5.7 percent, something that has not been observed in US history even during the Great Depression of 1929.

The high persistence in the beliefs process in the main model, combined with the preference for early resolution of uncertainty, yields a high equity premium and low risk-free rate, consistent with the data. In contrast, in the alternative model the low persistence of the beliefs process yields a low equity premium. Finally, consistent with the data, the main model generates strong time variation in the conditional mean and variance of the market return. Perhaps more impressive is the observation that it does so without relying on countercyclical heteroscedasticity of the consumption growth rate or the additional signal (the volatilities of consumption growth and the signal are set to be equal in the two states)—a phenomenon for which there is limited empirical evidence. Instead, the model generates time variation in the conditional moments of the market return from the heteroscedasticity of the beliefs process.

Early papers that recognize habit persistence through non-time-separable von Neumann–Morgenstern preferences include Marshall (1920), Duesenberry (1949), Pollak (1970), and Ryder and Heal (1973). Habit preferences have had some success in addressing the asset pricing agenda because habit tracks the business cycle. In a recession consumption is low relative to habit (average past consumption), the equity premium is high because of the induced high RRA, and the risk-free rate is low because of the precautionary demand for savings.

Constantinides (1990) models an economy in which the preferences of the representative consumer exhibit linear internal habit; that is, the

consumer takes into account the effect of current consumption on future habit. The calibrated model resolves the equity premium and risk-free rate puzzles with a low RRA coefficient and a low EIS while matching the mean and variance of consumption growth, albeit implying higher autocorrelation of aggregate consumption growth than the autocorrelation observed in the data. Ferson and Constantinides (1991) do not reject the model on a system of assets consisting of size-sorted equity portfolios, a Treasury bill, and a Treasury bond.

Campbell and Cochrane (1999) modify the Constantinides (1990) model in two ways. First, they model the habit as nonlinear, thereby rendering the interest rate volatility low or zero even when consumption growth is serially uncorrelated. Second, they model habit as external; that is, the consumer does not take into account the effect of current consumption on future habit. The calibrated model resolves the equity premium and risk-free rate puzzles while matching the mean, variance, and autocorrelation of consumption growth. However, the assumption that habit is external renders the RRA coefficient implausibly high and fluctuating between 60 and a few hundred. Ljungqvist and Uhlig (2015) point out that government interventions that occasionally destroy part of the aggregate endowment can lead to substantial welfare improvements.

In both the Constantinides (1990) and Campbell and Cochrane (1999) models the only source of innovation is consumption growth, and these models counterfactually imply that the correlation between the market price-dividend ratio and consumption growth is about 50 percent, unlike the correlation of close to zero in the data. Nevertheless, models with habit preferences and additional sources of innovation are free from this criticism. Chen and Ludvigson (2009) compare the Constantinides (1990) and Campbell and Cochrane (1999) models and conclude that habit is better described as nonlinear and internal rather than linear and external, and the model performs well in explaining the cross section of equity returns. Havranek, Rusnak, and Sokolova (2017) collect and examine estimates of habit from a wide range of studies and establish that habit is significant but varies widely across studies.

Rietz (1988) points out that the possibility of a macroeconomic crisis resulting in a catastrophic drop in consumption in principle resolves the equity premium puzzle because the MRS becomes very high at the catastrophic state, adding the caveat that the size of the required drop in consumption has never been observed in US history. Barro (2006), Barro and Ursùa (2008), Gabaix (2012), Nakamura et al. (2013), Wachter (2013), and others argue that rare disasters explain the equity premium and related puzzles. Barro (2006) and Barro and Ursùa (2008) present domestic and international evidence that macroeconomic crises are associated with a large and sustained drop in aggregate consumption. As Constantinides (2008) points out, Barro's calibrated model treats the peak-to-trough

decline in aggregate consumption during macroeconomic crises (which last, on average, 4 years) as if this decline occurs in 1 year, thereby magnifying by a factor of four the size of the observed annual disaster risk. Similar ad hoc magnification of the annual aggregate consumption decline during macroeconomic crises is employed in a number of papers that follow Barro (2006). Using an econometric methodology that allows the probabilities attached to different states of the world to differ from their sample frequencies and is therefore robust to the rare events problem in the data, Julliard and Ghosh (2012) reject the rare events explanation for the equity premium puzzle. They show that to explain the equity premium puzzle one should be willing to believe that economic disasters occur every 6.6 years, on average. Moreover, Backus, Chernov, and Martin (2011) demonstrate that options imply smaller probabilities of extreme outcomes than the probabilities estimated from international macroeconomic data.

Recent literature by Hansen and Sargent (2001), Epstein and Schneider (2003), Klibanoff, Marinacci, and Mukerji (2005), Maccheroni, Marinacci, and Rustichini (2006), and Johannes, Lochstoer, and Mou (2016), among others, addresses the uncertainty about the economic model and its parameters. This literature argues that a more challenging high-dimensional learning problem confronts investors where they need to learn not only about the current state but also about the true underlying model and its parameters and that such a learning problem plays an important role in enhancing the empirical performance of these models. This important class of issues is discussed in the essay by Lars Hansen, a major contributor to this literature.

Borrowing constraints address the equity premium and risk-free rate puzzles and provide a partial explanation for the limited participation of young consumers in the stock market and the demand for bonds in the context of households in different stages of their life cycle as young, middle-aged, and old. Constantinides, Donaldson, and Mehra (2002) consider an overlapping generations economy in which consumers live for three periods. In the first period, a period of human capital accumulation, consumers receive a relatively low income. In the second period, consumers are employed and receive income subject to great uncertainty. The consumers consume part of this income and save the rest by investing in stocks and bonds. In the third period, consumers consume the assets accumulated during the second period. The key feature is that the bulk of the future income of young consumers comes from wages that they earn during the second period, while the income of the elderly primarily comes from their savings in stocks and bonds during their middle age.

Young people would like to invest in equities, given the observed high equity premium. However, they are reluctant to reduce their current consumption in order to save by investing in stocks, because the bulk of their

lifetime income comes from their wages in their middle age. They want to borrow against their future income, but the borrowing constraints prevent them from doing so. Human capital alone cannot be used as collateral for large loans in modern economies for reasons of moral hazard and adverse selection. The model explains why many consumers do not participate in the stock market when they are young. Middle-aged consumers earn income that they partly consume and partly save by purchasing equities and bonds. The old earn no income and consume their savings. Therefore, the risk of stock and bond ownership is concentrated in the hands of middle-aged consumers who save. This concentration of risk generates the high equity premium and the demand for bonds, in addition to the demand for shares by the middle-aged. The model acknowledges and addresses at the same time the issue of the limited participation in the stock market and the demand for bonds.

References

Attanasio, Orazio P., and Steven J. Davis. 1996. "Relative Wage Movements and the Distribution of Consumption." *J.P.E.* 104:1227–62.

Backus, David, Mikhail Chernov, and Ian Martin. 2011. "Disasters Implied by Equity Index Options." *J. Finance* 66:1967–2009.

Bansal, Ravi, and Amir Yaron. 2004. "Risks for the Long Run: A Potential Resolution of Asset Pricing Puzzles." *J. Finance* 59:1481–1509.

Barro, Robert J. 2006. "Rare Disasters and Asset Markets in the 20th Century." *Q.J.E.* 121:823–66.

Barro, Robert J., and José F. Ursùa. 2008. "Macroeconomic Crises since 1870." *Brookings Papers Econ. Activity* (Spring): 255–335.

Beeler, Jason, and John Y. Campbell. 2012. "The Long-Run Risks Model and Aggregate Asset Prices: An Empirical Assessment." *Critical Finance Rev.* 1:141–82.

Blundell, Richard, Luigi Pistaferri, and Ian Preston. 2008. "Consumption Inequality and Partial Insurance." *A.E.R.* 98:1887–1921.

Brav, Alon, George M. Constantinides, and Christopher C. Geczy. 2002. "Asset Pricing with Heterogeneous Consumers and Limited Participation: Empirical Evidence." *J.P.E.* 110:793–824.

Campbell, John Y., and John H. Cochrane. 1999. "By Force of Habit: A Consumption-Based Explanation of Aggregate Stock Market Behavior." *J.P.E.* 107:205–51.

Chen, Xiaohong, and Sydney C. Ludvigson. 2009. "Land of Addicts? An Empirical Investigation of Habit-Based Asset Pricing Models." *J. Appl. Econometrics* 24:1057–93.

Cochrane, John H. 1991. "A Simple Test of Consumption Insurance." *J.P.E.* 99: 957–76.

Constantinides, George M. 1982. "Intertemporal Asset Pricing with Heterogeneous Consumers and without Demand Aggregation." *J. Bus.* 55:253–67.

———. 1990. "Habit Formation: A Resolution of the Equity Premium Puzzle." *J.P.E.* 98:519–43.

———. 2008. "Comment on Barro and Ursùa." *Brookings Papers Econ. Activity* (Spring): 341–50.

Constantinides, George M., John B. Donaldson, and Rajnish Mehra. 2002. "Junior Can't Borrow: A New Perspective on the Equity Premium Puzzle." *Q.J.E.* 117:269–96.

Constantinides, George M., and Darrell Duffie. 1996. "Asset Pricing with Heterogeneous Consumers." *J.P.E.* 104:219–40.
Constantinides, George M., and Anisha Ghosh. 2011. "Asset Pricing Tests with Long-Run Risks in Consumption Growth." *Rev. Asset Pricing Studies* 1:96–136.
———. 2017. "Asset Pricing with Countercyclical Household Consumption Risk." *J. Finance* 73:415–59.
Duesenberry, James S. 1949. *Income, Saving, and the Theory of Investment Behavior*. Cambridge, MA: Harvard Univ. Press.
Epstein, Larry G., and Martin Schneider. 2003. "Recursive Multiple-Priors." *J. Econ. Theory* 113:1–31.
Epstein, Larry G., and Stanley Zin. 1989. "Substitution, Risk Aversion, and the Temporal Behavior of Consumption and Asset Returns: A Theoretical Framework." *Econometrica* 57:937–69.
———. 1991. "Substitution, Risk Aversion, and the Temporal Behavior of Consumption and Asset Returns: An Empirical Analysis." *J.P.E.* 99:263–86.
Ferson, Wayne E., and George M. Constantinides. 1991. "Habit Persistence and Durability in Aggregate Consumption: Empirical Tests." *J. Financial Econ.* 29:199–240.
Gabaix, Xavier. 2012. "Variable Rare Disasters: An Exactly Solved Framework for Ten Puzzles in Macro-Finance." *Q.J.E.* 127:645–700.
Ghosh, Anisha, and George M. Constantinides. 2017. "What Information Drives Asset Prices." Working paper, Univ. Chicago.
Guvenen, Fatih, Serdar Ozkan, and Jae Song. 2014. "The Nature of Countercyclical Income Risk." *J.P.E.* 122:621–60.
Hansen, Lars Peter, John Heaton, and Nan Li. 2008. "Consumption Strikes Back: Measuring Long-Run Risk." *J.P.E.* 116:260–302.
Hansen, Lars Peter, and Thomas J. Sargent. 2001. "Robust Control and Model Uncertainty." *A.E.R.* 91:60–66.
Hansen, Lars Peter, and Kenneth J. Singleton. 1982. "Generalized Instrumental Variables Estimation of Nonlinear Rational Expectations Models." *Econometrica* 50:1269–86.
———. 1983. "Stochastic Consumption, Risk Aversion, and the Temporal Behavior of Asset Returns." *J.P.E.* 91:249–65.
Havranek, Tomas, Marek Rusnak, and Anna Sokolova. 2017. "Habit Formation in Consumption: A Meta-Analysis." *European Econ. Rev.* 95:142–67.
Johannes, Michael, Lars Lochstoer, and Yiqun Mou. 2016. "Learning about Consumption Dynamics." *J. Finance* 71:551–600.
Julliard, Christian, and Anisha Ghosh. 2012. "Can Rare Events Explain the Equity Premium Puzzle?" *Rev. Financial Studies* 25:3037–76.
Klibanoff, Peter, Massimo Marinacci, and Sujoy Mukerji. 2005. "A Smooth Model of Decision Making under Ambiguity." *Econometrica* 73:1849–92.
Kreps, David M., and Evan L. Porteus. 1978. "Temporal Resolution of Uncertainty and Dynamic Choice Theory." *Econometrica* 46:185–200.
Ljungqvist, Lars, and Harald Uhlig. 2015. "Comment on the Campbell-Cochrane Habit Model." *J.P.E.* 123:1201–13.
Lucas, Robert E., Jr. 1978. "Asset Prices in an Exchange Economy." *Econometrica* 46:1429–45.
Maccheroni, Fabio, Massimo Marinacci, and Aldo Rustichini. 2006. "Ambiguity Aversion, Robustness, and the Variational Representation of Preferences." *Econometrica* 74:1447–98.
Mankiw, N. Gregory. 1986. "The Equity Premium and the Concentration of Aggregate Shocks." *J. Financial Econ.* 17:211–19.
Marshall, Alfred. 1920. *Principles of Economics: An Introductory Volume*. 8th ed. London: Macmillan.

Mehra, Rajnish, and Edward C. Prescott. 1985. "The Equity Premium: A Puzzle." *J. Monetary Econ.* 15:145–61.
Nakamura, Emi, Jón Steinsson, Robert J. Barro, and José F. Ursùa. 2013. "Crises and Recoveries in an Empirical Model of Consumption Disasters." *American Econ. J.: Macroeconomics* 5:35–74.
Pollak, Robert A. 1970. "Habit Formation and Dynamic Demand Functions." *J.P.E.* 78:745–63.
Rietz, Thomas A. 1988. "The Equity Risk Premium: A Solution." *J. Monetary Econ.* 22:117–31.
Ryder, Harl E., Jr., and Geoffrey M. Heal. 1973. "Optimum Growth with Intertemporal Dependent Preferences." *Rev. Econ. Studies* 40:1–33.
Shiller, Robert J. 1982. "Consumption, Asset Markets and Macroeconomic Fluctuations." Working paper, NBER, Cambridge, MA.
Wachter, Jessica A. 2013. "Can Time-Varying Risk of Rare Disasters Explain Aggregate Stock Market Volatility?" *J. Finance* 68:987–1035.
Weil, Philippe. 1989. "The Equity Premium Puzzle and the Risk-Free Rate Puzzle." *J. Monetary Econ.* 24:401–21.

Finance at the University of Chicago

Eugene F. Fama

University of Chicago

Research in finance has two main areas: (i) corporate—theory and empirical work on optimal investment and financing decisions by firms (the demand side of capital formation)—and (ii) asset pricing—portfolio theory and related models of risk and expected return (the supply side of capital formation). The University of Chicago Booth School (formerly the Graduate School of Business) has long been front and center in corporate finance, and it is joined by the Chicago Economics Department as an asset pricing ringleader.

Corporate finance was kick-started by the "irrelevance of capital structure" theorems of Franco Modigliani and my longtime Chicago Booth mentor, colleague, and friend, Merton Miller (Modigliani and Miller 1958; Miller and Modigliani 1961; the latter published, like many seminal papers in finance, in the now defunct *Journal of Business*). I have

The comments of George Constantinides and many years of interaction on the topics of this paper with Kenneth French and John Cochrane are gratefully acknowledged

worked some in corporate finance, but more of my work is in asset pricing, and I largely focus here on Chicago's role in asset pricing research. I give my (business school) perspective, which has a strong eye toward applications. This is in contrast to the macro-finance perspective on asset pricing, which is more concerned with integrating asset pricing and macroeconomics. Lars Hansen (my 2013 co–Nobel laureate along with Robert Shiller) is a massive contributor to macro-finance. Aside from a few comments, I leave that area to him and George Constantinides in this anniversary edition of the *JPE*.

I emphasize that this is a perspective piece, biased toward Chicago's contributions, not a literature review meant to give everyone just due. For readers offended by my Chicago-centric overview, mea culpa in advance.

The asset pricing research I discuss has two main poles: (i) work on market efficiency—the proposition that asset prices reflect all available information—and (ii) models of market equilibrium—the nature of risk and the relation between risk and expected return. In my Nobel lecture (Fama 2014) I call these the Siamese twins of asset pricing because, as outlined below, they are inseparable pieces of asset pricing models based on rational behavior. I begin by discussing the work on market efficiency and then turn to models of market equilibrium.

Market Efficiency

Research on the behavior of commodity prices and prices of financial assets has a long history going back at least to Bachelier (1900), but the coming of computers in the late 1950s produced an explosion of work, primarily on the behavior of stock prices and stock returns. Interest was concentrated at the University of Chicago's Booth School and the Massachusetts Institute of Technology Economics Department and Sloan School. At Chicago, the early players were Larry Fisher (creator of the now-ubiquitous Center for Research in Security Prices [CRSP] data files), Merton Miller, Harry Roberts, and Lester Telser, with Benoit Mandelbrot as an occasional visitor. At MIT, Sydney Alexander, Paul Cootner, Franco Modigliani (Merton Miller's longtime coauthor), and Paul Samuelson carried the ball.

When I finished PhD prelims and it came time to write a thesis in 1962, I was twice a father and anxious to finish quickly. A PhD student could get faculty attention with a thesis on the behavior of stock prices, which explains Fama (1964), published in the *Journal of Business*. More of my thesis tests Mandelbrot's hypothesis that stock returns conform better to the nonnormal (fat-tailed) class of symmetric stable distributions than to the normal distribution, but about a third of it is on what I later dub market efficiency.

Market efficiency is the hypothesis that asset prices reflect all available information. But what are the testable implications? The early answer is that in an efficient market, prices follow random walks and future returns are unpredictable from currently available information. Mandelbrot (1966), another seminal paper in the *Journal of Business*, and Samuelson (1965) show that the random walk hypothesis is too strong. In their models, prices in an efficient market are submartingales. In simple terms, the presumption is that the price of an asset at time t, p_t, is set to deliver an equilibrium expected return at $t + 1$, $E(R_{t+1}|\phi_{t,m})$, where $\phi_{t,m}$ is the information embedded in the asset's time t price. If $\phi_{t,m}$ is all available information, ϕ_t, then

$$E(p_{t+1}|\phi_t) = p_t[1 + E(R_{t+1}|\phi_{t,m})], \tag{1}$$

and

$$E(R_{t+1}|\phi_t) - E(R_{t+1}|\phi_{t,m}) = 0.0. \tag{2}$$

Equation (2) implies that $R_{t+1} - E(R_t|\phi_{t,m})$, the deviation of the return from the equilibrium expected return, $E(R_t|\phi_{t,m})$, is unpredictable from information available at t. But (1) and (2) hold only when $\phi_{t,m} = \phi_t$, that is, when the information embedded in the price p_t is all available information. If some information is missed in setting prices at t, then $R_{t+1} - E(R_t|\phi_{t,m})$ is predictable from the broader information set ϕ_t.

Implicit in the submartingale models of Samuelson (1965) and Mandelbrot (1966) is what I call the joint hypothesis problem: tests of market efficiency are conditional on an assumed model for equilibrium prices and expected returns, which means that tests of efficiency are joint tests of efficiency and the assumed asset pricing model (Fama [1970], spelled out better in chap. 5 of Fama [1976]). Though not commonly acknowledged, the reverse is also true: asset pricing models based on rational behavior implicitly or explicitly assume a strong form of market efficiency: investors agree on the joint distributions of future asset payoffs and they get them right, which means that prices reflect all available information. In short, market efficiency and models of market equilibrium are the Siamese twins of asset pricing (Fama 2014).

The early work (1950s and 1960s) on market efficiency focuses on the time-series properties of stock returns. Worried that the newly minted CRSP data would not find their way into academic research, James Lorie, the founder of CRSP, suggested I do a study of the adjustment of stock prices to stock splits. The resulting paper, Fama et al. (1969) coauthored with Lawrence Fisher and two of the all-time best Chicago finance PhD students, Michael Jensen and Richard Roll, is the first event study. Event

studies subsequently play a major role in tests of market efficiency in the finance and accounting literatures, and they are often used to assess damages in court cases.

An important issue in the market efficiency literature is whether professional investors have private information about asset prospects that they use to enhance returns. The most complete data are for mutual funds, and there is a large literature on the performance of actively managed funds, that is, funds that attempt to enhance returns by choosing underpriced stocks. The seminal early study is Michael Jensen's (1968) University of Chicago PhD thesis. He uses the capital asset pricing model (CAPM) of Sharpe (1964) and Lintner (1965) as the model for equilibrium expected returns, and as in many studies thereafter, he finds that actively managed funds generally underperform the expected return predictions of the CAPM.

There is a large literature on mutual fund performance with many excellent papers. I cannot resist touting the one I like best, Fama and French (2010). In this paper, we use the stochastic properties of individual fund returns from 1984 to 2006 to simulate the cross-section distribution of risk-adjusted average returns when the market is efficient and active fund managers have no private information. The models for risk adjusting are the CAPM, the three-factor model of Fama and French (1993), and Carhart's (1997) four-factor extension. The results are striking. About 3 percent of actively managed funds seem to have enough private information to cover the costs (reported management fees and expenses) imposed on investors, which means that going forward, we expect them to perform like low-cost passive funds. The remaining 97 percent do not produce returns that suggest they have enough private information to cover costs. When we examine fund returns before costs, there is a near-perfect balance of funds in the extreme right tail of risk-adjusted average returns that do better than would be expected by chance and losers in the extreme left tail that do worse. The aggregate portfolio of active funds matches the overall US market return before costs, and its monthly return is correlated .99+ with the market return. In other words, in aggregate, active mutual funds are an expensive way to hold the market portfolio.

Though mutual funds are not the entire active investor universe, our mutual fund results are in line with Sharpe's (1991) observation that active investing is a zero-sum game before costs: winners eat losers. The logic is that an active investor holds an unbalanced portfolio, overweighting some assets relative to their market caps and underweighting others. This means that in aggregate, other investors have to take offsetting positions, underweighting the assets overweighted by the active investor and overweighting the assets underweighted. But passive investors hold cap-weight portfolios,

either of the entire market or of subsets (value stocks, growth stocks, energy stocks, etc.). Thus, the unbalanced portfolio of an active investor has to be balanced by offsetting positions of other active investors—a zero-sum game before costs. After costs, active investing is a negative-sum game. French (2008) provides estimates of the costs.

Asset Pricing

Finance as an area of scientific research had its birth at the University of Chicago. The parting shot is Harry Markowitz's Economics Department PhD thesis on portfolio theory, subsequently published as a journal article (Markowitz 1952) and then as a magnificent Cowles Foundation monograph (Markowitz 1959). When I started teaching investments in 1963, the textbooks of the day focused on the futile task of teaching students to pick stocks. In my early teaching years, Markowitz (1959) was the main reading in my investments course.

The mean-variance-efficient set of Markowitz's portfolio model is the foundation of the first formal asset pricing model, the CAPM of Sharpe (1964) and Lintner (1965). The CAPM provides the first rigorous analysis of asset risk and the equilibrium relation between risk and expected return. The CAPM is a one-period model. It gets a tour de force multiperiod extension in Robert Merton's (1973a) intertemporal CAPM, the ICAPM, which offers a theoretical framework for recent multifactor models, for example, the three-factor model of Fama and French (1993).

The CAPM predicts that market β, the slope in the regression of an asset's return on the market return, suffices to describe the cross section of expected asset returns. The initial tests of the model are cross-section regressions of average asset returns on estimates of their β's and other variables. The model predicts that the slopes for other variables are indistinguishable from zero. Black, Jensen, and Scholes (1972) argue that ordinary least squares standard errors for slope estimates from such regressions are too low because they do not adjust for cross-correlation of the regression residuals (return correlation beyond that associated with regression explanatory variables).

Fama and MacBeth (1973) provide a simple cure. Instead of cross-section regressions of average monthly returns on β estimates and other variables, the regressions are run month by month. Averages of monthly slopes and *t*-statistics for the averages are used to test the CAPM prediction that β suffices to describe expected asset returns. This approach is in effect repeated sampling in which the time-series variation in month-by-month regression slopes picks up the effects of cross-correlation of residuals without requiring an estimate of the residual covariance matrix. The approach has been used so much and so long in tests of asset pricing

models it is often referenced as Fama-MacBeth, without citation of the *JPE* source. The approach is generally applicable in panel regressions, balanced and unbalanced, when it is sensible to weight periods (rather than observations) equally.

Asset pricing, as represented by the CAPM and the ICAPM, is rather divorced from other areas of economics. This changes with the consumption-based CCAPM of Economics Department standout Robert Lucas (1978) (see also Breeden 1979). The CCAPM, with its simple but powerful economic insights, took macroeconomics by storm and gave birth to a large body of empirical work and theoretical extensions by the most talented macro-finance researchers. Excellent and much-cited *JPE* examples are Epstein and Zin (1991) and Campbell and Cochrane (1999). Lucas's Nobel citation is for his pathbreaking work in macro models with rational expectations, but on the basis of its impact on research in the intersection of macroeconomics and finance, the CCAPM is as important.

Work on asset pricing models is currently in a bit of a bind. The CAPM had a 25-year run when it was widely accepted as the ruling paradigm. The slow accumulation of anomalies—patterns in average returns that violate the model's predictions—led to the general conclusion that the model is an empirical failure. (The death knell is Fama and French [1992].) Despite the efforts of many talented macro-finance empiricists, the consumption-based CCAPM rests in empirical limbo, and we are unaware of anyone who suggests that it can be useful in applications (e.g., evaluating portfolio performance or determining the cost of capital).

The recent alternatives offered by asset pricing research in finance (as opposed to macro-finance) are factor models, which expand the CAPM by including factors beyond the CAPM market factor. The three-factor model of Fama and French (1993) had a 20+-year run, but in the face of accumulating anomalies, it now has a five-factor extension (Fama and French 2015). There are other examples. Merton's (1973a) ICAPM provides a ready-made theoretical framework that can accommodate factor-based asset pricing models. But until research identifies and empirically validates the state variables captured by model factors, the motivation for factor models remains empirical: the factors are chosen to capture patterns in average returns.

An important question (posed forcefully by John Cochrane) is, What is the discipline that prevents factor models from degenerating into mindless data dredging? The answer offered by Fama and French (1993, 2012, 2015, 2016, 2017) has three parts. (1) Robustness: A model should compete well on samples for different periods and different markets. (2) Parsimony: Though the guidelines remain to be drawn, other things equal, models with fewer factors are better. (3) Factors should have some motivation from theory, even if the theory is somewhat loose. For example, the

three-factor and five-factor models of Fama and French (1993, 2015) are consistent with the dividend discount valuation model widely used in finance and accounting. In the future we can hopefully do better. Time will tell.

Time-Varying Expected Returns

The expected return on a security is the expected compensation for holding the security. It is a price, and like other prices, it almost surely varies through time. Beginning with Fama (1975), research that attempts to measure variation in expected returns typically focuses on time-series regressions of returns on forecasting variables. For stocks, for example, regressions of returns on lagged dividend yields (ratios of annual dividends to price) are common. Fama and French (1988a, 1989) are examples, and Cochrane (2011) provides an insightful summary.

Fama and French (1988b) take a different tack. They observe that if expected stock returns are highly autocorrelated but slowly mean-reverting, stock returns will have negative autocorrelation that increases with the return horizon. We estimate that 30–40 percent of the variation of 3–5-year returns is due to time-varying expected returns, but the small samples for long horizons mean that the estimates are imprecise. Estimate imprecision is a general plague in the literature on time-varying expected returns.

The Economics of Organizations

Spurred by the pathbreaking paper of Jensen and Meckling (1976), my research in the late 1970s took a detour into agency theory. Earlier papers emphasize agency problems. I was interested in how competitive forces lead to mechanisms to mitigate agency problems. The first paper, Fama (1980), argues that managerial labor markets, inside and outside of firms, act to control managers faced with the temptations created by diffuse residual claims that reduce the incentives of individual residual claimants to monitor managers.

Michael Jensen and I then collaborated on two papers in a *JPE* sister journal (Fama and Jensen 1983a, 1983b) that study more generally how competition leads to different mechanisms to mitigate agency problems associated with separation of management and residual risk bearing, and how an organization's activities, and the special agency problems they pose, affect its residual claims and control mechanisms. For example, we argue that the redeemable residual claims of a financial mutual (e.g., an open-end mutual fund) provide strong discipline for its managers, but redeemability is cost-effective only when assets can be sold quickly with low transactions costs. We also argue that the nonprofit form, in which no

agents have residual claims to net cash flows, is a response to the agency problem associated with activities in which there is a potential supply of donations that might be expropriated by residual claimants.

Options Pricing and the *JPE*

Finally, though derivatives pricing is not my focus here, the options pricing papers of Black and Scholes (1973) and Merton (1973b), which resulted in Nobel Prizes for Merton and Scholes in 1997, warrant acknowledgment. These papers are familiar to students of finance and economics more generally, and they underpin a vibrant derivatives industry. One is hard-pressed to find other papers with such a combination of academic and applied impact. Myron Scholes is a member of the outstanding cohort of Chicago PhD students of the late 1960s and early 1970s, he and Fischer Black were Booth faculty for extended periods, and Robert Merton has an honorary Chicago PhD (granted before his Nobel). Black and Scholes (1973) is one of many fundamental *JPE* papers in finance.

References

Bachelier, L. J. B. A. 1900. *Theorie de la speculation*. Paris: Gauthier-Villars. Reprinted in *The Random Character of Stock Market Prices*, edited by Paul H. Cootner. Cambridge, MA: MIT Press, 1964.

Black, Fischer, Michael C. Jensen, and Myron S. Scholes. 1972. "The Capital Asset Pricing Model: Some Empirical Tests." In *Studies in the Theory of Capital Markets*, edited by Michael C. Jensen. New York: Praeger.

Black, Fischer, and Myron S. Scholes. 1973. "The Pricing of Options and Corporate Liabilities." *J.P.E.* 81 (3): 637–54.

Breeden, Douglas T. 1979. "An Intertemporal Asset Pricing Model with Stochastic Consumption and Investment Opportunities." *J. Financial Econ.* 7 (3): 265–96.

Campbell, John Y., and John H. Cochrane. 1999. "By Force of Habit: A Consumption-Based Explanation of Aggregate Stock Market Behavior." *J.P.E.* 107:205–51.

Carhart, Mark M. 1997. "On Persistence in Mutual Fund Performance." *J. Finance* 52:57–82.

Cochrane, John H. 2011. "Discount Rates: American Finance Association Presidential Address." *J. Finance* 66:1047–1108.

Epstein, Larry G., and Stanley Zin. 1991. "Substitution, Risk Aversion, and the Temporal Behavior of Consumption and Asset Returns: An Empirical Analysis." *J.P.E.* 99:263–86.

Fama, Eugene F. 1964. "The Behavior of Stock-Market Prices." *J. Bus.* 38:34–105.

———. 1970. "Efficient Capital Markets: A Review of Theory and Empirical Work." *J. Finance* 25:383–417.

———. 1975. "Short-Term Interest Rates as Predictors of Inflation." *A.E.R.* 65:269–82.

———. 1976. *Foundations of Finance*. New York: Basic Books.

———. 1980. "Agency Problems and the Theory of the Firm." *J.P.E.* 88:288–307.
———. 2014. "Two Pillars of Asset Pricing." *A.E.R.* 104:1467–85.
Fama, Eugene F., Lawrence Fisher, Michael Jensen, and Richard Roll. 1969. "The Adjustment of Stock Prices to New Information." *Internat. Econ. Rev.* 10:1–21.
Fama, Eugene F., and Kenneth R. French. 1988a. "Dividend Yields and Expected Stock Returns." *J. Financial Econ.* 22:3–25.
———. 1988b. "Permanent and Temporary Components of Stock Prices." *J.P.E.* 96:246–73.
———. 1989. "Business Conditions and Expected Returns on Stocks and Bonds." *J. Financial Econ.* 25:23–49.
———. 1992. "The Cross-Section of Expected Stock Returns." *J. Finance* 47 (2): 427–65.
———. 1993. "Common Risk Factors in the Returns on Stock and Bonds." *J. Financial Econ.* 33 (1): 3–56.
———. 2010. "Luck versus Skill in the Cross-Section of Mutual Fund Returns." *J. Finance* 65 (5): 1915–47.
———. 2012. "Size, Value, and Momentum in International Stock Returns." *J. Financial Econ.* 105 (3): 457–72.
———. 2015. "A Five-Factor Asset Pricing Model." *J. Financial Econ.* 116 (1): 1–22.
———. 2016. "Dissecting Anomalies with a Five-Factor Model." *Rev. Financial Studies* 29:70–103.
———. 2017. "International Tests of a Five-Factor Asset Pricing Model." *J. Financial Econ.* 123:441–63.
Fama, Eugene F., and Michael C. Jensen. 1983a. "Agency Problems and Residual Claims." *J. Law and Econ.* 26:327–49.
———. 1983b. "Separation of Ownership and Control." *J. Law and Econ.* 26:301–25.
Fama, Eugene F., and James D. MacBeth. 1973. "Risk, Return, and Equilibrium: Empirical Tests." *J.P.E.* 81 (3): 607–36.
French, Kenneth R. 2008. "The Cost of Active Investing." *J. Finance* 63:1537–73.
Jensen, Michael C. 1968. "The Performance of Mutual Funds in the Period 1945–1964." *J. Finance* 23:389–416.
Jensen, Michael C., and William H. Meckling. 1976. "Theory of the Firm: Managerial Behavior, Agency Costs and Ownership Structure." *J. Financial Econ.* 3:305–60.
Lintner, John. 1965. "The Valuation of Risk Assets and the Selection of Risky Investments in Stock Portfolios and Capital Budgets." *Rev. Econ. and Statis.* 47:13–37.
Lucas, Robert E., Jr. 1978. "Asset Prices in an Exchange Economy." *Econometrica* 46 (6): 1429–45.
Mandelbrot, Benoit. 1966. "Forecasts of Future Prices, Unbiased Markets, and Martingale Models." *J. Bus.* 39 (suppl.): 242–55.
Markowitz, Harry. 1952. "Portfolio Selection." *J. Finance* 7:77–91.
———. 1959. *Portfolio Selection: Efficient Diversification of Investments.* Cowles Foundation Monograph no. 16. New York: Wiley.
Merton, Robert C. 1973a. "An Intertemporal Capital Asset Pricing Model." *Econometrica* 41:867–87.
———. 1973b. "Theory of Rational Option Pricing." *Bell J. Econ.* 4 (1): 141–83.
Miller, Merton H., and Franco Modigliani. 1961. "Dividend Policy, Growth, and the Valuation of Shares." *J. Bus.* 34 (October): 411–33.

Modigliani, Franco, and Merton H. Miller. 1958. "The Cost of Capital, Corporation Finance, and the Theory of Investment." *A.E.R.* 48 (June): 261–97.
Samuelson, Paul. 1965. "Proof That Properly Anticipated Prices Fluctuate Randomly." *Indus. Management Rev.* 6 (Spring): 41–49.
Sharpe, William F. 1964. "Capital Asset Prices: A Theory of Market Equilibrium under Conditions of Risk." *J. Finance* 19:425–42.
———. 1991. "The Arithmetic of Active Management." *Financial Analysts J.* 47: 7–9.

Behavioral Economics

Richard H. Thaler

University of Chicago

Exactly 100 years ago, the *JPE* was poised to be at the forefront of the field that would eventually come to be called behavioral economics. John Maurice Clark, a *JPE* editor, University of Chicago faculty member, and son of John Bates Clark, authored the lead article of the January 1918 issue titled "Economics and Modern Psychology: I." (Part II appeared in the next issue.) His message was a simple one: "The economist may attempt to ignore psychology, but it is a sheer impossibility for him to ignore human nature. . . . If the economist borrows his conception of man from the psychologist, his constructive work may have some chance of remaining purely economic in character. But if he does not he will not thereby avoid psychology. Rather he will force himself to make his own, and it will be bad psychology" (4).

A few years later Clark left Chicago to take the position his father had once held at Columbia, and it seems fair to say that the subsequent editors of the *JPE* did not take up his call to arms. Behavioral economics papers have made only scattered appearances in the journal in the subsequent century.[1]

Thanks to Alex Imas, Emir Kamenica, and Jesse Shapiro for helpful comments.

[1] To put a tiny bit of data behind this assertion, I counted the number of papers published in a few top journals that are cited in Stefano DellaVigna's recent survey paper in the *Journal of Economic Literature*. The tally is *Q JE* 32, *AER* 21, *Journal of Finance* 16, and *JPE* 10. And my informal sense is that the 10 *JPE* papers contain a greater proportion that is not behavioral, as compared to those in the *Q JE* or *AER*.

As Herbert Simon once said, the term behavioral economics is a bit strange. "What 'non-behavioral' economics can we contrast with it?" he asked (Simon 1987, 221). One answer to this question is the style of economics that the *JPE* is perhaps best known for: price theory à la Chicago School led by the intellectual giants Gary Becker, Milton Friedman, and George Stigler. Becker's research goal was to apply the standard tools of maximizing behavior to study a wide variety of topics that were not then part of the domain of economics including addiction, crime, discrimination, marriage, divorce, childbearing, and social interactions.

Becker acknowledged that by applying the tools of economics to such topics he was pushing the envelope. In his Nobel address he discusses this explicitly (Becker 1993). "I have intentionally chosen certain topics for my research—such as addiction—to probe the boundaries of rational choice theory. . . . My work may have sometimes assumed too much rationality, but I believe it has been an antidote to the extensive research that does not credit people with enough rationality" (402).

Becker's last sentence suggests an alternative definition of behavioral economics: crediting people with just the right amount of rationality and human foibles. The trick is in figuring out what is just the right amount. The approach taken by most behavioral economists has been to focus on a few important ways in which humans diverge from *homo economicus*.

The basic assumption of standard economic theory is that among all the affordable consumption bundles, people choose the best one. One way that assumption might fail is if the utility maximization problem is too hard to solve; this is the problem of bounded rationality. Another cause of nonmaximizing behavior is a lack of willpower. The morning after, many decide that the previous night included at least one drink too many. Such self-control problems are the subject of my first publication in the *JPE* (Thaler and Shefrin 1981), and one of the first behavioral economics papers published in the journal since Clark's.[2]

Shefrin and I tried to modify the standard approach as little as possible to accommodate the struggle that people commonly face when choosing between an immediate small pleasure and larger delayed reward. Following Adam Smith's *Theory of Moral Sentiments* (1759), our model endows people with two conflicting sets of preferences, one belonging to a myopic "doer" and the other to a farsighted "planner." The doer lives just for one period and cares only about consumption in that period. The planner seeks to maximize the integral of doer utilities and so sometimes wishes to constrain or influence the doer's choices.

[2] I cite one earlier paper below. This is a good time to acknowledge that I have likely missed some important behavioral papers both before and after 1981. My apologies to the authors of the papers I have missed. Blame it on bounded memory and attention.

One implication of the planner-doer model is that individuals can be helped by market-supplied commitment strategies. Thaler and Benartzi (2004) provide evidence to support this prediction. Benartzi and I created a strategy to help reluctant savers that we called "Save More Tomorrow." Organizations offer their employees an opportunity to sign up for a program (starting in a few months) in which their pension contribution rates are increased each year when they get a pay raise. Standard economic theory predicts that no one would join such a program (they would not think they needed it) and that if they did, it would not change their savings rates (since they were already saving the optimal amount). The paper, published in a special issue of the *JPE* honoring my advisor Sherwin Rosen, reported the effects of the program in the first firm to try the idea. The results were striking: 80 percent of those offered Save More Tomorrow chose to join, and those who joined more than tripled their savings rates in just 4 years.

Kaur, Kremer, and Mullainathan (2015) study another type of commitment strategy offered by an employer, this time in the context of increasing output. The article reports on a yearlong experiment in which piece rate workers were offered a dominated contract on randomly chosen days. The employees could set a daily goal for themselves with the proviso that if they meet the goal they are paid normally, but if they fail to meet the goal they are paid only half the usual rate. Workers chose such contracts fully 36 percent of the time, and they were wise to do so. For those who opted in to the dominated contract, output (and thus pay) increased by 6 percent.

One of the most powerful findings of behavioral economics is "loss aversion," the psychological tendency to feel losses more acutely than gains. As Adam Smith (1759) put it, "Pain . . . is, in almost all cases, a more pungent sensation than the opposite and correspondent pleasure" (1981, III, ii, 176–77). Although Daniel Kahneman and Amos Tversky (1979) and I (1980) had earlier written about this phenomenon, its empirical validity was still very much in question when Kahneman, Jack Knetsch, and I submitted an experimental paper on the subject to the *JPE*, later published in 1990.

In the experiment we randomly assigned half the subjects to receive some object (often a coffee mug), with the other half getting nothing. We then conducted a market for the mugs in which both buyers and sellers stated their reservation prices. Since transaction costs were negligible and the objects were randomly assigned, the Coase theorem predicts that roughly half the mugs will change hands so that subjects who value mugs the most end up owning them. Our hypothesis was that fewer than half the mugs would trade because owners would regard a trade as a loss. This hypothesis was strongly supported. In a typical experiment, the expected

number of trades was 11 but the empirical average was only 3.4. As predicted by loss aversion, median reservation prices for selling the mug were roughly twice the median prices for buying the mug.

The editor handling this paper was George Stigler. He sent us back a rejection letter based on a highly critical referee report from someone Stigler described as a "heavyweight in the field." The referee said that income effects could explain our results since those who received the mugs had received a windfall relative to those who did not. After taking a few days to calm myself down (a good self-control strategy) I wrote back on behalf of my coauthors (who were both away traveling) explaining why the referee's comments could not be taken seriously, either theoretically or empirically. First, the marginal propensity to spend windfalls on university insignia coffee mugs must be minuscule. Second, one experiment explicitly tested and rejected this explanation. Stigler wrote back in his usual witty style saying that *JPE* stands for *Journal of Perspicacity and Equity*, and he offered to send both my letter and the original referee report to another referee to adjudicate. That referee said that if forced to choose between our view and that of the original referee, he would side with us, which is how the paper came to be accepted.[3]

Perhaps the subfield of economics in which the behavioral approach has had the greatest impact is finance, and although the *JPE* has published quite a few influential articles on the subject of financial economics, not many have been behavioral. One exception is the paper by De Long et al. (1990), "Noise Trader Risk in Financial Markets," which takes on a frequent misconception about the possible role of less than fully rational investors—"noise traders"—in well-functioning asset markets. De Long et al. quote the conventional Chicago wisdom (e.g., Friedman 1953; Fama 1965) that noise traders can have little effect on prices and that any mispricing cannot last long before being wiped out by rational arbitrageurs.

De Long et al. make the crucial observation that arbitrageurs are likely to be risk averse and to have short horizons (in part because they are usually managing other people's money). Thus when attempting to exploit mispricing caused by noise traders, arbitrageurs run the risk that whatever bias is inducing the noise traders to be excessively optimistic or pessimistic about a security might continue or even strengthen before the arbitrageurs have made their profits. This "noise-trader risk" prevents ar-

[3] The self-control paper also involved quite a bit of back and forth with the editor Sam Peltzman, who somewhat reluctantly agreed to accept it rather than continue to exchange letters. Both papers were published as the last paper in the issue, which I took as a signal that they were considered the paper the editors were most ashamed to publish. It is gratifying that both papers were ranked highly on the list of most-cited papers compiled by the editors for this issue. Perhaps people read the *JPE* from back to front.

bitrage from eliminating the price effects of noise traders. Indeed, in the De Long et al. model, noise traders actually make more money than rational traders because they inadvertently bear more noise trader risk, which because of the risk aversion of the rational traders pays a positive risk premium. So in this model noise traders can affect prices and they do not necessarily go broke—they might even get rich!

It is one thing to demonstrate that noise traders matter in a theoretical model; showing that noise traders influence actual market prices is another matter. How does one prove that a price is "wrong"? One approach is to exploit the basic building block of modern finance, the law of one price: two identical assets must sell for the same price. One counterexample cited by De Long et al. is the case of closed-end funds in which the price of a fund's shares should be equal to the net asset value of the securities the fund owns. But in fact closed-end funds typically sell at a discount relative to net asset value and occasionally sell at a premium.

Owen Lamont and I (2003) published a paper on this theme in the *JPE* with the obnoxious title "Can the Stock Market Add and Subtract?" As you might guess by now, the answer to the question posed by the title is "no." Lamont and I study equity carve-outs, focusing on the prominent example of Palm and 3Com. Here is the story in brief. Palm, a maker of then-sexy hand-held computers, was owned by 3Com, a profitable technology company. On March 2, 2000, 3Com sold a small fraction of its stake in Palm via an initial public offering (IPO). In this carve-out, 3Com retained 95 percent of the shares of Palm but announced that, pending an expected approval by the Internal Revenue Service, the remaining shares would be distributed to 3Com shareholders. At that point, 3Com shareholders would receive about 1.5 shares of Palm for every share of 3Com that they owned.

The law of one price implies in this case that the price of 3Com must be at least 1.5 times the price of Palm, since equity prices can never be negative. However, on the day of the Palm IPO, Palm's shares traded at $95.06 a share, but 3Com ended the day trading at $81.81, well short of the lower bound of $145 implied by the law of one price. Implicitly, the market was pricing the 3Com "stub" (the company once Palm was gone) at negative $22 billion!

Though it did not continue to invest much in the topic, the *JPE* published an early and influential paper on nonstandard beliefs, which arise when people do not use information optimally as traditional economic theory says they should.[4] Camerer, Loewenstein, and Weber (1989) demonstrate that people display a "curse of knowledge," in the sense that they

[4] A recent theoretical paper in the broad theme of biased beliefs by Bordalo, Gennaioli, and Shleifer (2013) studies cases in which "salient" features of the environment are given excessive weight.

have a hard time recognizing that others do not know what they know.[5] If Sally has written the code for some app, she is likely to underestimate the difficulty neophytes will have learning how to use the app. The problem is that Sally is unable to simulate how she would think in the absence of her expertise. Camerer et al. demonstrate the curse of knowledge in experimental markets in which traders with more information make systematic errors that affect market prices.

The editors gave us a limited amount of space for these essays, but that has not proved to be a major problem for the topic to which I was assigned. I cannot say for sure whether the small number of behavioral economics papers published in the *JPE* was a shortage of supply or demand, but there are entire branches of behavioral economics that have not made (much of) an appearance. Looking over DellaVigna's (2009) review article, one notices many themes that have been largely or entirely absent from the pages of the *JPE*, such as framing effects, menu effects (suboptimal diversification, effect of defaults, choice overload), peer pressure, and emotions.

When I came into the profession the *JPE* had a well-deserved reputation for having eclectic tastes. This was one reason Shefrin and I submitted our paper on self-control to the *JPE*. As the field of behavioral economics continues to grow, it will be a shame if the *JPE* does not include more behavioral research in its pages. I suggest the editors all read that paper by John Maurice Clark. But if the *JPE* continues to eschew papers on such topics, one can always quote Stigler and Becker (1977): "De gustibus non est disputandum."

References

Becker, Gary S. 1993. "Nobel Lecture: The Economic Way of Looking at Behavior." *J.P.E.* 101 (June): 385–409.

Bordalo, Pedro, Nicola Gennaioli, and Andrei Shleifer. 2013. "Salience and Consumer Choice." *J.P.E.* 121 (October): 803–43.

Camerer, Colin, George Loewenstein, and Martin Weber. 1989. "The Curse of Knowledge in Economic Settings: An Experimental Analysis." *J.P.E.* 97 (October): 1232–54.

Clark, John Maurice. 1918. "Economics and Modern Psychology: I." *J.P.E.* 26 (1): 1–30.

DellaVigna, Stefano. 2009. "Psychology and Economics: Evidence from the Field." *J. Econ. Literature* 47 (2): 315–72.

De Long, J. Bradford, Andrei Shleifer, Lawrence H. Summers, and Robert J. Waldmann. 1990. "Noise Trader Risk in Financial Markets." *J.P.E.* 98 (August): 703–38.

[5] A special case of the curse of knowledge is "hindsight bias" (Fischhoff 1975). Once people know that something happened, they remember thinking that they knew it all along.

Fama, Eugene F. 1965. "The Behavior of Stock-Market Prices." *J. Bus.* 38 (January): 34–105.
Fischhoff, Baruch. 1975. "Hindsight ≠ Foresight: The Effect of Outcome Knowledge on Judgment under Uncertainty." *J. Experimental Psychology: Human Perception and Performance* 1 (August): 288–99.
Friedman, Milton. 1953. *Essays in Positive Economics*. Chicago: Univ. Chicago Press.
Kahneman, Daniel, Jack L. Knetsch, and Richard H. Thaler. 1990. "Experimental Tests of the Endowment Effect and the Coase Theorem." *J.P.E.* 98 (December): 1325–48.
Kahneman, Daniel, and Amos Tversky. 1979. "Prospect Theory: An Analysis of Decision under Risk." *Econometrica* 47 (March): 263–92.
Kaur, Supreet, Michael Kremer, and Sendhil Mullainathan. 2015. "Self-Control at Work." *J.P.E.* 123 (December): 1227–77.
Lamont, Owen A., and Richard H. Thaler. 2003. "Can the Market Add and Subtract? Mispricing in Tech Stock Carve-outs." *J.P.E.* 111 (April): 227–68.
Simon, Herbert A. 1987. "Behavioural Economics." In *The New Palgrave: A Dictionary of Economics*, vol. 1, edited by John Eatwell, Murray Milgate, and Peter Newman. London: Macmillan.
Smith, Adam. [1759] 1981. *The Theory of Moral Sentiments*. Reprint, edited by D. D. Raphael and A. L. Macfie. Indianapolis: Liberty Classics
Stigler, George J., and Gary S. Becker. 1977. "De Gustibus Non Est Disputandum." *A.E.R.* 67 (March): 76–90.
Thaler, Richard H. 1980. "Toward a Positive Theory of Consumer Choice." *J. Econ. Behavior and Org.* 1 (March): 39–60.
Thaler, Richard H., and Shlomo Benartzi. 2004. "Save More Tomorrow™: Using Behavioral Economics to Increase Employee Saving." *J.P.E.* 112, no. 1, pt. 2 (February): S164–S187.
Thaler, Richard H., and H. M. Shefrin. 1981. "An Economic Theory of Self-Control." *J.P.E.* 89 (April): 392–406.

Corporate Finance

Robert Vishny

University of Chicago

Luigi Zingales

University of Chicago

The *Journal of Political Economy* and Chicago economists have played a major role in the development of the modern field of corporate finance,

pioneering the agency cost and property rights approach. This brief essay is a tribute to those contributions rather than a comprehensive review of the field. We apologize to those authors excluded by our singular focus.

Over 50 years ago, papers by Modigliani and Miller (1958) and Coase (1960) provided the foundations by establishing a remarkable set of "irrelevance" propositions. Coase established that, under the assumption of zero transaction costs, the allocation of property rights does not affect the ability of participants to achieve an efficient outcome. Modigliani and Miller showed that, with investment policy held constant and with zero taxes and transaction costs, the mix of securities issued by the firm does not affect the total value of the firm. These propositions provided a serious challenge to those who thought institutional arrangements could be easily explained.

Early Work

One of the fundamental assumptions behind the Modigliani-Miller irrelevance proposition is tax neutrality. Modigliani and Miller (1958, 1963) immediately recognized that in the United States (and most of the world) debt is tax favored at the corporate level. Thus, firm value increases when debt replaces equity. Modigliani and Miller's irrelevance proposition and its tax implications have been incredibly influential in the practice of finance. The fundamental valuation techniques (from the weighted average cost of capital to the adjusted present value approach) are based on the Modigliani-Miller proposition: they start from the cash flow available for investors (regardless of whether they are debt or equity holders), and then they adjust for the effect of taxes and possibly the cost of financial distress.

Even in a world in which debt is tax advantaged at the corporate level, capital structure can still be irrelevant for firm value if debt is tax disadvantaged from the perspective of personal taxes, as is the case in the United States (and particularly so before the 1986 tax reform). As Miller (1977) points out, if there is an interior equilibrium, it will have the characteristic that one dollar of pretax profits paid as interest should deliver investors the same value as a dollar of pretax profits paid as dividends or capital gains. Thus, the structure of taxes affects the average leverage in the economy, but any single company is still indifferent between issuing debt and equity.

The Agency Cost Approach to Corporate Finance

Starting in the 1950s, the influential managerialist literature (Baumol 1959; Simon 1959; Marris 1964; Williamson 1964) challenged the assump-

tion of value maximization and often abandoned optimizing models altogether in favor of ad hoc descriptive models. The Chicago response was to focus on managerial agency problems to explain deviations from value maximization as well as to understand real-world capital structures.

Alchian and Demsetz (1972) are early adopters of the agency cost perspective. They argue that the distribution of cash flow rights is determined to minimize the expected cost of shirking associated with team production. The firm's owner is a centralized monitor who can measure the productivity of team members and reward them accordingly. Ownership of the residual cash flows provides this centralized monitor with the incentive to be vigilant.

Jensen and Meckling (1976) also adopt an agency cost approach but try to match the reality of the modern corporation. The corporation is described as a "nexus of contracts" among various parties including owner-managers, employees, suppliers, outside equity holders, and bondholders instead of a monolith endowed with a single objective such as value maximization. Jensen and Meckling emphasize the conflicts between the various parties, especially management versus outside shareholders and bondholders versus shareholders.

For Jensen and Meckling, the essence of the manager-shareholder conflict is the dichotomy between cash flows and firm value on the one hand and perquisites or nonpecuniary benefits on the other. Their notion of perquisites has proved flexible enough to encompass various phenomena including diversion of cash flows, costs of effort, and empire building. Managers consume too many perquisites because they bear only a fraction of their cost. Outside equity is costly because it drives a wedge between the benefits of perquisites to the manager and the cash flow cost of those perquisites, although it allows the firm to pursue valuable investment opportunities. Borrowing may allow the manager to pursue some of these investments while still bearing the residual cash flow consequences, but leverage has its own associated agency costs. Jensen and Meckling outline how the optimal scale of the firm is determined by the trade-off between costly external finance and pursuit of positive net present value investments along with the optimal mix of manager-owned equity, outside equity, and debt. Their seminal paper illustrates the sheer explanatory power of a simple agency cost framework in corporate finance.

The main weakness of Jensen and Meckling's model is the omission of control rights of debt or outside equity as a means of limiting agency costs. Fama and Jensen (1983) argue that separation of residual risk bearing from decision management necessitates systems that also separate decision management from decision control. In large corporations, we typically have an explicit mechanism such as the board of directors for monitoring and decision ratification. Their paper has stimulated a thriving empirical literature on the role of boards of directors.

Jensen (1986) argues that corporate cash flow in excess of investment needs (free cash flow) runs the risk of being wasted by managers through self-aggrandizing negative net present value projects. The presence of debt in the capital structure has the benefit of reducing this kind of waste. This "free cash flow theory" provided a compelling rationale for the 1980s leveraged buyout wave.

Up to this point, the literature still lacks a good theory of ownership as distinct from claims to profit shares. Grossman and Hart (1986) fill this void. Their notion of ownership as the right to make decisions when contracts are incomplete has proved very powerful. It has influenced research on many topics including corporate governance and ownership patterns around the world, state-contingent financial contracting, venture capital contracting, and the role of public enterprises.

There is a large empirical literature on these topics. Early papers include Demsetz and Lehn (1985), Morck, Shleifer, and Vishny (1988), and Kaplan (1989).

Demsetz and Lehn (1985) study the variation of ownership concentration across large US firms and also find that there is no significant relationship between ownership concentration and profitability. Morck et al. (1988) find a nonmonotonic relationship between management ownership and market valuation, possibly indicating the interplay of an incentive effect as well as an entrenchment effect that sets in when management holds a large block of shares. Kaplan (1989) finds that ownership and leverage realignments via management buyouts have a positive impact on profitability, consistent with improved incentives.

The Market for Corporate Control

Manne (1965) is the first to point out the role of the market for corporate control in promoting efficiency and protecting atomistic shareholders from self-interested managerial behavior. Writing at a time when the antitrust consensus against horizontal mergers is very strong, he advocates taking into account the efficiency benefits of corporate takeovers. He explains how a low stock price resulting from inefficient management provides an opportunity for profit that acts as a stronger force for change than internal governance mechanisms.

Grossman and Hart (1980) take a similar view on the benefits of the market for corporate control but point out the limitations of the mechanism due to a free-rider problem. Atomistic shareholders have no incentive to tender their shares for anything less than the full value of those shares after managerial improvements. This makes it impossible for a more efficient buyer to profit on buying their shares. Shleifer and Vishny (1986) document the prevalence of large block holdings in US corpora-

tions and analyze the role of large shareholders as activist investors who can help overcome this free-rider problem.

The Labor Market

The pressure from the corporate control market is not the only remedy against agency problems. As Fama (1980) points out, an important form of discipline comes from the labor market, because the current marginal product of managerial labor contains information about future expected marginal products. Thus, the wage revision process imposed by the managerial labor market will reward well-performing managers and penalize poorly performing ones. In some special cases, this ex post settling up eliminates completely the costs of separation of ownership and control.

For this mechanism to work, however, managerial salaries should vary significantly on the basis of past performance and so should the probability of dismissal. Jensen and Murphy (1990) do not observe much evidence for these two predictions, a finding they attribute to unspecified "political forces" that constrain "the type of contracts that can be written between management and shareholders" (227).

International Dimension

Corporate finance theory was developed in the United States, inspired by US stylized facts, and for the first 30 years mostly tested on US data. Chicago economists have played a significant role in internationalizing the field. Rajan and Zingales (1995) confronted US-based capital structure theories with international evidence. They document that corporate leverage is fairly similar across developed countries. The differences in leverage reflect the way bankruptcy is designed rather than the divide between bank-centered and market-based economies. Where bankruptcy favors liquidation, firms appear more hesitant to lever up.

In corporate finance, it is often difficult to determine the direction of causality: does the law drive the behavior or does the behavior drive the law? La Porta et al. (1998) show that long-standing differences between legal systems can explain much of the variation in key investor protection laws across countries. These differences between legal systems can be traced to their families of origin, which in turn resulted from "a combination of conquest, imperialism, outright borrowing, and more subtle imitation" (1115). The authors document that the laws of common law countries (originating in English law) are more protective of outside investors than those of civil law countries (originating in Roman law) and that this difference can explain a significant fraction of the variation in financial development around the world. Their methods of coding inves-

tor protection laws have influenced much research by academics and policy makers and sparked a lively debate. Subsequent research has shown that the family of legal origin is highly correlated with a wide range of laws governing economic activity.

At Chicago, ideas are subjected to intense scrutiny, even among colleagues. While recognizing the importance of the legal origins argument, Rajan and Zingales (2003) thought that a more variable factor is needed to explain both the time-series variation and the cross-sectional differences in financial development. Their explanation is the opposition by incumbents, who feel threatened by financial markets, because they breed competition. They claim that incumbents' ability and willingness to resist financial development are reduced by free trade and capital flows. Indeed, financial market development accelerates with open borders and retrenches in periods of protectionism and restrictions to capital flows.

Given the two Chicago irrelevance propositions mentioned above, it is important to establish that these observed differences in financing patterns and financial market development are not just a neutral mutation, but matter for real economic variables. To this end, Rajan and Zingales (1998) show that differences in financial development can explain a country's ability to reallocate resources from sectors in excess to sectors in need, accelerating the growth of these sectors in need. The difference-in-difference strategy adopted in their paper has since become a standard in the literature.

New Directions

Chicago economists have also helped to broaden finance research beyond its traditional focus on large corporations. Entrepreneurial finance and household finance are two important new research areas. The field of entrepreneurial finance has grown along with the growth of private equity as an asset class and its increased role in driving innovation. In an early paper, Kaplan and Strömberg (2003) use the agency cost perspective to understand the allocation of cash flows and control rights in a large cross section of venture capital contracts. They find a strong correspondence between the predictions of agency theory and the structure of real-world venture capital contracts.

Household finance is increasingly recognized as perhaps the key link between finance and macroeconomics. Amit Seru and his coauthors have shown how securitization has led to lax lending standards as well as difficulty renegotiating bad loans (Keys et al. 2010; Piskorski, Seru, and Vig 2010). Atif Mian and Amir Sufi have established the close association between household debt and economic fluctuations using an array of data sets including detailed zip code–level data in the United States and a

1960–2012 panel data set covering 30 countries (see Mian and Sufi 2014; Mian, Sufi, and Verner 2017).

References

Alchian, Armen, and Harold Demsetz. 1972. "Production, Information Costs and Economic Organization." *A.E.R.* 62:777–95.

Baumol, William. 1959. *Business Behavior, Value and Growth.* New York: Macmillan.

Coase, Ronald. 1960. "The Problem of Social Cost." *J. Law and Econ.* 3:1–44.

Demsetz, Harold, and Kenneth Lehn. 1985. "The Structure of Corporate Ownership: Causes and Consequences." *J.P.E.* 93:1155–77.

Fama, Eugene. 1980. "Agency Problems and the Theory of the Firm." *J.P.E.* 88: 288–307.

Fama, Eugene, and Michael Jensen. 1983. "Separation of Ownership and Control." *J. Law and Econ.* 26:301–25.

Grossman, Sanford, and Oliver Hart. 1980. "Takeover Bids, the Free Rider Problem, and the Theory of the Corporation." *Bell J. Econ.* 11:42–64.

———. 1986. "Costs and Benefits of Ownership: A Theory of Vertical and Lateral Integration." *J.P.E.* 94:691–719.

Jensen, Michael. 1986. "Agency Costs of Free Cash Flow, Corporate Finance and Takeovers." *A.E.R.* 76:323–29.

Jensen, Michael, and William Meckling. 1976. "Theory of the Firm: Managerial Behavior, Agency Costs and Ownership Structure." *J. Financial Econ.* 3:305–60.

Jensen, Michael, and Kevin J. Murphy. 1990. "Performance Pay and Top Management Incentives." *J.P.E.* 98:225–64.

Kaplan, Steven. 1989. "The Effects of Management Buyouts on Operating Performance and Value." *J. Financial Econ.* 24:217–54.

Kaplan, Steven, and Per Strömberg. 2003. "Financial Contracting Theory Meets the Real World: An Empirical Analysis of Venture Capital Contracts." *Rev. Econ. Studies* 70:281–315.

Keys, Benjamin, Tanmoy Mukherjee, Amit Seru, and Vikrant Vig. 2010. "Did Securitization Lead to Lax Screening? Evidence from Subprime Loans." *Q.J.E.* 125:307–62.

La Porta, Rafael, Florencio Lopez-de-Silanes, Andrei Shleifer, and Robert Vishny. 1998. "Law and Finance." *J.P.E.* 106:1113–55.

Manne, Henry. 1965. "Mergers and the Market for Corporate Control." *J.P.E.* 73:110–20.

Marris, Robin. 1964. *The Economic Theory of "Managerial" Capitalism.* Glencoe, IL: Free Press.

Mian, Atif, and Amir Sufi. 2014. *House of Debt: How They (and You) Caused the Great Recession and How We Can Prevent It from Happening Again.* Chicago: Univ. Chicago Press.

Mian, Atif, Amir Sufi, and Emil Verner. 2017. "Household Debt and Business Cycles Worldwide." *Q.J.E.* Electronically published May 26.

Miller, Merton. 1977. "Debt and Taxes." *J. Finance* 32:261–75.

Modigliani, Franco, and Merton Miller. 1958. "The Cost of Capital, Corporate Finance and the Theory of Investment." *A.E.R.* 48:261–97.

———. 1963. "Corporate Income Taxes and the Cost of Capital: A Correction." *A.E.R.* 53:433–43.

Morck, Randall, Andrei Shleifer, and Robert Vishny. 1988. "Management Ownership and Market Valuation: An Empirical Analysis." *J. Financial Econ.* 20:293–315.

Piskorski, Tomasz, Amit Seru, and Vikrant Vig. 2010. "Securitization and Distressed Loan Renegotiation: Evidence from the Subprime Mortgage Crisis." *J. Financial Econ.* 97:369–97.
Rajan, Raghuram, and Luigi Zingales. 1995. "What Do We Know about Capital Structure? Some Evidence from International Data." *J. Finance* 50:1421–60.
———. 1998. "Financial Dependence and Growth." *A.E.R.* 88:559–86.
———. 2003. "The Great Reversals: The Politics of Financial Development in the Twentieth Century." *J. Financial Econ.* 69:5–50.
Shleifer, Andrei, and Robert Vishny. 1986. "Large Shareholders and Corporate Control." *J.P.E.* 94:461–88.
Simon, Herbert. 1959. "Theories of Decision Making in Economics and Behavioral Science." *A.E.R.* 49:253–83.
Williamson, Oliver. 1964. *The Economics of Discretionary Behavior: Managerial Objectives in a Theory of the Firm.* Englewood Cliffs, NJ: Prentice Hall.

Banking and the Evolving Objectives of Bank Regulation

Douglas W. Diamond

University of Chicago

Anil K. Kashyap

University of Chicago

Raghuram G. Rajan

University of Chicago

Views on the role played by banks in the economy have evolved greatly over the last 125 years, as have arguments on the need, as well as the best way, to regulate them. Some of the key insights in the debate have been published in the *Journal of Political Economy*. In what follows, we will outline the main contributions to the debate in recent years, with an emphasis on work done at the University of Chicago or published in the *JPE*. We

These views are those of the authors only and are not necessarily shared by the Bank of England or any other institutions with which we are affiliated.

want to emphasize work that has relevance today, but despite this caveat, we will probably end up doing injustice to work published long ago.

We begin with a framework for organizing the theories of intermediation. We then draw out the implications for what the theories say about regulation and note that in many respects the motivation for regulation has been only loosely tied to the theory of intermediation. We close with some open questions for regulators and economists interested in banking. We do not survey the research that has followed up on work published in the *JPE*, nor will we attempt to provide a detailed overview of the entire academic literature on banking. For that, we refer the reader to the excellent work by Gorton and Winton (2003) and Freixas and Rochet (2008).

Theoretical Overview

We define banks as financial institutions with a substantial fraction of illiquid assets financed with demandable liabilities payable at par. Banking theories typically have focused on one side of the bank's balance sheet as critical to its economic role and then argued why the other side of the bank's balance sheet has to take the form it does. In general, therefore, theories tend to emphasize the criticality of both sides of the balance sheet. Observing that bank-issued certificates of deposit pay market interest rates and that banks must hold central bank–issued reserves against them that pay a below-market rate, Fama (1985) argues that banks must provide some valuable services in order to bear this implicit tax. What might these services be? Let us start with various forms of liquidity provision.

Liquidity Provision

In environments in which the government does not issue sufficient payment media, banks could issue short-term demandable paper payable at par to fill the gap. An argument against privately issued bank money is that a bank will be tempted to overissue bank notes at par to unsuspecting clients and then default. Such "wildcat banking," critics argue, justifies a role for government-provided fiat money. Gorton (1996) analyzes the prices of private bank notes issued in the American Free Banking Era (1838–63) and finds that the risk of failure was priced into the bank note; for example, notes of new banks are discounted far more, and the discount declines as banks make payments over time (as predicted by Diamond [1989]; see later). Market discipline was a deterrent against overissue, though an open question is whether it sufficiently accounted for the risks of default.

Krishnamurthy and Vissing-Jorgenson (2012) show that even in modern times, the private sector supplies more short-term debt when govern-

ment debt issuance contracts, suggesting some degree of "gap filling." This has led for calls for the central bank (or government) to "crowd out" private short-term liability issuance by issuing risk-free excess reserves or short-term central bank paper (see, e.g., Carlson et al. 2016).

In a related vein to banks issuing payment media, banks may also provide depositors insurance against liquidity needs (Bryant 1980; Diamond and Dybvig 1983). Consider, for example, the three-date Diamond-Dybvig world starting at date 0 with depositors, some of whom may need to consume early (at date 1) and others who may need to consume late (at date 2). Also, the available investment projects at date 0 return a positive net return if left to mature at date 2 but return only the capital invested if liquidated beforehand. If depositors do not know what type they are a priori and invest directly in the project, the early consumers will have to settle for less consumption than late consumers. A bank can, however, act as a risk-sharing mechanism by promising to pay those who have to withdraw early a little more than the capital invested, while paying those who withdraw late a little less than the return to maturity of the project (but more than early withdrawers to incentivize them to stay for the long run). Unfortunately, though, this can expose banks to runs, because if everyone decides to withdraw early, the bank is committed to paying more than the early liquidation value of its assets. In such a situation, because the bank will not have anything left to pay those who do not demand to withdraw their money early, it makes sense for everyone to indeed withdraw their money early.

In sum, Diamond and Dybvig (1983) show that the liquidity risk sharing benefit that they identify is subject to panic-based runs (dominated Nash equilibria) even when illiquid assets are risk-free. They identify contractual solutions, such as suspension of convertibility of deposits to cash that can deter runs, but also show that government deposit insurance can do better in some circumstances because of the taxation authority of the government. The Diamond and Dybvig model has become the workhorse model in banking, in part because with a very simple framework, it rationalizes much of the structure as well as fragility of banks.[1]

Holmström and Tirole (1998) consider another way intermediaries can share and alleviate liquidity risk. In their model, the demand for liquidity comes from firms that get shocks that require them to infuse new

[1] The literature on runs has explored many other possibilities. In Bryant (1980) runs are based on depositor information and occur when the bank will be insolvent for any level of withdrawals. These information-based runs (in contrast to the liquidity or panic-based runs in Diamond and Dybvig [1983]) are also studied in Jacklin and Bhattacharya (1988). Postlewaite and Vives (1987) consider runs based on noisy information about when a depositor and other depositors will need their deposits. Diamond and Rajan (2005) and Goldstein and Pauzner (2005) examine runs in which depositors consider both information about solvency and its implications for bank losses due to illiquidity.

funds to protect existing investments. Moral hazard at the firm level ensures that there are positive wedges between the amount the firms need to infuse and what they can actually raise from the markets once the shock hits. Thus firms need to carry extra value ex ante that they can dip into if they have liquidity needs, so that positive net present value infusions can be funded.

Holmström and Tirole go on to ask how firms can do this if cash cannot be stored. One possibility is to hold claims against other firms. The problem is that this could be an inefficient way to store value since some firms will not need to infuse much, and will have "excess" liquid assets while others will have to infuse a lot and have too little. A better solution is to obtain committed lines of credit from an intermediary who holds shares in all firms. Only firms that need liquidity will draw down these lines. The intermediary can thus allocate available scarce liquidity to those who need it, avoiding trapped pools of liquidity. Holmström and Tirole go on to ask what would happen if the need for liquidity is aggregate: every firm needs to infuse more at the same time, and the amount each firm can raise (given the moral hazard wedge between what it earns and what it can promise to pay out) is too low. In such situations, even an intermediary cannot help since there is an aggregate shortage of pledgeable value relative to what needs to be raised. Holmström and Tirole point to a role for the government, which can get access to the value generated by the firm (because of the taxation authority of government) that a private-sector financier cannot access. The government can then lend more to firms than the private sector. Alternatively, firms can buy and hold government bonds, as a reliable source of value, to be sold for funds when the need arises. Holmström and Tirole suggest a liquidity premium for safe government claims, which is verified by Krishnamurthy and Vissing-Jorgensen (2012).

Kashyap, Rajan, and Stein (2002) also argue that banks may be useful in pooling demands for liquidity and that as long as demands to withdraw deposits are not perfectly positively correlated with drawdowns on lines of credit, banks may be in the best position to optimally use any given pool of liquidity. Indeed, they find that banks make more loan commitments than other intermediaries, and within the banking sector, banks with high ratios of transaction deposits to total deposits also have high ratios of total commitments to total loans.

Finally, much of the literature is about using intermediaries to divide up existing asset liquidity in better ways to meet the needs of investors. Diamond and Rajan (2001) argue that banks create additional liquidity because of their special capabilities. In their model, entrepreneurs need to raise money to fund projects. However, the entrepreneur's specific abilities are important to generate value. Because the entrepreneur cannot commit to stay with the project, he has the ability to hold up lenders,

which creates a wedge between the revenues he can generate and the amount he can borrow. This is the source of project illiquidity. Bankers, by learning alternative or second-best uses for the project assets without the entrepreneur, have a greater ability to extract repayment from him. This allows them to lend more to him. But what if the banker needs to raise money himself? Does the chain of illiquidity reassert itself since the banker can similarly hold up investors by threatening not to put his special collection skills to work?

Diamond and Rajan argue that funding through demandable deposits prevents such banker holdup and tie the banker's collection skills to the loans he has made, thus enabling him to borrow against the full value of the loans. Intuitively, if the banker threatens to pay depositors less than they are owed, they run on the bank. Importantly, because the banker is only an intermediary transferring value from the entrepreneur to the depositor and does not generate independent value, he can be shut out of any postrun negotiation. This implies that the run does squeeze out any intermediary rents and the threat of a run acts as an effective disciplinary device on the banker. Note that because the entrepreneur adds value to the project, he cannot tie his human capital to the project by issuing demandable debt directly to investors: demand deposits discipline intermediaries, not firms, which distinguishes Diamond and Rajan (2001) from Calomiris and Kahn (1991), where demand deposits discipline all.

Diamond and Rajan thus argue that demandable debt is a feature of banks, not a bug, and enables the bank to raise money whenever needed to fund firms with more than they can borrow from markets. This implies that bank fragility cannot be eliminated without eliminating bank funding cost advantage. Diamond and Rajan (2000) explore the role of bank capital in reducing fragility, which has to be traded off against the enhanced cost of bank funding.

Allen and Gale (1997) suggest a different form of risk sharing in intermediaries: smoothing intergenerational risk. They study a standard overlapping generations model with risk-averse investors. The young get an endowment when born. There is a fixed-supply risky asset that is infinitely lived and pays a nonnegative but independent and identically distributed dividend, as well as safe assets paying zero dividends that can serve as a store of value. Because the representative young agent solves the same problem each period, the equilibrium price of the risky asset is constant. Given that it pays a nonnegative, and sometimes positive, dividend, its total return dominates that of the safe asset. The safe asset is therefore never used in the market equilibrium. This means that each generation bears a substantial amount of risk, since its last-period consumption varies with the entire dividend paid by the risky asset. Allen and Gale rule out any possibility of market-based intergenerational insurance since

the old know the dividend they will get before the young are born, and there is no scope for risk sharing after the risk is realized.

The key then to risk sharing is an intermediary, in whom generations invest their residual endowment (after consumption). The intermediary holds the risky asset and builds up reserves when the dividend is high, only to run them down at times of low dividend. Each generation then gets a smoothed return, which improves on market-based outcomes. This is probably a better model of intermediary dominated financial systems (such as what Germany or Japan used to be) than of single intermediaries. More recent work by Dang et al. (2017) develops a related idea in which banks keep the information about the realized value of their investments secret so as to facilitate risk sharing. Secrecy provides incentives for new depositors to deposit to provide liquidity to the previous generation even when the low value of bank assets, if known, would leave the bank unable to offer them a rate of return that matches that available in the market.

Banks as Monitors

We have already reviewed papers that emphasize the role of banks in monitoring or managing assets to enhance liquidity provision. Other work emphasizes how monitoring can alter the availability of funding to borrowers.

Diamond (1984) argues that costly monitoring by banks can resolve moral hazard or adverse selection problems at firms. But then who monitors the monitor? Does this not simply push the problem one step back: will investors in the bank not have to engage in costly monitoring to ensure the bank monitors? Diamond argues that when the bank is diversified across a large number of loans, bank asset values will be less sensitive to the private information in each loan. If investors in the bank hold debt claims, they will not need to have information about the bank's portfolio value to enforce those claims and, if the claims are sufficiently safe, will not have to individually monitor the bank to see that it is doing its job. Furthermore, the need to service the debt claims forces the bank to monitor the loans and to repay the depositors. Banks are special because diversification reduces the importance of private information attached to each loan the bank makes and makes the bank's overall balance sheet more transparent. Banks in Diamond (1984) are thus the original form of pooling (diversification) and tranching (issuing senior claims to outside depositors and retaining junior claims inside the bank) structures that have proliferated in securitization vehicles.

Subsequent work in the *JPE* contributes to the characterization of the borrowers who would benefit most from monitoring and be most depen-

dent on banks as a source of finance. The analysis is based on differences in the severity of borrower moral hazard when borrowers consider the effect of their current choices on their future access to finance.

Diamond (1989) examines the role of reputation as a way of reducing borrower moral hazard, and Diamond (1991) extends this to examine the interaction of monitoring and reputation effects. These models predict that new borrowers would be subject to severe moral hazard and that the severity would be reduced over time for borrowers who survive and acquire a good reputation for repaying investors. The analysis differs from then-existing models of reputation by predicting that it may take time to acquire a reputation. The earlier models (Fama 1980; Holmström 1982; Kreps and Wilson 1982; Milgrom and Roberts 1982) focused on the effects of a prospect of having a reputation in the future rather than on the costs of losing one's current reputation.

Diamond (1989) considers a model of borrower moral hazard (distorted incentives for real investment risk choice) in which the project choice is private information, debt contracts are optimal financial contracts, and the only thing investors observe is the realized payment they receive. Borrowers who repay debt over time acquire a better reputation (a better credit rating), and this reputation becomes an asset that they lose if they subsequently default. Those who repay debt over time consist of those who always choose safe investments (have no moral hazard) and those who are subject to moral hazard but whose risky investments had a realization sufficient to repay investors. These repayments separate the borrowers from those whose risky investments have realizations insufficient to repay investors. This learning about survivors improves their reputation, and their moral hazard is reduced sufficiently until they prefer to avoid risky investments to maintain their reputation. The reputation of borrowers in this model is measured by their credit rating (their probability of default for a given level of borrowing).

Diamond (1989) ignores the possibility of borrowing from a lender who can monitor the investment choices of a borrower. Diamond (1991) investigates when monitoring will and will not be valuable. A separation emerges in which new borrowers without a long track record need monitoring from banks, while others who always repay such debt for a long enough time acquire a sufficiently good reputation to borrow directly without monitoring. The second group can issue debt directly to public markets. Although their record of successful repayments is made while investment choices have been monitored by banks, it helps future lenders learn about them: they separate themselves from borrowers who choose risky investments despite monitoring (those who have only very risky investments available or, more generally, those with stronger conflicts of interest). This produces a life cycle theory of borrowing: young borrowers (small and medium-sized businesses) that borrow from banks and ma-

ture ones that acquire a good enough credit rating switch to unmonitored borrowing and no longer depend on bank finance.

A small extension of this model could consider a third, extremely risky set of borrowers with a limited track record. Suppose that bank monitoring is effective but imperfect: either a competing higher level of monitoring exists (from family members) or some initial borrowers have such strong moral hazard that monitoring cannot improve their incentives. This implies that these very young and risky borrowers first self-finance or borrow from family members who can monitor more closely than banks. Only those who survive this start-up period can borrow from banks. This identifies the set of borrowers whose access to finance depends on the banking system and its financial health.

The possibility that banks may have private information about borrowers acquired over time raises the question of costs of bank-firm relationships. Rajan (1992) argues that because firms might be seen as lemons if they exit such a relationship, banks have holdup power over firms. This possibility, in turn, can influence firms' choice of financing between long-term arm's-length financing and bank financing and the number of banks they may want to borrow from. Bolton and Freixas (2000) present another private information–based model, adding costs of restructuring debt and of issuing bank equity to provide a theory of the type of firms that issue bonds, equity, and bank debt.

Aggregate Liquidity Shortages, Fire Sales, and Contagion

If banks have special skills in evaluating and monitoring loans, then there is a small group of entities with similar skills that have the ability to purchase such loans. This raises the possibility that when there is an aggregate shortage of financing in the market and a bank has to sell the loans on its books to repay the short-term debt it has coming due, there will be a limited pool of buyers with limited resources to buy those loans, and loans will sell, not for the full value that buyers can collect, but for what they can pay. The possibility that loans are priced at a fire sale discount value because the best buyers have limited financing is an important source of risk in banking. Shleifer and Vishny (1992) provide an early discussion of this phenomenon and Allen and Gale (1994) discuss the effects of limited participation in financial markets. To the extent that banks do not internalize the fact that they will be consuming the limited common pool of resources available to buy loans when there is a future aggregate shortage, thereby lowering the fire sale price and, importantly, the ability of other banks to finance themselves, there is a fire sale externality that causes banks to overissue short-term debt (see Stein 2012).

In Allen and Gale (2000), aggregate liquidity shortages can result in a contagion of bank failures. Interestingly, these arise from an attempt to share aggregate liquidity through interbank deposits, as opposed to a single bank that owns each firm as in Holmström and Tirole (1998). Essentially, as long as there is no aggregate liquidity shortage, a liquidity shortage in one region can be smoothed over by a bank drawing down on deposits it has made in banks in liquidity-surplus regions. However, if there is an aggregate liquidity shortage, banks could fail in a region that is short, since enough liquidity is not available elsewhere. This will imply a collapse in deposit values, which will reduce the value of banks that hold interbank deposits in the failing bank and transmit the failure to regions connected by interbank deposits. Interestingly, Allen and Gale argue that partial interconnection can be worse than full interconnection between banks across regions since the latter allows better use of the available common pool of aggregate liquidity.

In Diamond and Rajan (2005), all banks have access to the common pool of liquidity, and in contrast to Allen and Gale (2000), there are no ex ante interbank claims. Nevertheless, even in this structure bank failures can be contagious; bank insolvencies precipitate runs that cause them to dump all their assets on the market, thus shrinking the available pool of liquidity even more, causing other banks to become insolvent and run. In these views of crises, the failure of banks disrupts lending relationships and causes firms to face credit crunches. Investment and production then collapse. If banks are special, central bank liquidity provision can help keep banks solvent in the face of panics (see the evidence in Carlson, Mitchener, and Richardson [2011]) and avert wider systemic distress.

An alternative view of crises espoused by Friedman and Schwartz (1963) is that they stem from a shortage of payment media as bank failures result in a contraction of bank deposits. Once the payment system collapses, spending declines and falling prices are inevitable. In their view a series of banking collapses is just one way in which the money supply could contract, and as in all contractions in the money supply, the results for the real economy and deflation would be bad.

Financial Regulation

The evolution of financial regulation was dramatically overhauled during the Great Depression. The chaos associated with runs and massive number of bank closures and failures spurred a number of policy proposals to prevent that from occurring again. For example, Freidman and Schwartz (1963) emphasized the critical role of the central bank in not allowing sharp money supply contractions. Other regulations were aimed more at the financial system itself, imposing constraints on different institutions or agents and their ability to set interest rates or capital structures.

Many of these important regulatory interventions occurred before the theories surveyed above had been formally developed, but the intuition behind them played a role.

One prominent reaction, championed by Henry Simons and others at the University of Chicago (Simons 1933, 1936), was to call for an end to fractional reserve banking in what came to be known as the "Chicago Plan."[2] This idea, which in recent incarnations would be called narrow banking, is consistent with the view that the value added of the banking system comes from the payment services it provides. If the liabilities are what is special, then restricting the asset side of the banks to be boring and safe is a logical proposal.

While the Chicago Plan attracted many followers and appears to have received considerable attention at all levels of government, the Banking Act of 1933 instead created a national deposit insurance scheme. Modern theories recognize that deposit insurance brings some stabilizing benefits but also creates distortions by creating incentives to take risks that might be borne by a deposit insurance fund. Indeed, to mitigate risks, banks were subject to a plethora of additional regulations including limitations on interest payments. While these were relatively benign when market interest rates were low, they became problematic as market interest rates shot up in the 1970s and 1980s.

The next pair of major banking reforms in the United States, the adoption of the first Basel capital standards in 1988 and the FDIC Improvement Act (FDICIA) of 1991, can be viewed as the responses to the long-standing concerns about deposit insurance distortions. The Basel accords mandated that banks hold more capital in their liability structure, and FDICIA forced bank supervisors to close severely undercapitalized banks without delay so that fewer would operate while insolvent.

The latest wave of regulatory changes came in response to the global financial crisis of 2007–9. The policy responses appear to reflect different interpretations of the root causes of the crisis. Some new regulations (for instance, the Dodd-Frank Act in the United States) are most naturally viewed as concluding that some banks were "too big" or "too interconnected" to fail and that reforms to prevent a replay were needed. Others (e.g., some aspects of the latest Basel reforms) suggest that widespread runs that were mostly outside the traditional commercial banking system were the problem.

The lack of a unified diagnosis of what went wrong has led to four major regulatory innovations. First, the push for higher capital requirements was accelerated, especially given the view that financial institutions had

[2] Simons describes the plan briefly as an example of how to implement the ideas in his classic 1936 *JPE* paper on rules vs. discretion.

gamed existing capital requirements before the crisis. Capital requirements for all banks have risen substantially, with special additional rules for large global entities. While these are understandable given the extent of leverage before the crisis, it is important to note that banking theories do suggest that the cost of bank financing can go up as more bank capital is required. Unlike a pure Modigliani-Miller view of bank capital structure, this more nuanced view suggests there are trade-offs in setting bank capital. What the optimal level of bank capital should be is still an open question.

Second, liquidity regulations are also being imposed for the first time. These rules force banks with more illiquid assets to use more long-term funding and also mandate that banks with more runnable funding should hold more liquid assets that could easily be sold to meet outflows. The theoretical emphasis on liquidity outlined earlier and the externality caused by excessively illiquid assets funded by runnable claims support this focus. However, calibration of these regulations is proceeding with little theoretical guidance.

Third, new regulators have been created, with responsibilities for looking at the stability of the entire financial system rather than individual institutions or sectors. The range of tools and authorities for these "macroprudential" regulators vary greatly. These efforts are in their infancy, so it is too early to tell whether this approach will succeed in delivering extra stability and, if so, how and why. Once again, though, the ideas that liquidity and solvency are interlinked both within and across financial institutions and that the location of excess liquidity as well as the quantum of aggregate liquidity matters lend support to efforts at macroprudential interventions. More research, both theoretical and empirical, however, is needed to guide regulations.

Finally, banking regulators everywhere have begun "stress-testing" large banks. These exercises simulate how banks will fare under different macroeconomic scenarios. Motivated by the same concerns as FDICIA, that the book value of bank equity is a lagging indicator of bank health, these tests have become the primary supervisory tool in many jurisdictions.

The postcrisis regulations raise some obvious research questions. For instance, will macroprudential regulation prove to be a mirage or will it really change the riskiness of the financial system and its resilience? In most jurisdictions, activist macroprudential policies have not been pursued, and given the effect they could have on firms' profitability and business practices, some political pressure to avoid acting is likely to be present. What will be the most effective way to design a macroprudential regulator? Which tools are needed to deliver on the mandate? How do (and should) these policies interact with monetary policy?

This leads to a related set of questions. To what extent does monetary policy interact with bank lending and bank liquidity? Do central bank promises of abundant aggregate liquidity if the system is stressed cause

banks to create overly illiquid balance sheets (see, e.g., Diamond and Rajan 2012; Farhi and Tirole 2012)? To what extent do bank regulation and monetary policy work at counterpurposes?

Another issue is whether the international regulatory coordination that has featured so prominently since the crisis will prove to be successful. International monetary and fiscal policy coordination has rarely been sustained, and many question whether it is desirable except in very extreme circumstances. Is there some reason why financial regulation demands coordination, and is the harmonization that has been pursued working? Would it make more sense to allow countries to have more regulatory independence, while improving the frameworks for dealing with cross-border insolvencies and spillovers?

Finally, what exactly is the role of liquidity regulations? The Modigliani-Miller propositions (Modigliani and Miller 1958; Miller and Modigliani 1961) serve as a starting point for our thinking on capital regulation, from which departures have to be justified. There is no equivalent benchmark that describes whether the financial system is producing too much or too little liquidity. Could the new regulations destroy value? Are two liquidity requirements necessary? How do the liquidity and capital regulations interact? See Diamond and Kashyap (2016) for discussion of some of these issues.

Before the recent financial crisis, there was a broad sense that the financial system was well understood and relatively stable, especially in developed countries. The financial crisis has refocused attention on many of the issues that financial economists were pondering earlier. Despite the intense research and policy advances made recently, many open questions remain. Given the importance of the financial system in both benefiting and harming the overall economy, these are areas where more work will be very valuable.

References

Allen, Franklin, and Douglas Gale. 1994. "Limited Market Participation and Volatility of Asset Prices." *A.E.R.* 84 (4): 933–55.

———. 1997. "Intermediaries and Intertemporal Smoothing." *J.P.E.* 105 (3): 523–46.

———. 2000. "Financial Contagion." *J.P.E.* 108 (1): 1–33.

Bolton, Patrick, and Xavier Freixas. 2000. "Equity, Bonds, and Bank Debt: Capital Structure and Financial Market Equilibrium under Asymmetric Information." *J.P.E.* 108 (2): 324–51.

Bryant, John. 1980. "A Model of Reserves, Bank Runs and Deposit Insurance." *J. Banking and Finance* 4 (4): 335–44.

Calomiris, Charles W., and Charles M. Kahn. 1991. "The Role of Demandable Debt in Structuring Optimal Banking Arrangements." *A.E.R.* 81 (3): 497–513.

Carlson, Mark, Burcu Duygan-Bump, Fabio Natalucci, et al. 2016. "The Demand for Short-Term, Safe Assets and Financial Stability: Some Evidence and Implications for Central Bank Policies." *Internat. J. Central Banking* 12 (4): 307–33.

Carlson, Mark, Kris James Mitchener, and Gary Richardson. 2011. "Arresting Banking Panics: Federal Reserve Liquidity Provision and the Forgotten Panic of 1929." *J.P.E.* 119 (5): 889–924.

Dang, Tri Vi, Gary Gorton, Bengt Holmström, and Guillermo Ordoñez. 2017. "Banks as Secret Keepers." *A.E.R.* 107 (4): 1005–29.

Diamond, Douglas W. 1984. "Financial Intermediation and Delegated Monitoring." *Rev. Econ. Studies* 51 (3): 393–414.

———. 1989. "Reputation Acquisition in Debt Markets." *J.P.E.* 97 (4): 828–62.

———. 1991. "Monitoring and Reputation: The Choice between Bank Loans and Directly Placed Debt." *J.P.E.* 99 (4): 689–721.

Diamond, Douglas W., and Philip H. Dybvig. 1983. "Bank Runs, Deposit Insurance, and Liquidity." *J.P.E.* 91 (3): 401–19.

Diamond, Douglas W., and Anil K. Kashyap. 2016. "Liquidity Requirements, Liquidity Choice, and Financial Stability." In *Handbook of Macroeconomics*, vol. 2, edited by John B. Taylor and Harald Uhlig, 2263–2303. Amsterdam: North-Holland.

Diamond, Douglas W., and Raghuram G. Rajan. 2000. "A Theory of Bank Capital." *J. Finance* 55 (6): 2431–65.

———. 2001. "Liquidity Risk, Liquidity Creation, and Financial Fragility: A Theory of Banking." *J.P.E.* 109 (2): 287–327.

———. 2005. "Liquidity Shortages and Banking Crisis." *J. Finance* 60 (2): 615–47.

———. 2012. "Illiquid Banks, Financial Stability, and Interest Rate Policy." *J.P.E.* 120 (3): 552–91.

Fama, Eugene F. 1980. "Agency Problems and the Theory of the Firm." *J.P.E* 88 (2): 288–30.

———. 1985. "What's Different about Banks?" *J. Monetary Econ.* 15 (1): 29–39.

Farhi, Emmanuel, and Jean Tirole. 2012. "Collective Morality Hazard, Maturity Mismatch, and Systemic Bailouts." *A.E.R.* 102 (1): 60–93.

Freixas, Xavier, and Jean-Charles Rochet. 2008. *Microeconomics of Banking*. Cambridge, MA: MIT Press.

Friedman, Milton, and Anna J. Schwartz. 1963. *A Monetary History of the United States, 1867–1960*. Princeton, NJ: Princeton Univ. Press.

Goldstein, Itay, and Ady Pauzner. 2005. "Demand-Deposit Contracts and the Probability of Bank Runs." *J. Finance* 60 (3): 1293–1327.

Gorton, Gary. 1996. "Reputation Formation in Early Bank Note Markets." *J.P.E.* 104 (2): 346–97.

Gorton, Gary, and Andrew Winton. 2003. "Financial Intermediation." In *Handbook of the Economics of Finance*, vol. 1A, edited by George M. Constantindes, Milton Harris, and René M. Stulz. Amsterdam: Elsevier.

Holmström, Bengt. 1982. "Managerial Incentive Problems—a Dynamic Perspective." In *Essays in Economics and Management in Honor of Lars Wahlbeck*. Helsinki: Swedish School Econ.

Holmström, Bengt, and Jean Tirole. 1998. "Private and Public Supply of Liquidity." *J.P.E.* 106 (1): 1–40.

Jacklin, Charles J., and Sudipto Bhattacharya. 1988. "Distinguishing Panics and Information-Based Bank Runs: Welfare and Policy Implications." *J.P.E.* 96 (3): 568–92.

Kashyap, Anil K., Raghuram G. Rajan, and Jeremy Stein. 2002. "Banks as Providers of Liquidity: An Explanation for the Co-existence of Lending and Deposit Taking." *J. Finance* 57 (1): 33–73.

Kreps, David M., and Robert Wilson. 1982. "Reputation and Imperfect Information." *J. Econ. Theory* 27 (2): 253–79.

Krishnamurthy, Arvind, and Annette Vissing-Jorgensen. 2012. "The Aggregate Demand for Treasury Debt." *J.P.E.* 120 (2): 233–67.

Milgrom, Paul, and John Roberts. 1982. "Predation, Reputation, and Entry Deterrence." *J. Econ. Theory* 27 (2): 280–312.
Miller, Merton H., and Franco Modigliani. 1961. "Dividend Policy, Growth, and the Valuation of Shares." *J. Bus.* 34 (October): 411–33.
Modigliani, Franco, and Merton H. Miller. 1958. "The Cost of Capital, Corporation Finance, and the Theory of Investment." *A.E.R.* 48 (June): 261–97.
Postlewaite, Andrew, and Xavier Vives. 1987. "Bank Runs as an Equilibrium Phenomenon." *J.P.E.* 95 (3): 485–91.
Rajan, Raghuram G. 1992. "Insiders and Outsiders: The Choice between Informed and Arm's-Length Debt." *J. Finance* 47 (4): 1367–1400.
Shleifer, Andrei, and Robert Vishny. 1992. "Liquidation Values and Debt Capacity: A Market Equilibrium Approach." *J. Finance* 47 (4): 1343–66.
Simons, Henry C. 1933. "Banking and Currency Reform (with Supplement and Appendix)." Manuscript, Dept. Econ., Univ. Chicago.
———. 1936. "Rules versus Authorities in Monetary Policy." *J.P.E.* 44 (1): 1–30.
Stein, Jeremy C. 2012. "Monetary Policy as Financial-Stability Regulation." *Q.J.E.* 127 (1): 57–95.

Monetary Economics

Fernando Alvarez

University of Chicago and National Bureau of Economic Research

Introduction

In this paper I will review contributions to monetary economics, notwithstanding Lucas's essay on Chicago developments of monetary theory by Friedman and Patinkin. I will discuss the most highly cited papers on monetary economics published in the *Journal of Political Economy*, together with related work not necessarily published in the *JPE* to place them in context. I will organize the discussion of these contributions using Jevons's functions of money: store of value, medium of exchange, and unit of account.

Store of Value

I interpret the function of store of value as best captured by the monetary equilibrium of the overlapping generations model. The seminal paper on this topic is the *JPE* piece by Samuelson (1958), a ground-breaking contribution. This model, either on the version worked out by Samuelson or in subsequent ones, is useful to address several important issues,

such as the desirability of a pay-as-you-go social security or the role of demographic factors and of public debt in capital accumulation; see, for instance, Blanchard (1985) for an example of such analysis. Besides those issues, Samuelson's overlapping generations model is perhaps the clearest framework to conceptualize fiat money as a store of value. In the monetary equilibrium, money is an asset backed only by its use as a store of value—a form of rational bubble. There are multiple uses of the overlapping generations models in monetary economics. One such use is the path-breaking paper "Expectations and the Neutrality of Money" by Lucas (1972). The popularity of the overlapping generations model comes and goes as different phenomena fit the role of money as a store of value, such as the current research on the effects of and causes for the increase in the demand for safe and/or liquid assets.

Medium of Exchange

There is a long tradition of work on monetary economics to understand the role of money as a medium of exchange. Earlier seminal contributions are Baumol (1952) and Tobin (1956) on inventory theoretical models of cash holdings. During the postwar period there were theoretical and empirical advances on the study of money demand and quantitative theory led by Friedman and students of his; see Lucas's paper and the references therein for details. Later on there were related dynamic versions such as, for example, those in the work by Sidrauski (1967a, 1967b) with money in the utility function. Even later, that is, post–rational expectations revolutions, there are new analysis and conceptualizations, such as the cash-credit model of Lucas and Stokey (1987). All these models give closely related frameworks to properly define money demand and quantify its effects. A recent important development is the one in Kiyotaki and Wright (1989) and follow-up work by Lagos and Wright (2005), both published in the *JPE*. These models provide microfoundations for fiat or commodity money as a medium of exchange, based on search and information frictions. Their later versions build a bridge with traditional money demand theory, as well as provide insights of transactions of other assets beyond fiat or commodity money.

Standard of Value or Unit of Account

I will consider two related sets of ideas on why monetary policy has an effect, which I regard as, roughly speaking, related to the standard of value of unit of account function of money. One is the models that use nominal rigidities, such as sticky prices, to understand the effect of monetary shocks. The other set of ideas is the models that use rational confusion between nominal and real shocks as vehicles for monetary policy to affect output.

There is a long tradition, certainly emphasized by Keynes, on the macroeconomic effects of nominal rigidities and its relationship with the effects of monetary and fiscal policy. Some form of nominal nonneutrality is also central to the view of the role of monetary policy portrayed in Friedman's (1968) presidential address, which very much inspires several features of contemporary modeling, as well as the contemporary understanding of how a monetary policy shock plays it out in the short run (5–7). One possible reason for the nonneutrality is the effect of monetary policy on expectations, as described in Friedman's (1977) Nobel lecture, also published in the *JPE*. In that lecture, Friedman discusses unexpected nominal changes as candidate explanations of the traditional Phillips curve type of relationships: "only surprises matter." Another prominent candidate for such nonneutrality is nominal wage or nominal price rigidities. I discuss these candidates separately.

I first turn to contributions published in the *JPE* on the effect of monetary policy due to rational confusion between real and nominal shocks. The work by Muth (1961), Lucas and Prescott (1971), and Lucas (1972) was the cornerstone for the rational expectations revolution. In addition to the important methodological questions addressed in Lucas (1972), this paper tackles the classical macroeconomic question of the effect of monetary disturbances on aggregate prices and output. In his Nobel lecture, Lucas (1996) gives a historical account of the question, describes the difficulties in assessing it purely on empirical grounds, and sketches the main ideas behind the nonneutrality of money based on dispersion of information. One of the features emphasized by the class of models in his lecture is the differential short-term impact on prices and output of expected versus unexpected monetary shocks. This differential impact produces a type of Phillips curve relationship, due to the agents' rational confusion between nominal and real shocks. These ideas have had a deep impact on how macroeconomists think about the classical question of the effect of monetary shock. For instance, there was an early and influential theoretical work published in the *JPE* exploring the circumstances under which systematic monetary policy does have on effect on output. Sargent and Wallace (1975) is one of the main early rational expectations contributions emphasizing the generality of the result that the systematic component of monetary policy does not affect output, as well as the classical Wicksell indeterminacy of the price level under interest rates rules. On the one hand, Fischer (1977) is an early example in which, in spite of rational expectations and as a result of nominal wages set in advance for periods longer than those for which monetary policy can react, systematic effects of monetary policy do indeed affect output. On a related matter, Taylor (1980) studies the effect of setting wages in a staggered fashion on the propagation of shocks. He argues that in equilibrium this friction implies adjustment to shocks that are even more persistent than the length

of their assumed lag.[1] Complementary to these theoretical papers, Barro (1978) estimates the contribution of the expected and unexpected component of monetary changes on output and prices for the United States, finding evidence consistent with Lucas's hypothesis. Indeed, the distinction between expected and unexpected shocks plays an important role in the method presented in a seminal paper by Sims (1980). This paper introduced vector autoregressions (VARs), which have become one of the main conceptual frameworks to measure the importance of the contributions of different macroeconomic shocks.[2]

Now I turn to models in which monetary policy nonneutrality is due to the presence of nominal rigidities on the price (and wage) setting, as opposed to informational ones. I also regard these features as examples of the role of "standard of value of money" or unit of account. The *JPE* published the seminal paper by Rotemberg (1982), constructing a quantitative rational expectations sticky price model. Prices are sticky in the sense that competitive monopolistic firms face quadratic adjustment costs of changing them, giving rise to a type of forward-looking Phillips curve kind of relationship. The author finds a better fit when these costs are positive, that is, when prices are sticky. An alternative mechanism, whose reduced form is almost identical, was that proposed by Calvo (1983) and further developed by Yun (1996). The Calvo price setting results from the aggregation of firms that can change their prices only at idiosyncratic random times. The models in Rotemberg and Woodford (1997) and in Woodford (2001) use Calvo pricing to obtain a Phillips curve type relationship and also clarify and simplify the rest of the environment, producing the core of the neo-Keynesian model.

As mentioned in the conclusion of Rotemberg (1982), estimating and testing these types of models using microdata is desirable, since they may offer more direct evidence of the key pricing friction. The Calvo pricing assumption goes in this direction, since one can map the exogenous probability of keeping prices at a fixed value as related to actual durations of unchanged prices. A key input to this task is to have high-quality measurements of the frequency of price changes. In this regard Bils and Klenow (2004), also published in the *JPE*, provided a key input for this measurement for the US economy. While this is not the first paper to present micro evidence of price rigidities, using the microdata underlying the construction of the consumer price index provided the level of aggregation proper for macroeconomic analysis and thus initiated a large empirical literature that uses a variety of data sets and approaches. Surprisingly, Bils and Klenow found a relatively high frequency of price changes, although such a conclusion may depend on a more subtle definition of price changes.

[1] See Chari, Kehoe, and McGrattan (2000), which revisited the conditions required to obtain substantial propagation in a modern general equilibrium setup.

[2] Regrettably, the *JPE* missed out on this contribution.

The 2000s has seen an explosion of work on the effect of monetary policy based on the combination of assuming rational expectations, modeling sticky prices and/or wages using Calvo price setting, and using the basic structure of the neo-Keynesian model as developed in Woodford (2001). One, if not the, leading quantitative contribution on this area is Christiano, Eichenbaum, and Evans (2005), published in the *JPE*. This paper merges the structure of the real business cycle model with the core neo-Keynesian model, incorporating additional real frictions that enhanced the effects of nominal shocks. Additionally, the authors use VAR evidence from identified monetary shocks to estimate key parameters of the model, which in turn is used for quantitative evaluation of different policies.[3]

While Calvo pricing does provide a connection with microdata, these data strongly suggest the presence of idiosyncratic shocks. Moreover, the exogenous nature of the timing of price changes in Calvo pricing, while very tractable, is clearly a shortcut that may have important consequences. Indeed, Golosov and Lucas (2007) use a quantitative menu cost model, a model in which firms are subject to idiosyncratic costs and in which they must pay a fixed menu cost to change nominal product prices. The calibrated version of their model shows that while money is not neutral, it is almost so. Their contribution, together with the availability of rich microdata, has also started a literature that aims at using such data to better measure the impact of the frictions on price setting and to reconcile the small size of monetary shocks on output with the larger values typically estimated using identified VARs.

References

Barro, Robert J. 1978. "Unanticipated Money, Output, and the Price Level in the United States." *J.P.E.* 86 (4): 549–80.

Baumol, William J. 1952. "The Transactions Demand for Cash: An Inventory Theoretic Model." *Q.J.E.* 66 (4): 545–56.

Bils, Mark, and Peter J. Klenow. 2004. "Some Evidence on the Importance of Sticky Prices." *J.P.E.* 112 (5): 947–85.

Blanchard, Olivier J. 1985. "Debt, Deficits, and Finite Horizons." *J.P.E.* 93 (2): 223–47.

Calvo, Guillermo A. 1983. "Staggered Prices in a Utility-Maximizing Framework." *J. Monetary Econ.* 12 (3): 383–98.

Chari, V. V., Patrick J. Kehoe, and Ellen R. McGrattan. 2000. "Sticky Price Models of the Business Cycle: Can the Contract Multiplier Solve the Persistence Problem?" *Econometrica* 68 (5): 1151–79.

Christiano, Lawrence J., Martin Eichenbaum, and Charles L. Evans. 2005. "Nominal Rigidities and the Dynamic Effects of a Shock to Monetary Policy." *J.P.E.* 113 (1): 1–45.

[3] Smets and Wouters (2007) is an important complementary, and largely confirmatory, study of a very closely related model using a different econometric methodology.

Fischer, Stanley. 1977. "Long-Term Contracts, Rational Expectations, and the Optimal Money Supply Rule." *J.P.E.* 85 (1): 191–205.
Friedman, Milton. 1968. "The Role of Monetary Policy." *A.E.R.* 58 (1): 1–17.
———. 1977. "Nobel Lecture: Inflation and Unemployment." *J.P.E.* 85 (3): 451–72.
Golosov, Mikhail, and Robert E. Lucas Jr. 2007. "Menu Costs and Phillips Curves." *J.P.E.* 115:171–99.
Kiyotaki, Nobuhiro, and Randall Wright. 1989. "On Money as a Medium of Exchange." *J.P.E.* 97 (4): 927–54.
Lagos, Ricardo, and Randall Wright. 2005. "A Unified Framework for Monetary Theory and Policy Analysis." *J.P.E.* 113 (3): 463–84.
Lucas, Robert E., Jr. 1972. "Expectations and the Neutrality of Money." *J. Econ. Theory* 4 (2): 103–24.
———. 1996. "Nobel Lecture: Monetary Neutrality." *J.P.E.* 104 (4): 661–82.
Lucas, Robert E., Jr., and Edward C. Prescott. 1971. "Investment under Uncertainty." *Econometrica* 39 (5): 659–81.
Lucas, Robert E., Jr., and Nancy L. Stokey. 1987. "Money and Interest in a Cash-in-Advance Economy." *Econometrica* 55 (3): 491–513.
Muth, John F. 1961. "Rational Expectations and the Theory of Price Movements." *Econometrica* 29 (3): 315–35.
Rotemberg, Julio J. 1982. "Sticky Prices in the United States." *J.P.E.* 90 (6): 1187–1211.
Rotemberg, Julio J., and Michael Woodford. 1997. "An Optimization-Based Econometric Framework for the Evaluation of Monetary Policy." In *NBER Macroeconomics Annual*, vol. 12, edited by Ben Bernanke and Julio J. Rotemberg, 297–346. Cambridge, MA: MIT Press.
Samuelson, Paul A. 1958. "An Exact Consumption-Loan Model of Interest with or without the Social Contrivance of Money." *J.P.E.* 66 (6): 467–82.
Sargent, Thomas J., and Neil Wallace. 1975. "'Rational' Expectations, the Optimal Monetary Instrument, and the Optimal Money Supply Rule." *J.P.E.* 83 (2): 241–54.
Sidrauski, Miguel. 1967a. "Inflation and Economic Growth." *J.P.E.* 75 (6): 796–810.
———. 1967b. "Rational Choice and Patterns of Growth in a Monetary Economy." *A.E.R.* 57 (2): 534–44.
Sims, Christopher A. 1980. "Macroeconomics and Reality." *Econometrica* 48 (1): 1–48.
Smets, Frank, and Rafael Wouters. 2007. "Shocks and Frictions in US Business Cycles: A Bayesian DSGE Approach." *A.E.R.* 97 (3): 586–606.
Taylor, John B. 1980. "Aggregate Dynamics and Staggered Contracts." *J.P.E.* 88 (1): 1–23.
Tobin, James. 1956. "The Interest Elasticity of Transactions Demand for Money." *Rev. Econ. and Statis.* 38 (3): 241–47.
Woodford, Michael. 2001. "The Taylor Rule and Optimal Monetary Policy." *A.E.R.* 91 (2): 232–37.
Yun, Tack. 1996. "Nominal Price Rigidity, Money Supply Endogeneity, and Business Cycles." *J. Monetary Econ.* 37:345–70.

Memories of Friedman and Patinkin

Robert E. Lucas Jr.

University of Chicago

During the period between the two world wars the University of Chicago produced an extraordinary group of monetary economists. For these notes, I will concentrate on two of them: Milton Friedman and Don Patinkin. I knew them both well, and both were important to my own economic growth. Both of them are remembered more for their books than for their journal articles, but the *Journal of Political Economy* published them both, including the interesting exchanges that I will discuss here.

Both Friedman and Patinkin did graduate work at Chicago, Friedman in the 1930s and Patinkin in the 1940s. Patinkin got his Chicago PhD in 1947, working under Oscar Lange. Friedman got his degree from Columbia in 1940, supervised by Simon Kuznets. The two did not overlap at Chicago, but both of them recalled classes on monetary economics with Henry Simons and Lloyd Mints and, less directly, Jacob Viner and Frank Knight.

In 1956 Friedman published *Studies in the Quantity Theory of Money—a Restatement*, a book consisting of a long introduction by Friedman himself, followed by the dissertations of four of his students: Phillip Cagan, John Klein, Eugene Lerner, and Richard Selden. These four dissertations—the first fruit of the Chicago Money and Banking Workshop—are stunning examples of economics at its best. I will come back to them, but first I want to review Friedman's introduction, which was focused almost entirely on clarifying and reviving a version of the quantity theory of money.

Friedman began with the concern that it "is clear that the general approach (the quantity theory) fell into disrepute after the crash of 1929 and the subsequent Great Depression and only recently has been slowly re-emerging into professional respectability" (3). One source of this disrepute was "the proponents of the new income-expenditure approach" who described versions of the quantity theory that were "an atrophied and rigid charicature." Friedman argued that Chicago economists—mainly Simons, Mints, and Knight—had formulated a more sophisticated and useful version.[1]

This was his first attack on the economics of Keynes. Later attacks came in his 1970 *JPE* paper "A Theoretical Framework for Monetary Analysis"

[1] This Chicago version of the quantity theory has been discussed in much more detail by many authors. See in particular Laidler (2010), Nelson (2017), and Tavlas (2017).

and again in a 1971 extension, also in the *JPE*, "A Monetary Theory of Nominal Income."

At this point Don Patinkin, who had followed all three of these statements of Friedman's, lost patience. In 1972 the *JPE* published Patinkin's "Friedman on the Quantity Theory and Keynesian Economics." Here is his abstract:

> The article is based on textual evidence from the quantity-theory and Keynesian literature. It shows, first, that the conceptual framework of a portfolio demand for money that Friedman denotes as the "quantity theory" is actually that of Keynesian economics. Conversely, Friedman detracts from the true quantity theory by stating that its formal short-run analysis assumes real output constant, while only prices change. Friedman also incorrectly characterizes Keynesian economics in terms of absolute price rigidity. He does this by overlooking the systematic analysis by Keynes and the Keynesians of the role of downward wage flexibility during unemployment, and of the "inflationary gap" during full employment. Otherwise Friedman's interpretation of Keynes is the standard textbook one of an economy in a "liquidity-trap" unemployment equilibrium. The author restates his alternative interpretation of Keynesian economics in terms of unemployment *dis*equilibrium. (1972, 883)

Patinkin went on to develop these assertions in detail. Later, in the text, he added that "it is obviously no criticism of Friedman—nor does it derogate from his stature as a monetary economist—to say that his analytical framework is Keynesian. All that is being criticized is Friedman's persistent refusal to recognize this is so" (886).

In fact this 1972 *JPE* issue contained, in addition to Patinkin, reactions to Friedman's 1971 paper from Karl Brunner and Allan Meltzer, James Tobin, and Paul Davidson. These were followed by an 85-page counterattack from Friedman. (The editors at the time were Robert Gordon and Harry Johnson.)

Friedman's 1970 and 1971 papers did not mention Simons, Mints, or Knight, nor did they refer explicitly to a Chicago tradition. This time Keynes was discussed at some length. But Friedman continued to refer to "the Keynesian challenge to the quantity theory" and to cast the two as incompatible contestants. A common reaction from Patinkin, Brunner and Meltzer, and Tobin was to argue that Keynesian ideas and the quantity theory can and should be parts of a single model. Reading it now one expects some kind of unification, and there are times when this seems about to happen. But Friedman would not have it. The long 1971 debate began with confusion and ended there.

It is surprising to me that Friedman did not connect "the proponents of the new income-expenditure approach" to the national income account time series that Simon Kuznets had created in the late 1920s (and that the US government has maintained ever since). These data opened up an exciting new world for economists who wanted measurements on the state of the economy as a whole. Kuznets's data included "real" series only: national money supply data were still in the future. The model-building pioneers of quantitative macroeconomics—Jan Tinbergen, Lawrence Klein, and soon many others—worked with Kuznets's data because they were the only good data they had. I think this was all there was to the "new income-expenditure approach."

If so, then what was needed to restore the quantity theory was to construct time series on the money supply at the same level of accuracy as Kuznets's data on real series. And this is exactly what Friedman's students did in the substantive chapters of *Studies in the Quantity Theory of Money*. All four dissertations constructed aggregate, economywide time series on some well-defined measure of a money supply and measures of nominal prices. Cagan's study of postwar hyperinflations provides monthly time series—suitable for his high-frequency data—carefully documented, for prices, measures of cash balances, and real per capita incomes for seven countries. There is an explicit theoretical model—set out and calibrated. Comparisons of theory and time-series data are shown graphically and assessed econometrically.

Cagan's dissertation was the crown jewel of *Studies in the Quantity Theory of Money*, a breathtaking achievement that is still drawn on. The other three students also produced new monetary time series, shedding light on interesting situations. John Klein analyzed Germany from 1932 to 1944. Eugene Lerner studied the Southern Confederacy of 1861–65 (see also his 1954 *JPE* paper). Richard Selden's "Monetary Velocity in the United States" covered 1839–1951. In his introduction, Friedman also provided a valuable discussion of the way these very different situations all served as natural experiments. What Friedman and his students had achieved in part, then, was to do for aggregate monetary theory what Kuznets had done for the aggregate real income and product accounts. They created a synthesized "money" consisting of many component assets that can be measured and add up to a whole, just as Kuznets had done with "consumption" and other real aggregates.

The typology of M0, M1, M2, and so forth was not available in 1956. The concept M1 was initiated by Homer Jones at the St. Louis Fed in 1960. Many people were involved in developing it further. Friedman, along with Allan Meltzer, Karl Brunner, and others, were supporters and users.[2] Pa-

[2] I thank William Barnett and Stephen Williamson for information on the role of the St. Louis Fed. It is a fascinating story just touched on here.

tinkin was right, I think, to insist that monetary theory can fit quite nicely with Keynesian ideas. Friedman and his students showed how to do it with actual time-series evidence.

References

Brunner, Karl, and Allan H. Meltzer. 1972. "Friedman's Monetary Theory." *J.P.E.* 80 (September/October): 837–51.
Davidson, Paul. 1972. "A Keynesian View of Friedman's Theoretical Framework for Monetary Analysis." *J.P.E.* 80 (September/October): 864–82.
Friedman, Milton, ed. 1956. *Studies in the Quantity Theory of Money—a Restatement.* Chicago: Univ. Chicago Press.
———. 1970. "A Theoretical Framework for Monetary Analysis." *J.P.E.* 78 (March/April): 193–238.
———. 1971. "A Monetary Theory of Nominal Income." *J.P.E.* 79 (March/April): 323–37.
———. 1972. "Comments on the Critics." *J.P.E.* 80 (September/October): 906–50.
Laidler, David. 2010. "Chicago Monetary Traditions." In *The Elgar Companion to the Chicago School of Economics*, edited by Ross B. Emmett. Northampton, MA: Elgar.
Lerner, Eugene. 1954. "The Monetary and Fiscal Programs of the Confederate Government." *J.P.E.* 62 (December): 506–22
Nelson, Edward. 2017. "Milton Friedman and Economic Debate in the United States." Manuscript, Board Governors, Fed. Reserve System.
Patinkin, Don. 1972. "Friedman on the Quantity Theory and Keynesian Economics." *J.P.E.* 80 (September/October): 883–905.
Tavlas, George S. 2017. "The Group: The Making of the Chicago Monetary Tradition, 1927–36." Manuscript, Bank of Greece.
Tobin, James. 1972. "Friedman's Theoretical Framework." *J.P.E.* 80 (September/October): 852–63.

Labor Markets

Robert Shimer

University of Chicago

The *Journal of Political Economy* has been a crucial outlet for pathbreaking articles on the determination of employment, wages, and inflation. The *JPE* published what was arguably the key paper forecasting the demise of

a stable Phillips curve relationship, that is, the end of the trade-off between employment and inflation. It also published the first theory of intertemporal substitution in labor supply and many subsequent papers that sought to quantify the elasticity of labor supply. It has been an early and important source of papers that examined the impact of information frictions on labor market outcomes. And the *JPE* has published key papers that improved our understanding of the sources of fluctuations in employment over the business cycle. The following paragraphs expand on those points.

Phelps (1968) sought to understand whether the "Phillips trade-off [was] real, serious, and not misleading" (681). Phelps recognized that low unemployment may induce firms to raise wages in an effort to attract and retain workers. In the most innovative part of his paper, Phelps developed a notion of a "macroequilibrium" in which actual and expected wage growth are equal to each other. This was a key component of the citation for Phelps's Nobel Prize in 2006. Although this paper does not have the mathematical sophistication of later studies using rational expectations, Phelps's argument predicted that maintaining an unemployment rate below the steady-state equilibrium value would necessitate spiraling increases in inflation. This notion of an expectations-augmented Phillips curve was of course also emphasized in Friedman's presidential address to the American Economic Association as well as in his Nobel lecture (Friedman 1968, 1977).

Given the *JPE*'s early recognition of the limitations of conventional Keynesian analysis of the Phillips curve and the economists at the University of Chicago's tight connection to the rational expectations revolution, it is equally interesting to note the papers on the employment-inflation trade-off that the *JPE* did not publish. The most prominent and important of these is Lucas (1972). This omission appears to reflect the reluctance of the journal at the time to engage in the types of mathematical arguments that rational expectations models with strong microeconomic foundations embraced. Indeed, papers with rational expectations that the *JPE* did publish at the time, such as Sargent and Wallace (1975), used ad hoc models rather than the optimizing framework embraced by Lucas.

Sophisticated models of labor market equilibrium do appear in many more recent papers published in the journal. And central to any notion of labor market equilibrium is the interaction between labor supply and labor demand. While there was historically relatively little controversy about the role of the marginal product of labor (or the marginal revenue product of labor) in determining labor demand, the history of models of labor supply is more interesting. In the 1960s, models of the long run, like Solow (1956), treated labor supply as fixed, while the Keynesian models that were used for understanding the short run treated labor supply

as perfectly elastic at some rigid real or nominal wage. In a seminal paper, Lucas and Rapping (1969) argued that a standard model of the household provides a strong microeconomic foundation for a dynamic theory of the elasticity of labor supply. The paper extends a utility-maximizing theory of the labor-leisure choice to a multiperiod framework and shows how current and future real wages, as well as interest rates and wealth, affect labor supply. This trade-off is central to all modern models of the labor market.

Subsequent to the largely theoretical contribution of Lucas and Rapping (1969), the *JPE* has been at the forefront of an empirical literature measuring how elastic labor supply is. In an important early contribution, MaCurdy (1981) developed an empirical framework for using variation in wages and hours worked over the life cycle to measure the Frisch (constant marginal utility of wealth) elasticity of labor supply. Using data on prime-aged men in the United States, he found that the Frisch elasticity was positive but small, in the range of 0.10–0.45. This paper spawned many follow-ups, some using similar sources of wage variation, others using different sources, such as changes in income tax rates. A prominent example is Altonji (1986). This paper deals with issues related to missing and mismeasured wage observations, as well as the components of future wage growth that are known to the worker but not to the econometrician. The careful analysis modestly lowers the range of plausible labor supply elasticities to 0–0.35.

Intertemporal elasticities of labor supply in this range present a challenge for business cycle models in which flexible wages equate labor supply and labor demand. These models rely on the assumption that the observed modest variation in wages induces substantial cyclical shifts in employment because labor supply is elastic (e.g., Kydland and Prescott 1982).

A possible resolution of this is that shifts between market and home production effectively raise the relevant elasticity. The modern theory of time allocation and home production is also closely linked to the *JPE* and economists at the University of Chicago. Becker (1965) proposed that time and market goods are inputs into the production of commodities, which in turn enter the welfare function. Becker's theory was both general and abstract, and so its implications for labor supply were initially unclear. Gronau (1977) parameterized this framework more tightly, in particular distinguishing between home production and leisure: "work at home (like work in the market) is something one would rather have somebody else do for one (if the cost were low enough), while it would be almost impossible to enjoy leisure through a surrogate" (1104). In turn, Benhabib, Rogerson, and Wright (1991) argued that introducing home production into the Kydland and Prescott (1982) real business cycle model significantly

improved its ability to match the business cycle facts, in large part by effectively raising the intertemporal elasticity of labor supply.[1]

In a paper that was instrumental to his 1982 Nobel Prize, Stigler (1961) pioneered the study of information frictions in markets. In Stigler (1962), he applied these ideas to the labor market, starting a research agenda that now offers a complementary explanation for the behavior of employment and the determination of wages. The 1962 paper notes the existence of pure wage dispersion, meaning that different firms offer the same worker different wages. He showed how dispersion in wage offers gives workers an incentive to get multiple offers so as to obtain higher wages on average. He also examines how the optimal number of offers depends on the amount of wage dispersion, the cost of search, and the number of years a job is expected to last. He observed that employers search for workers at the same time as workers search for employers. Finally, he argues that large firms pay higher wages because they face a higher cost of monitoring and screening workers. This employer size–wage effect was only carefully documented and explored nearly three decades later (Brown and Medoff 1989).

While Stigler (1962) discussed search both by workers and by firms, he was not able to articulate a framework in which both happened simultaneously. Doing so was the major achievement of Diamond (1982), a paper that features prominently in his 2010 Nobel Prize citation. Although the model in that paper was stylized and not explicitly cast in terms of the labor market, it has frequently been reinterpreted in those terms. The paper argues that thick market effects are a natural implication of search frictions. There is no point in searching for a job if firms are not looking to hire workers, and vice versa. The paper also shows that thick market effects can generate multiple Pareto-rankable steady-state equilibria. In one equilibrium, search intensity is low because everyone assumes that it will be difficult to find a trading partner and so engages in minimal search activity. In another, search intensity is high for the opposite reason. By construction, wages and prices are fixed in Diamond's model, and so the model can generate large swings in productive employment without needing any movement in wages and without relying on elastic labor supply.

Search is not the only important friction in the labor market. Oi (1962) stressed that training costs are substantial, making it costly to hire and fire a worker. In his language, this turns labor into a quasi-fixed factor.

[1] Other important *JPE* papers have followed up on other implications of Becker's and Gronau's work. For example, Aguiar and Hurst (2005) showed that time allocation and home production, rather than myopia, are important for understanding the drop in expenditures on consumption goods at retirement.

A number of papers examine how these two hiring frictions, search and training costs, affect the determination of employment and wages. Jovanovic (1979) studied an environment in which the quality of a job match is an experience good. Workers and firms jointly learn about the quality of the match from the evolution of the cumulative amount of output produced. Because of hiring frictions, a little bit of bad news will not be enough to cause the pair to separate. Nevertheless, with enough bad news, the worker prefers to quit and look for another job. Stochastic match quality and endogenous separations help break the link between individual employment histories and wages. This model of symmetric learning has proved very useful for understanding how turnover varies with job tenure.

Azariadis (1975) and Lazear (1979) examine implicit and explicit long-term contracts between workers and firms. The starting point for both papers is the puzzling question of why firms lay off workers rather than cut wages. For Azariadis, these layoffs happen when the firm contracts. For Lazear, they occur when a worker retires. Azariadis's argument is somewhat simpler, since it does not rely on the incentive issues emphasized by Lazear. Azariadis explained that long-term contracting considerations can induce firms to lay off workers while maintaining the wage of the workers they keep. One way to think about this is in terms of optimal contracting with (possibly limited) commitment. To the extent that workers are less able to insure themselves against idiosyncratic fluctuations in income, firms should compete for workers by promising to smooth wages as long as the employment relationship lasts. This promise has some chance of being credible if it is costly for the firm to fire one worker and rehire an identical one, that is, if there is a hiring friction. But if business conditions become too adverse, the optimal contract may necessitate firing the worker. Thus Azariadis's model again predicts smooth wages and large employment fluctuations, reflecting the fact that the decision to work at a particular firm is not determined in a spot market with reference only to the current wage.

A related issue is the structure of pay within organizations. Personnel managers believe that pay equality is important within an organization, at least for workers charged with doing similar tasks. A common view is that this reflects morale concerns within the firm. Lazear (1989) argued that this explanation is at best incomplete, since highly productive workers should feel discouraged by a compressed wage scale, and indeed their discouragement may be particularly costly to the firm. He instead emphasized the potential destructive role of competition between coworkers. When one coworker can sabotage another's output, tournaments may be counterproductive. This paper offers a distinct explanation for why wages are smooth within employment relationships.

Finally, Lilien (1982) proposed that sectoral shifts are important for understanding employment fluctuations. The basic idea is that a market

economy always moves workers from declining sectors into expanding ones. Owing to the presence of search frictions, this takes time. Indeed, even though the mean duration of job search may be on the order of 2 or 3 months in the US economy, when a worker moves from a declining sector to an expanding one, he may cycle through several jobs before finding a stable one, for the reasons that Jovanovic (1979) emphasized. This does not imply that worker reallocation is socially undesirable. But it does imply that if the amount of reallocation is unusually high at some point in time, employment and output will be low; that is, there will be a recession. Thus search frictions and sectoral shocks naturally give rise to aggregate fluctuations.

Lilien's emphasis on sectoral shifts rather than aggregate fluctuations has proved controversial. Arguably his most lasting insight may lie in how we think about and model fluctuations in employment. Echoing Phelps (1968) but armed with more modern terminology and modeling tools, Lilien wrote, "Given a definition of the natural rate [of unemployment] based on microeconomic foundations, much of cyclical unemployment is better described as fluctuations of the natural rate itself" (1982, 778). This is indeed the current view of equilibrium unemployment.

References

Aguiar, Mark, and Erik Hurst. 2005. "Consumption versus Expenditure." *J.P.E.* 113 (5): 919–48.

Altonji, Joseph G. 1986. "Intertemporal Substitution in Labor Supply: Evidence from Micro Data." *J.P.E.* 94, no. 3, pt. 2 (June): S176–S215.

Azariadis, Costas. 1975. "Implicit Contracts and Underemployment Equilibria." *J.P.E.* 83 (6): 1183–1202.

Becker, Gary S. 1965. "A Theory of the Allocation of Time." *Econ. J.* 75 (299): 493–517.

Benhabib, Jess, Richard Rogerson, and Randall Wright. 1991. "Homework in Macroeconomics: Household and Aggregate Fluctuations." *J.P.E.* 99 (6): 1166–87.

Brown, Charles, and James Medoff. 1989. "The Employer Size–Wage Effect." *J.P.E.* 97 (5): 1027–59.

Diamond, Peter A. 1982. "Aggregate Demand Management in Search Equilibrium." *J.P.E.* 90 (5): 881–94.

Friedman, Milton. 1968. "The Role of Monetary Policy." *A.E.R.* 58 (1): 1–17.

———. 1977. "Nobel Lecture: Inflation and Unemployment." *J.P.E.* 85 (3): 451–72.

Gronau, Reuben. 1977. "Leisure, Home Production, and Work—the Theory of the Allocation of Time Revisited." *J.P.E.* 85 (6): 1099–1123.

Jovanovic, Boyan. 1979. "Job Matching and the Theory of Turnover." *J.P.E.* 87, no. 5, pt. 1 (October): 972–90.

Kydland, Finn E., and Edward C. Prescott. 1982. "Time to Build and Aggregate Fluctuations." *Econometrica* 50 (6): 1345–70.

Lazear, Edward P. 1979. "Why Is There Mandatory Retirement?" *J.P.E.* 87 (6): 1261–84.

———. 1989. "Pay Equality and Industrial Politics." *J.P.E.* 97 (3): 561–80.

Lilien, David M. 1982. "Sectoral Shifts and Cyclical Unemployment." *J.P.E.* 90 (4): 777–93.
Lucas, Robert E., Jr. 1972. "Expectations and the Neutrality of Money." *J. Econ. Theory* 4 (2): 103–24.
Lucas, Robert E., Jr., and Leonard A. Rapping. 1969. "Real Wages, Employment, and Inflation." *J.P.E.* 77 (5): 721–54.
MaCurdy, Thomas E. 1981. "An Empirical Model of Labor Supply in a Life-Cycle Setting." *J.P.E.* 89 (6): 1059–85.
Oi, Walter Y. 1962. "Labor as a Quasi-Fixed Factor." *J.P.E.* 70 (6): 538–55.
Phelps, Edmund S. 1968. "Money-Wage Dynamics and Labor-Market Equilibrium." *J.P.E.* 76, no. 4, pt. 2 (August): 678–711.
Sargent, Thomas J., and Neil Wallace. 1975. "'Rational' Expectations, the Optimal Monetary Instrument, and the Optimal Money Supply Rule." *J.P.E.* 83 (2): 241–54.
Solow, Robert M. 1956. "A Contribution to the Theory of Economic Growth." *Q.J.E.* 70 (1): 65–94.
Stigler, George J. 1961. "The Economics of Information." *J.P.E.* 69 (3): 213–25.
———. 1962. "Information in the Labor Market." *J.P.E.* 70, no. 5, pt. 2 (October): 94–105.

Chicago Labor Economics

James J. Heckman

University of Chicago

Starting in the 1950s, economists influenced by the Chicago approach to economics revolutionized the field of labor economics and developed concepts and evidence that now shape the thought of the entire profession.[1] Not all economists working in this tradition were faculty, postdocs, or students at Chicago, although the dominant figures were.

Chicago labor economics was a natural by-product of Chicago economics. The same general principles were relentlessly applied in all fields, and labor was no exception. Chicago economics emphasized the value of economic models in interpreting and guiding collection of data and making forecasts and constructing policy counterfactuals. The contributions of labor economists working in the Chicago tradition are so extensive that no short survey can do justice to them. Instead, I will discuss the core prin-

[1] For a comprehensive history of Chicago labor economics, see Kaufman (2010).

ciples of the enterprise that have been relentlessly applied with great theoretical and empirical success.

The Chicago approach was in stark contract with the prevalent methods used in the labor economics of that time: industrial relations. This was a largely atheoretical institutionalist approach that focused on thick description, not explanation. It made generalizations to summarize data, but the summaries were largely disconnected from analytical economics and often from each other (see, e.g., Kerr et al. 1960). Its hostility to economic theory is well exemplified by the paper of Richard Lester (1946), who interviewed managers of firms, literally asking them if they set wages to marginal products. Finding they did not report doing so, he pronounced that marginal analysis was useless for the study of labor markets.[2]

The Chicago approach emphasized the value of price theory in interpreting data. Milton Friedman was a major influence in economics at Chicago, and his approach to theory and empirical research set the tone.[3] In an essay on Wesley Clair Mitchell, he crystallized the Chicago approach to scientific economics:[4]

> The ultimate goal of science in any field is a theory—an integrated "explanation" of observed phenomena that can be used to make valid predictions about phenomena not yet observed. Many kinds of work can contribute to this ultimate goal and are essential for its attainment: the collection of observations about the phenomena in question; the organization and arrangement of observations and the extraction of empirical generalizations from them; the development of improved methods of measuring or analyzing observations; the formulation of partial or complete theories to integrate existing evidence. (Friedman 1950, 465)

In a dig at the institutionalists in the same essay, he quoted Marshall: "'The most reckless and treacherous of all theorists is he who professes to let facts and figures speak for themselves.' [Marshall 1885] And, one might add 'The most reckless and treacherous of all empirical workers

[2] This paper provoked numerous responses. See, e.g., Machlup (1947) and Friedman (1953).

[3] Friedman made basic contributions to labor economics. Friedman and Kuznets (1945) pioneered the study of panel data income dynamics and introduced the notion of firm-specific human capital. Friedman (1951) studied the impact of unionism. Prior to the 1950s, Chicago economist Paul Douglas studied production functions, labor supply, and unionism. Frank Knight and Henry Simons wrote polemical essays on monopoly unionism.

[4] He later expanded on this essay in his well-known essay "The Methodology of Positive Economics" (Friedman 1953).

is he who formulates theories to explain observations that are the product of careless and inaccurate empirical work'" (Friedman 1950, 465–66).

The interplay between data and theory was the hallmark of the Chicago approach. The problems investigated were important for understanding the economy and public policy. Careful documentation of these problems and new empirical regularities were highly valued.

However, economic analysis did not stop there. Interpretation of data—understanding the problems being studied and the mechanisms generating them—was a crucial part of policy analysis. All of these activities were essential for the scientific analysis of counterfactuals that is the basis for rigorous policy analysis. Both data and theory were taken very seriously. Economic theory was viewed as an engine for analysis and empirical discovery, and not as an end itself. At the same time, great value was placed on careful and comprehensive empirical work that produced convincing evidence. Great value was also placed on rigorous economic theorizing that made lasting contributions to the understanding of the economy. Models that were discordant with data were revised and tested on the same and new data.

Theory was subjected to rigorous testing against the data and was used to parsimoniously explain phenomena within and across fields in economics, including labor economics. Only when the standard tools failed would the theory be amended. This approach was in stark contrast to that of the institutionalists who often favored ad hoc generalizations to "let the facts speak for themselves." They typically made up new models one empirical finding at a time and lacked a common core of basic principles that applied across multiple domains.

Chicago-inspired labor economics applied price theory to the labor market and emphasized testing models on data and taking great care with empirical evidence. The first studies used consumer demand theory (the demand for leisure) to understand the relationship of labor supply with wages and asset income (Lewis 1956; Mincer 1962) and to estimate income and substitution effects.[5] Lewis (1963) investigated counterfactual union wage impacts.[6] A study that addressed the challenging policy question of whether there was an undersupply or oversupply of education led Becker (1962, 1964), a Chicago PhD, to create the edifice of modern human capital theory.

[5] Douglas (1934) and Schoenberg and Douglas (1937) were pioneering empirical papers on relations of wages (or earnings) to measures of labor supply. See Pencavel (1986) for a survey of labor supply.

[6] His neglected and densely written book defined and estimated economically grounded counterfactual union "effects" for a variety of market scenarios, including partial and general equilibrium. Lewis's (1963) framework is far richer than just the framework he used to measure union gaps that he estimated in his later work (Lewis 1986).

Becker's work entailed more than a direct application of consumer demand theory. New analyses were required to explain the relationship of human capital to earnings and other phenomena. True to Chicago tradition, human capital theory was parsimonious and had general applications. It interpreted empirical regularities on earnings, on-the-job training, life cycle wage growth, the quit and layoff decisions of workers and firms, and patterns of trade across countries using a core set of basic principles consistent with prior theory, but supplementing it when needed. It is a brilliant operationalization of Friedman's vision of Chicago economics.

Becker's work on fertility (Becker and Lewis 1973; Becker 1991) grew out of an early failure to explain the time series of fertility and income.[7] The systematic study of labor supply and fertility led to novel papers on the allocation of time and the creation of a new field of household economics.[8] At each step, theory was tested, and if it failed, it was modified, but only as needed. The core principles remained.

Becker was, of course, a uniquely creative figure. But the methodological principles of the Chicago approach were applied by many labor economists to explain a variety of phenomena in the labor markets. For example, Walter Oi (1962) and Sherwin Rosen (1968) applied and expanded production theory to analyze the determinants of labor demand using the Chicago approach.[9] Hedonic models of pricing of heterogeneous goods (and skills) explained variation in quality in both product markets and labor markets.[10] It explained apparent deviations from the law of one price by invoking quality variations as the source of these differences. In the same vein, Stigler's (1961) search theory explained price (and wage dispersion) not as a failure of competitive markets—as had the institutionalists—but as a consequence of costly search. Rosen's paper (with Ed Lazear) on tournaments (Lazear and Rosen 1981) was a basic contribution to understanding the determinants of compensation of workers.

Research on the economics of the family synthesized and expanded the Chicago portfolio (see Becker 1991). The three basic ingredients of Chicago economics were central to the field in his analyses: (*a*) stable preferences for agents, (*b*) agents responding to incentives, and (*c*) equilibrium. Studying equilibrium in marriage markets that matched indivisible agents required new tools not standard in conventional price theory but present in earlier work by Koopmans and Beckmann (1957). Chiappori (2017) defines the state of the art.

[7] See Heckman (2015) for a discussion of the evolution of Becker's thought.
[8] Mincer (1962, 1963) played a seminal role in this activity.
[9] See Hamermesh (1993) for a survey.
[10] See Lewis (1969), Welch (1969), and Rosen (1974) for early work.

Another example is the work of Becker and Tomes (1979, 1986) analyzing intergenerational mobility. Basic tools of economics were adapted and applied to the data to investigate the persistence of family influence across generations. As discussed by Mogstad (2017), this is an active area of research that integrates theory and empirical work and extends theory.

Many other examples of labor economics in the Chicago tradition could be given, but space constraints prevent this. The guiding principles in each of these studies are the same. Theory is used to interpret data. Data are used to test theory. Understanding the mechanisms producing empirically estimated "effects" is essential for principled counterfactual analysis and for explaining phenomena. It was never enough to say an intervention "worked." It was required that analysts understand the mechanisms producing "the facts," their generality across multiple empirical domains, and their relevance for public policy.

References

Becker, Gary S. 1962. "Irrational Behavior and Economic Theory." *J.P.E.* 70:1–13.
———. 1964. *Human Capital: A Theoretical and Empirical Analysis, with Special Reference to Education*. New York: NBER.
———. 1991. *A Treatise on the Family*. Enl. ed. Cambridge, MA: Harvard Univ. Press.
Becker, Gary S., and H. Gregg Lewis. 1973. "On the Interaction between the Quantity and Quality of Children." *J.P.E.* 81, no. 2, pt. 2 (April): S279–S288.
Becker, Gary S., and Nigel Tomes. 1979. "An Equilibrium Theory of the Distribution of Income and Intergenerational Mobility." *J.P.E.* 87 (6): 1153–89.
———. 1986. "Human Capital and the Rise and Fall of Families." *J. Labor Econ.* 4, no. 3, pt. 2:S1–S39.
Chiappori, Pierre-André. 2017. *Matching with Transfers: The Economics of Love and Marriage*. Princeton, NJ: Princeton Univ. Press.
Douglas, Paul H. 1934. *Theory of Wages*. New York: Macmillan.
Friedman, Milton. 1950. "Wesley C. Mitchell as an Economic Theorist." *J.P.E.* 58 (6): 465–93.
———. 1951. "Some Comments on the Significance of Labor Unions for Economic Policy." In *The Impact of the Union: Eight Economic Theorists Evaluate the Labor Union*, edited by David McCord Wright, 204–34. New York: Kelley & Millman.
———. 1953. "The Methodology of Positive Economics." In *Essays in Positive Economics*, edited by Milton Friedman. Chicago: Univ. Chicago Press.
Friedman, Milton, and Simon Smith Kuznets. 1945. *Income from Independent Professional Practice*. New York: NBER.
Hamermesh, Daniel S. 1993. *Labor Demand*. Princeton, NJ: Princeton Univ. Press.
Heckman, James J. 2015. "Introduction to a Theory of the Allocation of Time by Gary Becker." *Econ. J.* 125 (583): 403–9.
Kaufman, Bruce. 2010. "Chicago and the Development of Twentieth Century Labor Economics." In *The Elgar Companion to the Chicago School of Economics*, edited by Ross B. Emmett, 128–51. Northampton, MA: Elgar.

Kerr, Clark, Frederick H. Harbison, John T. Dunlop, and Charles A. Myers. 1960. *Industrialism and Industrial Man: The Problems of Labor and Management in Economic Growth.* Cambridge, MA: Harvard Univ. Press.

Koopmans, Tjalling C., and Martin Beckmann. 1957. "Assignment Problems and the Location of Economic Activities." *Econometrica* 25 (1): 53–76.

Lazear, Edward P., and Sherwin Rosen. 1981. "Rank-Order Tournaments as Optimum Labor Contracts." *J.P.E.* 89 (5): 841–64.

Lester, Richard A. 1946. "Shortcomings of Marginal Analysis for Wage-Employment Problems." *A.E.R.* 36 (1): 63–82.

Lewis, H. Gregg. 1956. "Hours of Work and Hours of Leisure." In *Annual Proceedings of the Industrial Relations Research Association,* 196–206. Madison, WI: Indus. Res. Assoc.

———. 1963. *Unionism and Relative Wages in the United States: An Empirical Inquiry.* Chicago: Univ. Chicago Press.

———. 1969. "Employer Interests in Employee Hours of Work." *Cuadernos de Economia* 18:38–54.

———. 1986. *Union Relative Wage Effects: A Survey.* Chicago: Univ. Chicago Press.

Machlup, Fritz. 1947. "Rejoinder to an Antimarginalist." *A.E.R.* 37 (1): 148–54.

Marshall, Alfred. 1885. *The Present Position of Economics: An Inaugural Lecture.* London: Macmillan.

Mincer, Jacob. 1962. "Labor Force Participation of Married Women." In *Aspects of Labor Economics*, edited by H. Gregg Lewis. Princeton, NJ: Princeton Univ. Press.

———. 1963. "Market Prices, Opportunity Costs, and Income Effects." *Measurement Econ.* 48 (1): 67–82.

Mogstad, Magne. 2017. "The Human Capital Approach to Intergenerational Mobility." *J.P.E.* 125 (6): 1862–68.

Oi, W. Y. 1962. "Labor as a Quasi-Fixed Factor." *J.P.E.* 70 (6): 538–55.

Pencavel, John H. 1986. "Labor Supply of Men: A Survey." In *Handbook of Labor Economics*, vol. 3, edited by Orley Ashenfelter and Richard Layard. Amsterdam: North-Holland.

Rosen, Sherwin. 1968. "Short-Run Employment Variation on Class-I Railroads in the U.S., 1947–1963." *Econometrica* 36 (3/4): 511–29.

———. 1974. "Hedonic Prices and Implicit Markets: Product Differentiation in Pure Competition." *J.P.E.* 82 (1): 34–55.

Schoenberg, Erika H., and Paul H. Douglas. 1937. "Studies in the Supply Curve of Labor: The Relation in 1929 between Average Earnings in American Cities and the Proportions Seeking Employment." *J.P.E.* 45 (1): 45–79.

Stigler, George J. 1961. "The Economics of Information." *J.P.E.* 69 (3): 213–25.

Welch, Finis. 1969. "Linear Synthesis of Skill Distribution." *J. Human Resources* 4 (3): 311–27.

Keeping the ECON in Econometrics: (Micro-)Econometrics in the *Journal of Political Economy*

Stephane Bonhomme

University of Chicago

Azeem M. Shaikh

University of Chicago

In 1970, John Siegfried wrote an instructive short note in the miscellany section of the *Journal of Political Economy* titled "A First Lesson in Econometrics" (Siegfried 1970). The author starts by writing the equation "1 + 1 = 2" but immediately argues that "every budding econometrician must learn early that it is never in good taste to express the sum of two quantities in [this] form" (1378). He then produces two pages of intricate derivations to arrive at an equivalent but extremely cumbersome expression.[1] From the publication of this note, it is reasonable to infer that the *JPE*'s editorial team at the time had some level of distrust in sophisticated econometric analysis.[2] Shortly thereafter, however, the journal began to play a key role in the development of several, novel econometric ideas.

Compared to many of its competitors, the type of econometric research the *JPE* has published has two distinctive features. The first one is the promotion of a type of econometric work that is tightly connected to economic models. In particular, the *JPE* has been a leading vehicle for structural econometric modeling. The second main feature is the emphasis on empirical applications of the methodology. The *JPE* seldom publishes abstract econometric theory. Instead, it promotes econometric analysis mainly through applications. In agreement with the motto of

While the title echoes Leamer's (1983) "taking the con out of econometrics," the expression "keeping the econ" was previously used by Ehrlich and Liu (1999) in a paper that appeared in the *Journal of Law and Economics*, also published at the University of Chicago.

[1] To add to the irony of that note, Eldridge (2014) points out that Siegfried's formula is wrong because of a matrix algebra mistake.

[2] For example, the *JPE* is mentioned only in passing in Christ's (1994) historical account of the Cowles Commission between 1939 and 1955, during which it was hosted at the University of Chicago.

the Cowles Commission, the *JPE*'s style of econometrics is one in which theory and measurement go hand in hand.

Since trying to review all of the econometrics research in the *JPE* would be a daunting task, we will focus on only a handful of contributions, each of which links economics and econometrics in particularly insightful ways. Such a choice necessarily means leaving aside a large number of equally important and influential contributions. In the same spirit, this review will be limited mainly to microeconometric applications, abstracting from key contributions to time-series econometrics, macroeconometrics, and finance that have appeared in the journal.

Econ Meets Metrics: An Econometric Model of Marriage

Becker's (1973, 1974) classical theory of marriage appeared in the *JPE* and is considered a landmark of the journal. Becker proposed a static model of the marriage market in which agents of different types, when matched, share surplus and can transfer utility to each other. Agents rank potential matches according to their preferences. In equilibrium, all matches are stable. Viewing marriage as a rational decision leading to an equilibrium distribution of matches has strong empirical appeal. For example, the model could be used to understand the effect of divorce laws or changes in contraception technology on marriage patterns. Devising an empirical counterpart to the Becker model, however, remained an unsolved question for a long time.

The *JPE* has been a pioneer in the structural econometric analysis of marriage markets, and more generally, it has published some of the most innovative and accomplished work in structural econometrics. The structural approach tries to build and exploit a tight link between the economic model and the empirical econometric model. Its main goal is to estimate primitive structural parameters with the hope that such parameters are invariant to policy and can be used for counterfactual predictions.

Choo and Siow (2006) proposed a structural econometric model of marriage. They completed the Becker theory to make it an econometric model that could be taken to the data. In doing so they faced several challenges: first, how to define an agent's type empirically and how to properly account for heterogeneity in preferences; second, how to deal with the fact that, typically, data on transfers within couples are not observed by the econometrician.

In the Choo and Siow framework, agents (i.e., men and women) have discrete types defined in terms of covariates such as age, education, ethnicity, or geography. Individuals of both genders have preferences for being married to different types of individuals. The structure of preferences is key to the tractability of the framework. Specifically, the utility of

a woman of type i (e.g., defined in terms of age and education) for marrying a man of type j is the sum of an (i, j)–specific systematic preference term, an (i, j)–specific transfer, and an idiosyncratic preference term. Building on McFadden (1974), the latter is assumed to follow a type I extreme value distribution, independent of the other terms and independent across options. For given values of systematic preferences and transfers, women's choices therefore take the form of logit demand models. Symmetrically, men's choices also take a logit form.

In contrast to single-agent choice models, in marriage markets two types of agents interact with each other and equilibrium constraints must be met. An innovation of the Choo and Siow framework is that transfers, which are not observed by the econometrician, are identified as the prices that clear the market and make women's and men's demands equal. As a result, the overall structure of the model is a two-sided logit demand model with equilibrium constraints.

A particular implication of the model is that it delivers a closed-form expression for the utility gains from marriage. A key equation in Choo and Siow (2006) shows that net gains from marriage for agents of types (i, j), relative to being single, can be written as a combination of quantities that are typically easy to estimate: the number of men and women of types (i, j) who are married, divided by the geometric average of the number of unmarried women of type i and unmarried men of type j. This transparent expression illustrates the power of a theory that delivers an economically interpretable quantity that can be directly estimated from the data. Taking advantage of this expression, in a way that is typical of many *JPE* papers, Choo and Siow illustrate the empirical relevance of their framework by estimating net gains from marriage in 1970–71 and 1981–82 by gender, age, and age of the spouse. In addition, they estimate how gains from marriage evolved after the legalization of abortion in *Roe v. Wade* by exploiting variation across states in a difference-in-differences fashion. This exercise nicely showcases the type of applications that can be studied with the framework.

Choo and Siow's (2006) seminal paper has already spurred a long legacy. Important work building on their framework has also appeared in the *JPE* (e.g., Chiappori and Oreffice 2008; Chiappori, Oreffice, and Quintana-Domeque 2012; Dupuy and Galichon 2014; Chiappori, Costa Dias, and Meghir, forthcoming; Fox, Yang, and Hsu, forthcoming).

Structural Econometric Models of the Labor Market

Among the many studies using the structural approach that have appeared in the *JPE*, models of education decisions and career choices have particularly benefited from the development of novel econometric methods. In such models, individuals choose to select into different ca-

reers depending on the costs they face and their expected returns. The econometrician must deal with the fact that returns and costs are largely unobserved. A central challenge is thus how to estimate rates of return to college or to a type of occupation in the presence of self-selection.

The 1970s and 1980s saw great progress on the understanding of selection models. Some of the key contributions appeared in the *JPE* (Gronau 1974; Heckman and Sedlacek 1985). Here, we focus on two contributions to structural econometrics that have built on this work. We note that there are several important *JPE* articles that are closely related to these two papers, such as Cameron and Heckman (1998, 2001), which we do not discuss here because of space constraints.

Willis and Rosen (1979) is an early example of a structural econometric model of education decisions.[3] The empirical model builds on the theory of comparative advantage. This work contains a number of strikingly modern econometric insights that are still relevant to today's research. A notable aspect concerns the way the authors specify and analyze the counterfactual—or "potential"—outcomes corresponding to different education choices. Their classical discussion of the role of exclusion restrictions, which are needed for credible identification, includes an exposition of the distinction between the marginal rate of return to investment and the marginal cost of funds due to Gary Becker. The role of functional form assumptions is also carefully discussed.

Another noteworthy aspect of the analysis is the way the economic model and the econometrics are linked to each other. The model's predictions are assessed for two outcomes: initial earnings in the life cycle and growth rates of earnings. The authors test several of the main structural restrictions of the model, but they do not interpret the fact that those restrictions are not violated as definitive success for the structural model. In the conclusion of the paper, the authors go one step further and include a small out-of-sample prediction exercise as a validation check.

Willis and Rosen's work was extended by Keane and Wolpin (1997), also published in the *JPE*, in several dimensions. Keane and Wolpin build a dynamic life cycle model of human capital investment in which individuals go to school and work in various occupations. Agents, who face uncertainty in the returns to their choices (i.e., wages), are forward looking and have rational expectations. Keane and Wolpin work under the constraint that the restrictions from the theory must be fully imposed in estimation. This structural approach to policy evaluation then allows them to perform counterfactual policy exercises. Taking such a setup to the data raises a number of econometric challenges. Setting up a coherent

[3] The reader may wonder about the unusual ordering of the authors' names. The initial footnote informs us that it was "selected by a random device."

structure that is rich enough to fit the complex heterogeneity in individual trajectories, while keeping the model tractable, remains a very difficult task today.

A central feature of Keane and Wolpin's econometric model is its dynamic nature. Experience is treated as a state variable, and agents form expectations about streams of income conditional on education and career choices. Estimation is based on maximum likelihood. Unlike previous structural dynamic discrete choice models, however, observed wages in their model are self-selected since work and experience are choices—and therefore endogenous, just like schooling. Endogeneity complicates estimation since it is not possible to proceed sequentially. For computation, Keane and Wolpin develop an approximate solution to the dynamic programming problem that allows them to address the computational curse of dimensionality.

A second key feature of the model is the presence of unobserved types. Borrowing from Heckman and Sedlacek (1985), Keane and Wolpin allow for self-selection in multiple dimensions of skill endowments. They deal with the presence of multidimensional heterogeneity using a finite mixture approach, which disciplines the different dimensions of heterogeneity.

Since its publication, Keane and Wolpin's framework has become a blueprint for structural econometric analysis in labor economics and elsewhere. Dynamic structural econometric modeling is still a vibrant research area, and some of the best research in this field is appearing in the *JPE*, such as two recent contributions by Adda, Dustmann, and Stevens (2017) and Heckman, Humphries, and Veramendi (forthcoming).

Partial Identification Meets Economic Theory

Partial identification is one of the most prominent recent themes in econometrics. The *JPE* played an early and important role in promoting the use of such methods in economics. The defining feature of a partially identified model is that the parameter of interest is not uniquely determined by the distribution of the observed data. Instead, it is limited only to a set of values. As we will see below, one of the main attractions of such methods is that they permit researchers to avoid making assumptions that may be deemed unpalatable for one reason or another but that might have been previously made for tractability.

While not fitting within our theme of microeconometrics, an early and influential example of partial identification can be found in Hansen and Jagannathan's (1991) landmark paper on the implications of security market data for asset pricing models. In the case of Hansen and Jagannathan, the parameters of interest are the means and standard deviations of the intertemporal marginal rates of substitution and the observed

data consist of security market data. They observe that under weak assumptions one can restrict the set of possible values for the parameters of interest using the Cauchy-Schwarz inequality. In contrast to previous approaches, Hansen and Jagannathan need not specify parametric functional forms for the intertemporal marginal rates of substitution. In fact, their analysis allows them to conclude that certain specifications are inconsistent with the observed data.

A more recent example of partial identification and one that fits more closely within our theme is Haile and Tamer's (2003) analysis of English or oral ascending auctions. In the case of Haile and Tamer, the parameter of interest is the distribution of bidders' (private) valuations and the observed data consist of bids. Instead of relying on a particular model of bidding behavior, such as the "button model" found in Milgrom and Weber (1982), which they argue may be inconsistent with the observed data, they instead propose assuming only that (i) bidders do not bid more than they are willing to pay and (ii) bidders do not allow an opponent to win at a price they are willing to beat. Using these minimal assumptions on bidder behavior and well-known results from the theory of order statistics, Haile and Tamer derive bounds on the distribution of valuations, which, in turn, permit them to construct bounds on the optimal reserve price in such auctions.

A common criticism of partial identification is that weak assumptions are often accompanied by limited ability to draw meaningful conclusions from the data. Hansen and Jagannathan and Haile and Tamer both show that this need not always be the case. In fact, in both settings, weak assumptions lead to remarkably sharp conclusions. In this way, both papers illustrate clearly the usefulness of approaching empirical work through the combined lens of economic theory and partial identification and have provided ample motivation for further applications of partial identification as well as the development of the accompanying theory for estimation and inference. Recent work in this spirit is the estimation of a structural voting model with deliberation using data from US appellate courts in Iaryczower, Shi, and Shum (forthcoming).

Conclusion

In this brief and partial review of microeconometrics in the *JPE*, we have highlighted the journal's focus on econometric frameworks that propose novel ways of taking fundamental economic theories to the data. Influential examples that we have not discussed include hedonic models (Rosen 1974; Ekeland, Heckman, and Nesheim 2004) and collective models (Chiappori 1992; Browning et al. 1994). In addition to this focus on the interplay between economic models and empirical analysis, we note that the *JPE* has also published several key contributions to traditional areas of economet-

rics such as instrumental variables (Altonji, Elder, and Taber 2005) and measurement error models (Erickson and Whited 2000).

Despite John Siegfried's warning against unnecessarily complicated econometrics, the recent history of the *JPE* demonstrates the power of careful econometric thinking in order to blend economic theory and empirical measurement. We hope that going forward the journal will continue and reinforce its role as a promoter of pioneering econometric research.

References

Adda, J., C. Dustmann, and K. Stevens. 2017. "The Career Costs of Children." *J.P.E.* 125 (2): 293–337.

Altonji, J. G., T. E. Elder, and C. R. Taber. 2005. "Selection on Observed and Unobserved Variables: Assessing the Effectiveness of Catholic Schools." *J.P.E.* 113 (1): 151–84.

Becker, G. S. 1973. "A Theory of Marriage: Part I." *J.P.E.* 81 (4): 813–46.

———. 1974. "A Theory of Marriage: Part II." *J.P.E.* 82, no. 2, pt. 2 (April): S11–S26.

Browning, M., F. Bourguignon, P.-A. Chiappori, and V. Lechene. 1994. "Income and Outcomes: A Structural Model of Intrahousehold Allocation." *J.P.E.* 102 (6): 1067–96.

Cameron, S. V., and J. J. Heckman. 1998. "Life Cycle Schooling and Dynamic Selection Bias: Models and Evidence for Five Cohorts of American Males." *J.P.E.* 106 (2): 262–333.

———. 2001. "The Dynamics of Educational Attainment for Black, Hispanic, and White Males." *J.P.E.* 109 (3): 455–99.

Chiappori, P.-A. 1992. "Collective Labor Supply and Welfare." *J.P.E.* 100 (3): 437–67.

Chiappori, P.-A., M. Costa Dias, and C. Meghir. Forthcoming. "The Marriage Market, Labor Supply, and Education Choice." *J.P.E.*

Chiappori, P.-A., and S. Oreffice. 2008. "Birth Control and Female Empowerment: An Equilibrium Analysis." *J.P.E.* 116 (1): 113–39.

Chiappori, P.-A., S. Oreffice, and C. Quintana-Domeque. 2012. "Fatter Attraction: Anthropometric and Socioeconomic Characteristics in the Marriage Market." *J.P.E.* 120:659–95.

Choo, E., and A. Siow. 2006. "Who Marries Whom and Why." *J.P.E.* 114 (1): 175–201.

Christ, C. F. 1994. "The Cowles Commission's Contributions to Econometrics at Chicago, 1939–1955." *J. Econ. Literature* 32 (1): 30–59.

Dupuy, A., and A. Galichon. 2014. "Personality Traits and the Marriage Market." *J.P.E.* 122:1271–1319.

Ehrlich, I., and Z. Liu. 1999. "Sensitivity Analyses of the Deterrence Hypothesis: Let's Keep the Econ in Econometrics." *J. Law and Econ.* 42 (S1): 455–88.

Ekeland, I., J. J. Heckman, and L. Nesheim. 2004. "Identification and Estimation of Hedonic Models." *J.P.E.* 112 (S1): S60–S109.

Eldridge, D. S. 2014. "A Comment on Siegfried's First Lesson in Econometrics." *Econ. Inquiry* 52 (1): 503–4.

Erickson, T., and T. M. Whited. 2000. "Measurement Error and the Relationship between Investment and *q*." *J.P.E.* 108 (5): 1027–57.

Fox, J. T., C. Yang, and D. H. Hsu. Forthcoming. "Unobserved Heterogeneity in Matching Games." *J.P.E.*

Gronau, R. 1974. "Wage Comparisons—a Selectivity Bias." *J.P.E.* 82 (6): 1119–43.
Haile, P. A., and E. Tamer. 2003. "Inference with an Incomplete Model of English Auctions." *J.P.E.* 111 (1): 1–51.
Hansen, L. P., and R. Jagannathan. 1991. "Implications of Security Market Data for Models of Dynamic Economies." *J.P.E.* 99 (2): 225–62.
Heckman, J. J., J. E. Humphries, and G. Veramendi. Forthcoming. "Returns to Education: The Causal Effects of Education on Earnings, Health, and Smoking." *J.P.E.*
Heckman, J. J., and G. Sedlacek. 1985. "Heterogeneity, Aggregation, and Market Wage Functions: An Empirical Model of Self-Selection in the Labor Market." *J.P.E.* 93 (6): 1077–1125.
Iaryczower, M., X. Shi, and M. Shum. Forthcoming. "Can Words Get in the Way? The Effect of Deliberation in Collective Decision Making." *J.P.E.*
Keane, M. P., and K. I. Wolpin. 1997. "The Career Decisions of Young Men." *J.P.E.* 105 (3): 473–522.
Leamer, E. E. 1983. "Let's Take the Con Out of Econometrics." *A.E.R.* 73 (1): 31–43.
McFadden, D. 1974. "Conditional Logit Analysis of Qualitative Choice Behavior." In *Frontiers in Econometrics*, edited by P. Zarembka. New York: Academic Press.
Milgrom, P. R., and R. J. Weber. 1982. "A Theory of Auctions and Competitive Bidding." *Econometrica* 50:1089–1122.
Rosen, S. 1974. "Hedonic Prices and Implicit Markets: Product Differentiation in Pure Competition." *J.P.E.* 82 (1): 34–55.
Siegfried, J. J. 1970. "A First Lesson in Econometrics." *J.P.E.* 78 (6): 1378–79.
Willis, R. J., and S. Rosen. 1979. "Education and Self-Selection." *J.P.E.* 87, no. 5, pt. 2 (October): S7–S36.

Life Cycle Wage Dynamics and Labor Mobility

Derek Neal

University of Chicago and National Bureau of Economic Research

Introduction

The *Journal of Political Economy* has published a number of seminal papers on individual investments in human capital and how these investments vary with ability, preferences, age, and other individual characteristics.[1]

I thank Stephane Bonhomme, Ronni Pavan, Canice Prendergast, and Christopher Taber for useful comments.

[1] Examples include Mincer (1958), Becker (1962), Ben-Porath (1967), Heckman (1976), and Rosen (1976).

However, most of this literature has nothing to say about labor mobility. For the most part, jobs play no role in this literature. All human capital is general, and the wage earned by a given worker at a particular point in time is the product of the stock of human capital she possesses and the rental rate on human capital. There is one market price for the labor services of any given worker, and all firms must pay this price in order to receive her labor services. As a result, these models cannot address data on labor turnover.

Models of turnover propose mechanisms that explain how both the inside and outside options for particular workers evolve over their careers and therefore produce predictions about the relationships between wage dynamics and labor mobility that vary with worker skills, tenure, and labor market experience. While most of this literature has focused on worker mobility among firms, a more recent literature has explored links between wage dynamics and patterns of mobility among industries or careers.

Here I discuss four *JPE* papers that are intellectual cornerstones of the literature on individual wage dynamics and mobility decisions and also comment on how work published in the *JPE* and elsewhere built on these seminal papers.

Hiring Costs, Specific Skills, and Other Frictions

Becker (1962) and Oi (1962) are cornerstones of the literature that explores why the value of a worker to his current firm may diverge from the worker's best outside option, even in competitive labor markets. Oi discusses several fixed costs associated with hiring workers. New workers need training in order to understand how a given firm organizes work and communication. Further, they also need to learn how to use capital that is specific to their new firm. In addition, firms cannot hire workers or assign workers to tasks without screening them in some way. Screening activities may include reviewing applications, conducting interviews, requiring workers to audition during probationary employment periods, and so forth.

Oi (1962) argues that, in competitive labor markets, workers must pay for these training and screening costs by accepting wages less than their marginal products. One can imagine many different payment schedules, but among workers with some seniority in their current firms, their compensation may well exceed their best option elsewhere, since any new employer would require them to pay for new hiring costs associated with training and screening.

Oi (1962) provides some evidence that these hiring costs are greater among more skilled workers, and he argues that this is the reason that turnover rates among skilled workers are less cyclical. Assume for a mo-

ment that, in economic downturns, some but not all firms suffer transitory demand shocks that decrease the marginal product of their workers. Oi claims that those firms hit by negative shocks can ask their most skilled employees to accept temporary wage cuts without losing them since large hiring costs create an important wedge between their inside and outside options.

Becker (1962) stressed the idea that firms and workers may share the costs and returns associated with firm-specific training. During training periods, workers earn more than their marginal product but less than their outside option. After training, workers earn more than their outside option but less than their full value to the firm. Such sharing rules may promote efficient separation decisions in settings in which shocks arrive that change either firm productivity or the outside options of workers.

This framework also suggests that skilled workers, who benefit more from training and therefore receive more training, should exhibit lower turnover rates. Further, individual wages should increase with worker tenure, holding constant wage growth attributable to the accumulation of general human capital. This insight spawned a large empirical literature that sought to measure the impact of changes in worker seniority on wages holding constant years of education and total labor market experience. I return to this topic below.

Acemoglu and Pischke (1999) note that the types of frictions that Becker (1962) and Oi (1962) identify may make it profitable for firms to pay for general training as well. The existence of employment frictions may make it possible for firms and their workers to share in the costs and returns of training that produces skills that are valuable in many firms.

Lazear (2009) presents a model in which no skills are truly firm specific, but each firm employs a different combination of skills and subsidizes investment in the skill mix it desires. This is a two-period model, and in the second period, a worker remains in her initial firm if her skills are more valuable to this firm than to the outside firm that makes the best raiding offer. In this framework, Nash bargaining determines compensation, and separation decisions are efficient, but market thickness influences skill investment in the first period. Workers are more willing to invest in skills if their first-period firm employs a mix of skills that is highly valued by a large number of other potential employers.

These papers sharpen our thinking about how skill specificity affects both wages and turnover. They offer reasons why skilled workers should be less mobile than other workers, and they provide reasons why wages may not equal outside options, even in competitive labor markets. Still, none of these papers provide a complete characterization of both wage growth and job mobility over the entire career of a worker. Almost two decades after Becker and Oi produced their seminal work, another

economist trained at Chicago produced two papers that sparked a significant literature on life cycle wage growth and patterns of labor mobility.

Matching, Wage Growth, and Turnover

In 1979, Boyan Jovonavic published two chapters from his PhD thesis in the *JPE*. Both are seminal contributions to modern work on individual mobility decisions and how these decisions affect individual wage growth over the life cycle. In the late 1960s and throughout the 1970s, researchers began analyzing data from the first modern panel surveys of labor market experiences, and at least three patterns were apparent in these early panel surveys: (1) wages are higher, on average, among workers with greater tenure, holding total labor market experience constant; (2) separation rates decline with wage levels given various sets of controls for worker characteristics; and (3) separation rates decline with tenure. Both papers (Jovanovic 1979a, 1979b) address these empirical regularities.

Jovanovic (1979a) presents a continuous-time model, but to facilitate exposition, I discuss the discrete-time analogue. When a worker joins a firm, he draws a match from a distribution, and both worker and firm observe a signal that provides information about the quality of the match. If the worker remains with the firm, both worker and firm observe an additional signal concerning the match quality each period. The competitive equilibrium in this model is an implicit contract equilibrium that requires all firms to pay workers their expected marginal product based on the information available about the quality of their current match. Each period, the worker must decide whether to stay with his current firm or pay a search cost and start over with a new firm.

Workers live forever in this model, and the distribution of potential matches is the same in all firms. These features imply that workers follow a reservation wage rule and that reservation wages rise with seniority.

Once a worker has drawn a match, his current job has value for two reasons. First, he earns a wage equal to the expected value of the match, and second, his match has option value. If the match turns out to be great, he can keep it. If it turns out to be terrible, he can leave and try a new firm. In Jovanovic (1979a), this option value component declines monotonically with tenure because the posterior variance of the conditional distribution of match quality shrinks as each new signal arrives. For this reason, reservation wages rise with seniority.

Average wages are positively correlated with seniority in this model precisely because reservation wages rise with seniority. Any worker with τ periods of tenure has received signals that produced a sequence of posterior means that exceed more cutoffs and a more stringent set of cutoffs

than another worker with only $\tau - k$ periods of tenure. Thus, on average, wages must be higher for workers with more tenure.

Further, the probability that the next signal leads to a revision that ends the match decreases with the expected value of the current match. So, individual separation rates are negatively related to wage rates.

However, in Jovanovic (1979a), separation rates may rise with seniority in the early periods of a match. Early on, separation rates may be higher at tenure $t + 1$ than at t because the reservation wage is higher at $t + 1$. Nonetheless, separation rates must eventually decline with seniority because of the shrinking variance of the conditional distribution of match quality. At some point, a given worker becomes so confident about the actual quality of her match that there is almost no chance that future signals could produce wage decreases that would justify a separation.

In 1979, Jovanovic may have been disappointed that this model did not produce separation rates that declined monotonically with tenure. However, he had to be pleased when Farber (1994) reported results from the first decade of work history information in the National Longitudinal Survey of Youth—1979 (NLSY79). In contrast to earlier panel studies, the NLSY79 contained weekly rather than yearly information on separations, and Farber discovered that individual separation rates actually rose during the first 6 months of job tenure before falling monotonically, a pattern that one can easily produce using the model in Jovanovic (1979a). Since a significant fraction of separations occur in the first 6 months of job spells, Jovanovic managed to predict an important feature of hazard rates out of employment spells more than a decade before researchers documented its existence.

Extensions of Jovanovic's Matching Model

The optimal reservation wage policy in Jovanovic (1979a) is relatively simple in this model for two reasons. First, Jovanovic assumes that each worker has a common prior about the distribution of potential matches that she applies to all potential employers. Second, he assumes that workers learn about only one match at a time. The signals that one match produces provide no information about the potential matches that a worker might have with any other employer. Since workers live forever and the number of potential jobs is infinite, workers never return to any job they leave. This means that the information that a worker acquires on any given job that eventually ends never affects his next job choice.

Over the past four decades, economists have tried to generalize this model. The literature on multiarmed-bandit problems with correlated arms contains no analytical solutions to the most general formulations of this matching model. If workers begin their careers with different pri-

ors over each possible job and learn about the distribution of matches in many potential jobs from observing signals in their current match, the problem becomes intractable because optimal decision rules depend on too many state variables and laws of motion. So, subsequent work has explored different ways to relax the key assumptions in Jovanovic's framework while maintaining tractability.

Miller (1984) allows workers to have different prior beliefs about the potential matches associated with different jobs, but he does not allow the signals from one match to shape beliefs about potential matches in other jobs. Miller argues that his model explains why young workers are more likely to enter jobs in which success is rare but particularly lucrative.

In Neal (1999), I assume that jobs involve a firm component and a career component. I further assume that workers cannot shop effectively over different careers while working for one firm. This model predicts that workers should search in two stages. They should focus on finding a career match first and then search for a firm match. Job histories in the NLSY79 contain several patterns that are consistent with this two-stage search strategy (see Neal 1999; Gathmann and Schoenberg 2010; Pavan 2010).

Gibbons and Waldman (1999) adopt a different approach. They assume that all workers know the mappings between talent and expected output in every potential job, and they adopt a learning technology such that the rate of learning about worker talent is the same in all jobs. Here, the optimal policy is simple. Each period, each worker selects the job in which she expects to be most productive, given what past signals have revealed about her talent. This model stands in sharp contrast to that in Miller (1984) because optimal mobility patterns do not reflect differences in the value of information produced by different types of work experience. Nonetheless, this model has enjoyed noteworthy success empirically (see Gibbons and Waldman 1999; Gibbons et al. 2005).

The Role of Experience

Jovanovic (1979a) has been quite influential, but there are features of data on life cycle wage growth and mobility that it cannot address. Because workers live forever and learn nothing about themselves in one job that allows them to search more efficiently for future jobs, wages and separation rates are not functions of labor market experience among workers who share the same level of job seniority. One could add general learning by doing as a quick fix for the implied wage profiles, but each worker would still make mobility decisions facing an infinite horizon and any information a worker acquired while working at any one job would not affect her expected match quality in future jobs. So, total market experience would not affect mobility decisions holding seniority constant.

However, in Jovanovic (1979b), workers have finite work lives, and among workers with the same seniority, total market experience affects both wages and separation decisions. Here, workers can allocate time to three activities: work, investment, and search. Job offers arrive randomly, and workers decide whether to keep their current match or accept the new offer. Investment activities build the value of the current match, and search activities increase the arrival rate of new offers. The key insight of this model is that good matches last, in part, because of behavioral mechanisms that are not present in Jovanovic (1979a). Agents rationally invest more in matches that are harder to beat and devote more time to search when their current match is not that good.

In this model, wages rise with seniority relative to outside options, separation rates decline with current wages, and separation rates decline monotonically with tenure. Further, holding tenure constant, separation rates decline with total labor market experience since workers with more experience have had more time to receive high offers, and all else equal, workers devote less effort to search as they near the end of their work lives because they face a shorter horizon over which to enjoy the returns from successful search.

Sources of Wage Growth

Both Jovanovic papers highlight a reverse causality concern for the empirical literature on wage determination. Wages are positively correlated with seniority in both models. However, in Jovanovic (1979a), the value of matches does not grow with time, and this positive correlation arises solely because good matches last. And, even in Jovanovic (1979b), a component of the relationship between tenure and wage levels is driven by the fact that good initial matches last longer, in part, because they induce low search intensity. The selection effects highlighted by these models have frustrated many efforts to pin down the causal effect of seniority on wages. However, the ideas in these two papers have shaped decades-long debates concerning the best ways to measure the contribution of growth in firm-specific, industry-specific, or career-specific skills to life cycle wage growth.

This literature is too large to review here (see Altonji and Shakotko 1987; Topel 1991; Kambourov and Manovskii 2009). Yet Pavan (2011) deserves our attention. Pavan estimates a structural model of search and wage growth. In the model, workers search for both firm and career matches. Further, wages grow over the life cycle for four reasons: workers accumulate general human capital through schooling and experience, matches with careers evolve over time, matches with firms evolve over time, and workers change both firms and careers optimally. In Jovanovic (1979b),

good matches last, in part, because workers invest in them. In Pavan (2011), matches last if they grow fast enough.[2]

Pavan's estimates imply that both firm-specific and career-specific components of wages grow as workers gain more seniority in a firm and more experience in a career.[3] In the final section of the paper, Pavan simulates data from his model and then estimates an instrumental variable (IV) regression quite similar to one found in Kambourov and Manovskii (2009). The results of this regression imply that the firm-specific component of wages does not grow with seniority, which parallels the results reported in Kambourov and Manovskii's paper. This finding casts serious doubt on the value of such IV methods, since Pavan ran his IV regressions on data simulated from a model in which the firm-specific component of wages does grow substantially with tenure.

Pavan (2011) shows that researchers who seek to pin down the sources of life cycle wage growth must employ empirical models that explicitly treat wage growth and mobility decisions as joint outcomes. Jovanovic (1979a, 1979b) pointed this literature in the right direction, and Pavan (2011) demonstrates the value of this approach. However, more work remains. Pavan needed several restrictive assumptions on the processes that determine wage growth in order to make his model computationally tractable.

Conclusion

Models of life cycle investment in general human capital are powerful tools for explaining how the average earnings of particular types of workers evolve, rather smoothly, over the life cycle. However, individual wage and earnings histories are quite jagged. They often contain large increases or decreases, and these changes are usually coincident with promotions, changes of employer, or career changes.

Thus, researchers cannot explain individual wage histories without formulating models that treat both individual wage growth and individual mobility decisions as endogenous life cycle outcomes. Even in a setting in which demands for various types of workers are stationary, a worker must discover what type she is and what forms of employment allow her type to be most productive. In addition, a worker must constantly weigh the returns from investing in her current match against the possible gains from searching for better.

These search and investment activities often create ex post rents, that is, wedges between current productivity and outside options, but the im-

[2] Keane and Wolpin (1997) structurally estimate a dynamic model that allows young men to choose dynamically working in either the blue-collar or white-collar sector. This model includes sector-specific skill growth and shocks to individual skill stocks, but firms play no role. All firms pay the same prices for both skills.

[3] Pavan (2010) defines careers using industry and occupation codes. Neal (1995) and Poletaev and Robinson (2008) produce similar results using data on wage changes among displaced workers.

plicit contract equilibrium in Jovanovic (1979a) suggests that the existence of these rents does not necessarily imply that any jobs are rationed, that workers do not invest efficiently in skills, or that mobility decisions are inefficient.[4]

Further, if workers sort among jobs on the basis of information about their skills that econometricians cannot measure but employers observe, researchers may measure firm, industry, or career differences in residual wage levels and turnover rates that are not evidence of ex ante labor market rents but rather evidence that workers sort among firms and careers on valuable traits that are in relatively fixed supply (see Neal 1998).

The next century of *JPE* papers will likely contain important efforts to position explicit life cycle models of wage growth within models of labor market equilibrium that contain shocks to firms or sectors. Here, mobility and wage growth may reflect shocks to what a worker knows about herself or shocks to demand and technology that change the productivity of her type. This hybrid approach may help economists better understand the causes and consequences of the ex post rents that experienced workers enjoy in their current firms and careers.

References

Acemoglu, Daron, and Jörn-Steffen Pischke. 1999. "The Structure of Wages and Investment in General Training." *J.P.E.* 107 (3): 539–72.

Altonji, Joseph G., and Robert A. Shakotko. 1987. "Do Wages Rise with Job Seniority?" *Rev. Econ. Studies* 54 (3): 437–59.

Becker, Gary S. 1962. "Investment in Human Capital: A Theoretical Analysis." *J.P.E.* 70, no. 5, pt. 2 (October): 9–49.

Ben-Porath, Yoram. 1967. "The Production of Human Capital and the Life Cycle of Earnings." *J.P.E.* 75 (4): 352–65.

Farber, Henry S. 1994. "The Analysis of Interfirm Worker Mobility." *J. Labor Econ.* 12 (4): 554–93.

Gathmann, Christina, and Uta Schoenberg. 2010. "How General Is Human Capital? A Task-Based Approach." *J. Labor Econ.* 28 (1): 1–49.

Gibbons, Robert, Lawrence F. Katz, Thomas Lemieux, and Daniel Parent. 2005. "Comparative Advantage, Learning, and Sectoral Wage Determination." *J. Labor Econ.* 23 (4): 681–724.

Gibbons, Robert, and Michael Waldman. 1999. "A Theory of Wage and Promotion Dynamics Inside Firms." *Q.J.E.* 114 (4): 1321–58.

Heckman, James. 1976. "A Life-Cycle Model of Earnings, Learning, and Consumption." *J.P.E.* 84, no. 4, pt. 2 (August): S11–S44.

Jovanovic, Boyan. 1979a. "Job Matching and the Theory of Turnover." *J.P.E.* 87 (5): 972–90.

———. 1979b. "Firm-Specific Capital and Turnover." *J.P.E.* 87 (6): 1246–60.

Kambourov, Gueorgui, and Iourii Manovskii. 2009. "Occupational Specificity of Human Capital." *Internat. Econ. Rev.* 50 (1): 63–115.

Keane, Michael, and Kenneth Wolpin. 1997. "The Career Decisions of Young Men." *J.P.E.* 105 (4): 473–522.

[4] However, these ex post rents may create inefficiency. See Sanders and Taber (2012).

Lazear, Edward P. 2009. "Firm-Specific Human Capital: A Skill-Weights Approach." *J.P.E.* 117 (5): 914–40.
Miller, Robert A. 1984. "Job Matching and Occupational Choice." *J.P.E.* 92 (6): 1086–1120.
Mincer, Jacob. 1958. "Investment in Human Capital and Personal Income Distribution." *J.P.E.* 66 (4): 281–302.
Neal, Derek. 1995. "Industry-Specific Human Capital: Evidence from Displaced Workers." *J. Labor Econ.* 13 (4): 653–77.
———. 1998. "The Link between Ability and Specialization: An Explanation for Observed Correlations between Wages and Mobility Rates." *J. Human Resources* 33 (1): 173–200.
———. 1999. "The Complexity of Job Mobility among Young Men." *J. Labor Econ.* 17 (2): 237–61.
Oi, Walter Y. 1962. "Labor as a Quasi-Fixed Factor." *J.P.E.* 70 (6): 538–55.
Pavan, Ronni. 2010. "The Role of Career Choice in Understanding Job Mobility." *Labour* 24 (2): 107–27.
———. 2011. "Career Choice and Wage Growth." *J. Labor Econ.* 29 (3): 549–87.
Poletaev, Maxim, and Chris Robinson. 2008. "Human Capital Specificity: Evidence from the Dictionary of Occupational Titles and Displaced Worker Surveys, 1984–2000." *J. Labor Econ.* 26 (3): 387–420.
Rosen, Sherwin. 1976. "A Theory of Life Earnings." *J.P.E.* 84, no. 4, pt. 2 (August): S45–S67.
Sanders, Carl, and Christopher Taber. 2012. "Life-Cycle Wage Growth and Heterogeneous Human Capital." *Ann. Rev. Econ.* 4:399–425.
Topel, Robert. 1991. "Specific Capital, Mobility, and Wages: Wages Rise with Job Seniority." *J.P.E.* 99 (1): 145–76.

The Human Capital Approach to Intergenerational Mobility

Magne Mogstad

University of Chicago

Introduction

In two closely related papers, Becker and Tomes (1979, 1986) develop a model of the transmission of earnings, assets, and consumption from parents to children. The model is based on utility maximization of parents concerned about the income or welfare of their children, in contrast to

Thanks to Jorge Luis García, Jim Heckman, John Eric Humphries, Jack Mountjoy, and Alessandra Voena for helpful comments and suggestions.

contemporaneous empirical and statistical work that was not explicitly based on a model of maximizing behavior.[1]

Three decades later, the Becker-Tomes model remains the main building block of economic research on intergenerational mobility. Not only did the model help clarify key economic mechanisms that may be producing or preventing intergenerational mobility, but it also started a process in which new data, empirical evidence, and theoretical models were brought forward and scrutinized.[2] By now, this process has generated a large and growing literature on intergenerational mobility.

Different strands of the literature have different goals. Some analyses advance knowledge by addressing measurement challenges or uncovering new facts. Some seek to identify causal impacts of specific interventions or policy changes. Other analyses try to understand the mechanisms producing or preventing intergenerational mobility. Some studies even do all three. To give a full picture of the literature on intergenerational mobility is a daunting task and is not within the scope of this short article.[3] Instead, I outline a stripped-down Becker-Tomes model, discuss a few of its insights, and review some exciting recent advancements.

The Becker-Tomes Model

Becker and Tomes (1979, 1986) present multigenerational models with one period of childhood, one period of adulthood, one child per family (no fertility decisions), and a single parent. Parents begin with income Y_t, a combination of earnings and financial transfers they received from their parents. Parents spend on three items: their own consumption C_t, investments in the human capital of their child I_{t+1}, and financial transfers to their child X_{t+1}. Parents exogenously transmit ability or endowment A_{t+1} to their children through a stochastic linear autoregressive process. After observing the child's ability, the human capital of the child is determined by parental investment in human capital. In adulthood, labor is supplied inelastically.

Parents care about their own consumption and the income available for consumption and investment for their children. The optimization problem of the parent is

$$\max_{C_t, I_{t+1}} U(C_t, Y_{t+1}) \tag{1}$$

[1] Goldberger (1989) famously critiqued the Becker-Tomes model for having little added value relative to statistical, "nonoptimizing" models of intergenerational transmission. Becker (1989) replies to the criticism, while Mulligan (1999) and Solon (2004) clarify and discuss the predictions of the Becker-Tomes model.

[2] An early example is Behrman, Pollak, and Taubman (1982). While contemporaneous work on intergenerational mobility appealed to stylized facts in an attempt to rationalize or test the theories, Behrman et al. estimate a general preference model for analyzing parental allocations of resources among their children.

[3] For a recent review of the literature, see Heckman and Mosso (2014).

subject to

$$\begin{aligned} Y_t &= C_t + X_{t+1} + I_{t+1}, \\ Y_{t+1} &= w_{t+1} f(I_{t+1}, A_{t+1}) + (1 + r_{t+1}) X_{t+1} + U_{t+1}, \end{aligned} \tag{2}$$

and, possibly, the borrowing constraint

$$X_{t+1} \geq 0, \tag{3}$$

where w_{t+1} is the return to human capital in period $t + 1$, $f(\cdot)$ is the human capital production function, r is the return on financial assets, and U_{t+1} is the idiosyncratic component of children's income (market luck).

There are two distinct versions of the Becker-Tomes model. In both versions, the parent can affect the consumption allocation of the family by investing in children's human capital and by leaving bequests. A key difference between the models is the possibility that credit constraints may influence parental investment decisions and thereby alter the nature of intergenerational mobility as compared to a situation without credit constraints.

In the 1979 version of the model, the constraint (3) is not imposed, and intergenerational mobility is driven by persistence in ability and the variance of labor market shocks. Because this version of the model places no restrictions on the ability of parents to borrow against the earnings potential of their children, there is no role for parental income (or the magnitude of parental altruism) in determining the optimal level of investment. No matter their income, parents can borrow freely in the market to finance the optimal investment level. All parents will therefore choose to invest the privately efficient amount in children's human capital so that the marginal return is driven down to the opportunity cost of investments, which is the forgone interest on financial investments. As a consequence, equally able children will receive equal investments independent of their parent's income or human capital, and parental influence on intergenerational mobility is limited to the heritability of ability. This does not, however, mean that all children receive the same level of investments. In the Becker-Tomes model, the specification of $f(\cdot)$ assumes that the marginal return from investment in a child's human capital is positively related to the endowment he inherits. This assumption implies that children with greater endowments will receive larger investments, contributing to earnings inequality across families within a generation.

In the 1986 version of the model, the constraint (3) is invoked, restricting parents from borrowing against the earnings potential of their children. This constraint captures the idea that children cannot credibly commit to repay the loans parents would take on their behalf. If a parent is credit constrained, the child acquires less human capital, and so the re-

turn to such investments is higher than that on financial assets. In this setting, earnings persist across generations both because ability persists (as in the 1979 version) and because credit constraints limit human capital investments. Building on the insights of Becker and Tomes, researchers have tried to test for and quantify the distortions caused by credit constraints. As discussed by Heckman and Mosso (2014), the early literature did not make a compelling case for the importance of credit constraints for investments in human capital and intergenerational mobility. However, there is a now a growing body of evidence—based on more recent data—supporting the empirical importance of credit constraints in affecting educational attainment (see, e.g., Lochner and Monge-Naranjo 2012, 2015; Hai and Heckman 2017). In particular, the constrained seem to fall into two groups: those who are permanently poor over their lifetimes and a group of well-endowed individuals with rising high levels of acquired skills who are constrained early in their life cycles.

Beyond the Becker-Tomes Model

Since Becker and Tomes (1979, 1986), analyses of intergenerational mobility have made important progress by using multiple models and sources of data in a back and forth in which both models and measurements are augmented as learning evolves. Recently, particular attention has been devoted to three assumptions of the Becker-Tomes model: (i) investments at any stage of childhood are equally effective, (ii) earnings depend on a single skill in the form of human capital, and (iii) parental engagement with the child is in the form of investment in educational goods, analogous to firm investments in capital equipment. An active body of research suggests that these assumptions are at odds with the data and that the Becker-Tomes model misses key implications of richer models of intergenerational mobility. Work by Cunha and Heckman (2007), Heckman and Mosso (2014), and Lee and Seshadri (forthcoming) highlights three ingredients of particular significance for measuring and understanding intergenerational mobility: multiple periods of childhood and adulthood, multiplicity of skills, and several forms of investments.

Multiple Periods

Given the assumption that investments at any stage of childhood are equally effective, Becker and Tomes can model parental choices of investment in children through a simple lifetime budget. Therefore, what matters for parental investments is the lifetime or permanent income, and not the timing of receipt (or uncertainty) of income over the life cycle. In models with multiple periods of childhood and adulthood, how-

ever, the timing of income can be important as it interacts with restrictions on credit markets and the technology of skill formation. Parents may not be only restricted from borrowing against the earnings potential of their children (i.e., an intergenerational credit constraint, as in [3]), but also prevented from borrowing fully against their own future earnings (i.e., an intragenerational credit constraint). The intragenerational credit constraint induces a suboptimal level of investment (and consumption) in each period in which it binds. How harmful this constraint will be depends on the technology of skill formation and the life cycle profile of parental earnings. Cunha, Heckman, and Schennach (2010) estimate the elasticity of substitution parameters for inputs at different periods that govern the trade-off of investment between the early years and the later years. They present evidence of dynamic complementarities in the production of human capital, implying that early investment in children's skill development will have large returns because they raise the payoffs to future investments. As a consequence, if the intragenerational credit constraint is binding during the early periods, late investments will be lower, even if the parent is not constrained in later periods.

Multiple Skills

An active body of research measures various types of traits or skills and examines how well they predict or explain various socioeconomic outcomes (see Borghans et al. [2008] and the references therein). Recently, important progress has been made in accounting for measurement error and in trying to establish causal rather than merely predictive effects. The evidence points to the importance of including sufficiently broad and nuanced measures of skills in studies of intergenerational mobility. Both cognitive and noncognitive skills predict adult outcomes, but they have different relative importance in explaining different outcomes. For example, schooling seems to depend more strongly on cognitive skills, whereas earnings are equally predicted by noncognitive abilities. Importantly for models of intergenerational mobility, both cognitive and noncognitive skills can be affected by parental investment (Cunha and Heckman 2007). However, sensitive periods in which investments have particularly high returns appear to come earlier for cognitive as compared to noncognitive skills.

Many Forms of Investments

The Becker-Tomes model (and many of the extensions of the model) considers only a single child investment good. Recent evidence, however, points to the importance of allowing for multiple forms of investments and letting the returns to these investments vary over the life cycle of the

child. For example, Del Boca, Flinn, and Wiswall (2014) develop and estimate a model of intergenerational mobility in which parents make a number of specific input choices, ranging from various time inputs to child good expenditures, each with a child age-specific productivity. Their empirical results indicate that both parents' time inputs are important for the cognitive development of their children, particularly when the child is young. In contrast, the productivity of monetary investments in children has limited impacts on child quality no matter what the stage of development.

Concluding Remarks

The human capital approach considers how the productivity of people in market and nonmarket situations is changed by investments in education, skills, and knowledge. The approach was pioneered by scholars associated with the University of Chicago in the late 1950s and early 1960s, and many of the seminal contributions were published in the *JPE*. In 1962, for example, the *JPE* published a special issue on human capital with several landmark papers. Nearly two decades later, Becker and Tomes (1979, 1986) developed the human capital approach into a general theory for income inequality, both across families within a generation and between different generations of the same family. Much of the progress since, however, has focused on improving measurements, uncovering new facts, or identifying causal impacts of various determinants of mobility. With some notable exceptions, the *JPE* and the University of Chicago played a smaller role in this endeavor, possibly because of preferences for empirical work with tighter links to theory. Recently, however, important progress has been made by combining theory and empirics. In particular, Heckman and coauthors have adjusted the theories of intergenerational mobility in light of new evidence and then taken those theories to new data sets, getting us closer to fulfilling the goal of Becker and Tomes (1986, 3) of an "analysis that is adequate to cope with the many aspects of the rise and fall of families."

References

Becker, Gary S. 1989. "On the Economics of the Family: Reply to a Skeptic." *A.E.R.* 79 (June): 514–18.

Becker, Gary S., and Nigel Tomes. 1979. "An Equilibrium Theory of the Distribution of Income and Intergenerational Mobility." *J.P.E.* 87 (December): 1153–89.

———. 1986. "Human Capital and the Rise and Fall of Families." *J. Labor Econ.* 4 (July): 1–39.

Behrman, Jere R., Robert A. Pollak, and Paul Taubman. 1982. "Parental Preferences and Provision for Progeny." *J.P.E.* 90 (February): 52–73.

Borghans, Lex, Angela Lee Duckworth, James J. Heckman, and Bas ter Weel. 2008. "The Economics and Psychology of Personality Traits." *J. Human Resources* 43 (4): 972–1059.
Cunha, Flavio, and James J. Heckman. 2007. "The Technology of Skill Formation." *A.E.R.* 97 (2): 31–47.
Cunha, Flavio, James J. Heckman, and Susanne M. Schennach. 2010. "Estimating the Technology of Cognitive and Noncognitive Skill Formation." *Econometrica* 78 (3): 883–931.
Del Boca, Daniela, Christopher Flinn, and Matthew Wiswall. 2014. "Household Choices and Child Development." *Rev. Econ. Studies* 81 (January): 137–85.
Goldberger, Arthur S. 1989. "Economic and Mechanical Models of Intergenerational Transmission." *A.E.R.* 79 (June): 504–13.
Hai, Rong, and James J. Heckman. 2017. "Inequality in Human Capital and Endogenous Credit Constraints." *Rev. Econ. Dynamics* 25 (April): 4–36.
Heckman, James J., and Stefano Mosso. 2014. "The Economics of Human Development and Social Mobility." *Ann. Rev. Econ.* 6:689–733.
Lee, Sang Yoon, and Ananth Seshadri. Forthcoming. "On the Intergenerational Transmission of Economic Status." *J.P.E.*
Lochner, L. J., and A. Monge-Naranjo. 2012. "Credit Constraints in Education." *Ann. Rev. Econ.* 4:225–56.
———. 2015. "Student Loans and Repayment: Theory, Evidence and Policy." Working Paper no. 20849, NBER, Cambridge, MA.
Mulligan, Casey B. 1999. "Galton versus the Human Capital Approach to Inheritance." *J.P.E.* 107, no. 6, pt. 2 (December): S184–S224.
Solon, Gary. 2004. "A Model of Intergenerational Mobility Variation over Time and Place." In *Generational Income Mobility in North America and Europe*, edited by Miles Corak. Cambridge: Cambridge Univ. Press.

Health Economics: A Selective Historical Review for the 125th Anniversary of the *Journal of Political Economy*

Robert H. Topel

University of Chicago and National Bureau of Economic Research

Introduction

Only a small fraction of my own work can be categorized as health economics (HE), so the reader might consider my views on the field as those of a modestly informed outsider who participates on occasion. My taxon-

omy of HE contributions also reflects my background and "Chicago" tastes. Simplifying more than a bit, I group important HE papers appearing in the *JPE* into two main areas. The first area treats health as a form of human capital, with all that entails. Given the place of the University of Chicago in developing the human capital model, one might expect that a larger share of HE papers in the *JPE* would fall in this category than at other journals, and that is indeed the case.[1] One might also expect that I would find this category the most interesting and important, and that is also the case. The second main area is the economics of health insurance, including the design of public health insurance schemes and the effects of insurance on incentives, the demand for medical care, and the prices and allocation of health care services. I confine most of my discussion to these two areas.[2]

Because this is the 125th anniversary issue of the *JPE*, I shall engage in a bit of intellectual history, following the trail of HE back to the early 1900s, and spend some ink on articles unlikely to appear on modern graduate reading lists. Yet the data indicate that health economics is a fairly young field. A rough but fairly reliable method of dating its emergence is to find the first mentions of the word "health" in the title of research papers published in leading general-interest journals. The *JPE* began publication in 1892, but a paper with "health" in its title did not appear for 70 years, with the publication of Selma Mushkin's "Health as an Investment" (1962), a landmark (in my view, at least) contribution about which I will say more below. The same method of carbon dating yields similar results in other leading journals: the first appearances of "health" in the *Quarterly Journal of Economics* and the *American Economic Review* occurred in 1952 and 1954, respectively. A broader search for health-related terms (such as longevity, mortality, morbidity, disease, sickness, medical, medicine, and life span) does not alter my general conclusion that HE was not an especially active area of scholarship before the 1950s, though, as discussed below, important work on the desirability and design of universal health insurance—a very active topic in HE since 1970—appeared in the *JPE* as early as 1904.

Of necessity this is a very selective review, with a historical perspective. I emphasize original contributions in areas that I find interesting and important, which means that later and perhaps even better contributions are neglected. I apologize if yours is one of those.

[1] An unscientific review of PhD reading lists in health economics indicates that other general-interest journals also publish more HE papers, broadly defined. This suggests that the *JPE*'s human capital "slant" is greater than simple shares would suggest.

[2] Back in graduate school one of my professors explained that there are only two fields in economics: labor and industrial organization. My taxonomy might be an application of his more general principle.

Health as Human Capital

Investing in Health Capital

Mushkin's "Health as an Investment" (1962) appears in a special issue of the *JPE* titled "Investment in Human Beings," edited by T. W. Schultz and including now-classic papers by Becker, Mincer, Stigler, and Sjaastad. Becker (2007) describes it as "not particularly insightful," and he even attributes the slow growth of the health-as-human-capital field to this paper. I disagree with Gary. Mushkin's paper is a *tour de force* that places human health within the then-nascent theory of human capital.[3] Unlike later, more specialized, contributions in this area, Mushkin's paper is practically a survey of what health economists studying the production, maintenance, and value of health will work on in later decades, and most of her admittedly preliminary and nontechnical analysis is spot-on and (I think) remarkably insightful. Among other topics, she analyzes (1) complementarities between health and other forms of human capital, particularly education, at both the private and social levels;[4] (2) the difficulties of measuring health-related changes in the quality of life; (3) the contributions of health capital and increasing longevity to economic growth; (4) the public-good nature of health research and treatment, which motivates government support of both medical research and the supply of health care; (5) inherent difficulties in identifying the social returns to health research as an input to policy evaluation; and (6) methods of assigning a monetary value to health improvements.

Mushkin's discussion of methods for valuing health improvements is a bit primitive and even contradictory. Like many others before and after, she attempts to value improved health in terms of the increased labor supply and production or earnings that come with living longer or healthier, whereas the economically correct approach would measure the value in terms of individuals' willingness to pay for health improvements—which embeds any value derived from productivity and the ability to work. For example, her approach assigns zero value to reductions in mortality or morbidity that affect only the retired population, yet these individuals would nevertheless enjoy the benefits of living longer and better. That is valuable. In fact, this is the margin where health gains in advanced economies have been concentrated since about 1960: older individuals are living longer and better (Murphy and Topel 2006). Mushkin appears to

[3] She cites Schultz (1961), but *JPE* papers by Schultz (1960) and Mincer (1958) are equally relevant.

[4] "A lengthening of life expectancy through improved health reduces the rate of depreciation of investment in education and increases the return to it. . . . Improved education, on the other hand, increases the return on a lifesaving investment in health. . . . Educational levels determine to a large extent the seeking out of health services and the selection of appropriate kinds of services" (131).

have recognized the issue, however, and in another part of her paper she offers this prescient econometric strategy for valuing debilitating health risks and lost lives: "If workmen's compensation charges . . . can be used to measure the risk of death and disability [by occupation or industry], it may be possible to use [wage] premiums paid for extra occupational risks—or 'hazard pay'—as an index of the market evaluation of the risk of continuing debility. . . . Injury rates become one of the many factors considered in wage negotiation even when separate hazard premiums are not paid" (135). This is a concise description of methods that would be applied many years later, and are still widely used today, to estimate the "value of a statistical life," which became a key tool for government cost-benefit analyses of policies or programs that save lives.[5]

Human capital research accelerated in the 1960s, but beyond noting that health could be viewed as a form of human capital (e.g., Becker 1964), a full decade passed before an important paper followed Mushkin's early lead. Grossman (1972) is the first formal treatment of health as a form of capital and an object of an individual's life cycle choices, so that both health and longevity are endogenous. It is the most-cited health economics paper to have appeared in the *JPE*, so it is worth a somewhat closer look.

There is no uncertainty in Grossman's setup: he assumes perfect foresight. The stock of human capital at each age, H_a, produces an output of "healthy time" that enters life cycle utility. The stock follows a standard law of motion for a durable good, $H_{a+1} - H_a = I_a - \delta_a H_a$, where I_a and δ_a are gross investment in health and depreciation, respectively. Longevity is endogenous because death occurs when the stock of human capital falls below some exogenous level, $\bar{H}$. Gross investment is produced with purchased medical care and the stocks of health and other human capital, so the demand for medical care is not a form of consumption but is derived from the more primitive demand for health itself. With assumed constant returns in goods and time inputs, and thus constant marginal cost of investment, death can occur only if depreciation rises with age, which Grossman assumes. This yields additional predictions for the life cycle patterns of purchased medical care and gross investment, as people attempt to offset rising depreciation. Gross investment and the demand for medical care also rise with age if the elasticity of the marginal efficiency of capital is smaller than unity. Production of health capital and consumption of health rise with the wage because health time is more valuable, making health a normal good. Grossman also builds in complementarity between education, or other forms of human capital, and the production of health, so more educated individuals are also endogenously healthier.

[5] The literature is reviewed by Viscusi and Aldy (2003).

Grossman's (1972) paper stimulated a substantial literature treating health as a form of human capital, much of which appeared in the *JPE*.[6] But until recently research in this area remained a tiny fraction of the much larger human capital literature, which has mainly focused on education and training. Becker (2007) is a sweeping attempt to rectify this and also a useful summary of developments in the area.

Valuing Medical Advances and the Measurement of Health Capital

In 1900, 18 percent of newborn males in the United States died before their first birthday. By the end of the century it took until age 63 to reach the same level of cohort mortality. Early in the twentieth century reduced mortality rates were concentrated among the young, as infant mortality plummeted and a variety of infectious diseases affecting children were controlled or eradicated. Later gains were mainly concentrated at older ages as death rates from age-related diseases—especially cardiovascular diseases—declined after 1970.[7] Overall morbidity also declined, so that not only were people living longer, they were living better.

What drove these health improvements, and what are they worth? What would be the social value of further gains, and what might those gains cost? These questions are important for several reasons. First, improved health is an obvious contributor to human welfare, yet health improvements are not measured in traditional growth and welfare metrics such as per capita income. The uncounted gains can be large, so traditional measures can be very misleading. Second, basic medical research is a public good: a new idea or treatment can be extended across the current population and to future generations as well. This is one of the most powerful arguments for public support of medical research. In the United States, federal government support for medical research currently totals about $35 billion annually, almost all of which flows through the National Institutes of Health. As a matter of public policy, are the medical advances and improved health flowing from this research worth their cost? The answer requires that we connect medical research to treatments and health outcomes and that we properly value any changes in mortality and morbidity.

Weisbrod (1971) is the first attempt of which I am aware to measure the costs and benefits of medical research and to perform a cost-benefit

[6] See, e.g., Ben-Porath (1976), Cropper (1977), Rosenzweig and Schultz (1983), Wolfe (1985), Ehrlich and Chuma (1990), Ehrlich and Lui (1991), Kenkel (1991), Sah (1991), Philipson and Becker (1998), and Acemoglu and Johnson (2007).

[7] See Murphy and Topel (2006) for details.

analysis. Like many papers that followed, this is a case study of a particular disease (polio) for which (1) reasonable historical data on the flow of research support are available; (2) a preventative treatment (vaccination) was successfully discovered and widely applied, effectively eradicating the disease in the United States; and (3) once discovered and perfected, the marginal cost of treatment is fairly low.[8] Thus the connection between research, treatment, and outcomes is fairly clear. Weisbrod measures benefits in terms of increased productivity and reduced costs of care that would have been otherwise necessary. He recognizes that these measures are inadequate but at least conservative. He also recognizes that the benefit of a polio cure affected not just the current generation but all future ones as well, so future generations entered his calculations using various discount rates. Applying his model to the United States alone (other populations also gained because of the public-good character of new knowledge), he estimates an internal rate of return on polio research of 11 percent. It turned out to be a good investment, at least ex post.

Case studies such as this are important, and Weisbrod makes a convincing argument that his estimates are conservative, conditional on the successful research outcome that occurred and his assumption that research support caused it.[9] But even a retrospective assessment of the overall value of medical research requires that we integrate over successes and failures and that we have a more convincing method of valuing benefits and costs than Weisbrod was able to apply, including the widely varying costs of treating heterogeneous diseases. My paper with Kevin Murphy (Murphy and Topel 2006) makes some initial progress on these questions, offering evidence that the past and potential future gains from medical research may be quite large.

An important contribution of our paper is to recognize that the private economic value of health is determined by willingness to pay. With diminishing marginal utility of full consumption, which includes the value of leisure, life extension is valued because average utility exceeds marginal utility—consumption yields a flow of consumer surpluses over the life cycle—so the ability to spread consumption over more life years is valued (see also Usher 1973; Rosen 1988). This yields a formal expression for the value of a statistical life, which other research put at about $6 million at the time we wrote. We allocate that total to create a life cycle pat-

[8] Cutler and Kadiyala (2003) and Cutler, Rosen, and Vijan (2006) provide overviews of later literature. Chandra and Staiger (2007) study productivity in the treatment of heart attacks, explaining geographic differences in the use of intensive medical care.

[9] Even here, the attribution of costs and benefits to a particular research program is problematic. Weisbrod points out that Watson and Crick's work on DNA was partially supported by grants to study polio, so spillovers are important. And who knows what other grants polio researchers received.

tern of values for life years that declines at older ages. We then use these values to estimate the gains from past and potential future reductions in age-specific mortality rates from various diseases and overall.

The numbers are huge (see also Murphy and Topel 2003; Nordhaus 2003). Net of medical expenditures, the value of increased life expectancy between 1970 and 2000 in the United States totaled roughly $60 trillion, or a flow of about $2 trillion per year. Compare this to an average annual value of GDP of about $6 trillion over this period and the importance of improvements in health capital as a contributor to welfare is obvious. Becker (2007) extends these calculations to all OECD countries and estimates benefits that are triple the values for the United States (see also Becker, Philipson, and Soares 2005). Murphy and Topel (2006) also extend the analysis to measure the prospective gain (net of treatment costs and other possible distortions) to future generations from, say, a 10 percent reduction in cancer mortality. We find that a 10 percent reduction in cancer mortality would be worth about $500 billion to current and future US generations. Of course there are a number of caveats to these estimates, which we attempt to cover in some detail in the paper. Some are noted in the last section of this piece.

Health Insurance and Markets for Medical Care

On the basis of the flow of articles published, I claimed earlier that HE was not an active field of economic research until at least the 1950s. The extreme outlier is a series of papers on various types of social insurance by Isaac Rubinow that appeared in the *JPE* beginning in 1904. Rubinow (1875–1936) might lay claim to being the original health economist, and the *JPE* was his favorite outlet; he had 11 papers in the journal between 1904 and 1930.[10] A Russian immigrant, he earned his doctor of medicine degree from New York University and began his professional life treating poor immigrants in New York City. This experience alerted him to the complex interaction of poverty and health, and he concluded that he could have greater impact as an economist (it's true) than as a physician. So he earned his degree at Columbia in 1903 and began a long economics career in government and nonprofit organizations. His "Labor Insurance" (1904) was "the first publication by an American to call unequivocally for social insurance" (Kreader 1988, 56) and the first that I know of that explicitly argued for compulsory participation in government-subsidized health insurance; the intellectual debate over the "individual

[10] Health and social insurance were not his only topics. His *JPE* contributions also covered unemployment (1917) and even retail coupons and discounting (1905).

mandate" (as we call it today) was heated even then, and the issues were precisely the same. Foreshadowing the reasoning of the US Supreme Court (2012) concerning the individual mandate in the Affordable Care Act, Rubinow argued that compulsory participation is not forced consumption; it is simply a form of "tax" that the government is empowered to levy. He later greatly expanded these ideas in "Standards of Sickness Insurance" (1915). This paper is a nontechnical exercise in policy design, using the experiences of European health care programs to advocate for universal mandated coverage in the United States. Agree with him or not, Rubinow's work was way ahead of his time and his economic reasoning was not bad.

I am sure that Rubinow would have been very surprised at how things played out over the next 100 years.

Health insurance and the demand for health care did not materially reemerge in the literature until Arrow's influential 1963 paper in the *AER*. Publications from this strand did not appear in the *JPE* until the 1970s, and one is struck by the changes in style and methodology of economic research. The best of the lot is a classic paper by Feldstein (1973), who models and estimates the distortions and welfare implications of existing health insurance plans.[11] In modern parlance the paper might be described as "structural industrial organization," except that it is penetrable. The key idea is this. Virtually by definition, insurance plans include coinsurance rates such that the marginal price of medical care for an insured person is lower than the marginal cost of supplying that care. Then insurance increases the demand for both the quantity and quality of health care, which further increases the demand for insurance. With a rising supply price of medical services, this (stable) feedback process raises prices and allows suppliers to earn rents, while buyers of health insurance are overinsured. Feldstein formulates and estimates a structural model of health care demand and this dynamic process and then calibrates the welfare gains of raising the average coinsurance rate from 0.33 to 0.5 or higher. Because raising the coinsurance rate reduces the demand for medical care, equilibrium prices fall, which provides an offsetting benefit to consumers. The net welfare gain turns out to be positive and large: Feldstein estimates the gain from such reforms to be about $4 billion per year in early 1970s dollars. It is a very nice piece that foreshadows many others; the most recent example in the *JPE* is a fine piece by Clemens and Gottlieb (2017).

[11] Much of the health insurance literature seems to ebb and flow with public policy interests in creating national health insurance, perhaps driven by the supply of research funds. Several papers in the 1970s came out of the RAND health insurance project; see Mitchell and Phelps (1976).

What Can We Learn?

Of necessity this has been a highly selective review, in part because of my emphasis on the historical development of HE in one general journal and in part because of my own interests. I close with some comments on what I believe to be a very important area of research that is connected to both of the areas I discussed above: health as human capital and the role of public and private insurance schemes in allocating health care resources.

My work with Murphy suggests that the potential value of medical innovations is vastly larger than the flow of government-supported medical research, so the public-good case for greater research funding is somewhat compelling. But an important caveat is that the costs of research might be very small compared to the costs of treatments that are discovered, and these treatment costs could offset even the large potential surplus we calculate. Part of the reason has to do with insurance providers as downstream allocators of health care resources and technologies. In a nutshell, health care allocation via third-party payers, combined with consumers' desire for whatever care might help, creates an "invent it and they will use it" waste in the allocation of health care resources. Back upstream at the level of research, this distortion creates incentives for inefficiently costly health care advances. Going further, even technologies that would yield positive social surplus in a world of efficient downstream allocation can be a net social waste because the new technology is vastly overused. In these circumstances the major risk in medical research might not be the risk of failure to develop a new treatment after large up-front investment, but instead the risk of unaffordable success. This means that health care markets and the value of research are complements: greater efficiency in downstream allocation raises the value of medical research. Future research improving downstream allocations can therefore yield a sort of double dividend, so that is one area where I think substantial effort is warranted. Similarly, it is important to know the extent of the upstream distortion in research incentives, which I think is an even more challenging problem.

References

Acemoglu, Daron, and Simon Johnson. 2007. "Disease and Development: The Effect of Life Expectancy on Economic Growth." *J.P.E.* 115 (6): 925–85.

Arrow, Kenneth. 1963. "Uncertainty and the Welfare Economics of Medical Care." *A.E.R.* 53 (5): 941–73.

Becker, Gary S. 1964. *Human Capital: A Theoretical and Empirical Analysis, with Special Reference to Education.* New York: NBER.

———. 2007. "Health as Human Capital: Synthesis and Extensions." *Oxford Econ. Papers* 59:379–410.

Becker, Gary S., Tomas J. Philipson, and Rodrigo R. Soares. 2005. "The Quantity and Quality of Life and the Evolution of World Inequality." *A.E.R.* 95: 277–91.

Ben-Porath, Yoram. 1976. "Fertility Response to Child Mortality: Micro Data from Israel." *J.P.E.* 84, no. 4, pt. 2 (August): S163–S178.
Chandra, Amitabh, and Douglas O. Staiger. 2007. "Productivity Spillovers in Health Care: Evidence from the Treatment of Heart Attacks." *J.P.E.* 115 (1): 103–40.
Clemens, Jeffrey, and Joshua Gottlieb. 2017. "In the Shadow of a Giant: Medicare's Influence on Private Physician Payments." *J.P.E.* 125 (1): 1–39.
Cropper, M. L. 1977. "Health, Investment in Health, and Occupational Choice." *J.P.E.* 85 (December): 1273–94.
Cutler, David M., and S. Kadiyala. 2003. "The Return to Biomedical Research: Treatment and Behavioral Effects." In *Measuring the Gains from Medical Research: An Economic Approach*, edited by Kevin M. Murphy and Robert H. Topel. Chicago: Univ. Chicago Press.
Cutler, David M., Allison B. Rosen, and Sandeep Vijan. 2006. "Value of Medical Innovation in the United States: 1960–2000." *New England J. Medicine* 355 (9): 920–27.
Ehrlich, Isaac, and Hiroyuki Chuma. 1990. "A Model of the Demand for Longevity and the Value of Life Extension." *J.P.E.* 98 (4): 761–82.
Ehrlich, Isaac, and Francis T. Lui. 1991. "Intergenerational Trade, Longevity, and Economic Growth." *J.P.E.* 99 (5): 1029–59.
Feldstein, Martin. 1973. "The Welfare Loss of Excess Health Insurance." *J.P.E.* 81, no. 2, pt. 1 (March/April): 251–80.
Grossman, Michael. 1972. "On the Concept of Health Capital and the Demand for Health." *J.P.E.* 80 (March/April): 223–55.
Kenkel, Donald S. 1991. "Health Behavior, Health Knowledge, and Schooling." *J.P.E.* 99 (2): 287–305.
Kreader, J. L. 1988. "America's Prophet for Social Security: A Biography of Isaac Max Rubinow." PhD diss., Univ. Chicago.
Mincer, Jacob. 1958. "Investment in Human Capital and the Personal Distribution of Income." *J.P.E.* 66 (August): 281–302.
Mitchell, Bridger M., and Charles E. Phelps. 1976. "National Health Insurance: Some Costs and Effects of Mandated Employee Coverage." *J.P.E.* 84 (June): 553–72.
Murphy, Kevin M., and Robert H. Topel. 2003. "The Economic Value of Medical Research." In *Measuring the Gains from Medical Research: An Economic Approach*, edited by Kevin M. Murphy and Robert H. Topel. Chicago: Univ. Chicago Press.
———. 2006. "The Value of Health and Longevity." *J.P.E.* 114 (5): 871–904.
Mushkin, Selma J. 1962. "Health as an Investment." *J.P.E.* 70, no. 5, pt. 2 (October): 129–57.
Nordhaus, William. 2003. "The Health of Nations: The Contribution of Improved Health to Living Standards." In *Measuring the Gains from Medical Research: An Economic Approach*, edited by Kevin M. Murphy and Robert H. Topel. Chicago: Univ. Chicago Press.
Philipson, Tomas J., and Gary S. Becker. 1998. "Old-Age Longevity and Mortality-Contingent Claims." *J.P.E.* 106 (3): 551–73.
Rosen, Sherwin. 1988. "The Value of Changes in Life Expectancy." *J. Risk and Uncertainty* 1:285–304.
Rosenzweig, Mark R., and T. Paul Schultz. 1983. "Estimating a Household Production Function: Heterogeneity, the Demand for Health Inputs, and Their Effects on Birth Weight." *J.P.E.* 91 (5): 723–46.
Rubinow, I. M. 1904. "Labor Insurance." *J.P.E.* 12 (June): 362–81.
———. 1905. "Premiums in Retail Trade." *J.P.E.* 13 (September): 574–86.
———. 1915. "Standards of Sickness Insurance: I." *J.P.E.* 23 (March): 221–51.

———. 1917. "Medical Benefits under Workmen's Compensation: I." *J.P.E.* 25 (6): 580–620.
Sah, Raaj K. 1991. "The Effects of Child Mortality Changes on Fertility Choice and Parental Welfare." *J.P.E.* 99 (3): 582–606.
Schultz, T. W. 1960. "Capital Formation by Education." *J.P.E.* 68 (December): 571–83.
———. 1961. "Investment in Human Capital." *A.E.R.* 51 (March): 1–17.
Usher, D. 1973. "An Imputation of the Measure of Economic Growth for Changes in Life Expectancy." In *The Measurement of Economic and Social Performance*, edited by Milton Moss, 193–225. Conference on Research in Income and Wealth, Studies in Income and Wealth, vol. 38. New York: Columbia Univ. Press (for NBER).
Viscusi, W. K., and J. E. Aldy. 2003. "The Value of a Statistical Life: A Critical Review of Market Estimates throughout the World." *J. Risk and Uncertainty* 27:5–76.
Weisbrod, Burton A. 1971. "Costs and Benefits of Medical Research: A Case Study of Poliomyelitis." *J.P.E.* 79 (3): 527–44.
Wolfe, John R. 1985. "A Model of Declining Health and Retirement." *J.P.E.* 93 (6): 1258–67.

Agency Issues

Canice Prendergast

University of Chicago

Modern agency theory begins with Mirrlees (1976) and Holmstrom (1979) with a general prescription for how compensation can be used to alleviate agency issues. It proposes that an agent's pay should vary whenever there is information about her effort and that all information on performance should be used. Furthermore, the ability to resolve agency problems is limited only by risk-sharing considerations associated with random variation in performance measures. With some parametric restrictions, this also has the implication that greater randomness (uncertainty) reduces the intensity between pay and performance measures.

This work has been appropriately feted as the root from which modern agency theory derives. It does, however, suffer from one important problem: most people do not get paid this way. Instead, the current pay of most workers is insensitive to pretty much anything, relevant information is consciously not used, and the relationship between uncertainty and the

intensity of pay for performance if anything goes the wrong way. Because of this, the literature published in the journal has instead originated a series of avenues that—while respecting the logic of these contributions—at times better accord with observed practices. These avenues largely fit into two categories: (i) there are other ways to skin the cat of motivating workers, and (ii) in many settings, some of the assumptions of the canonical model do not hold. I consider each in turn.

Other Ways of Motivating

Workers are often embedded in firms, in markets, and in long-lasting relationships. Each of these issues has been explored to offer alternatives to the pay for performance logic of the canonical model.

First, most of us work in organizations that are hierarchical, with workers sorted into positions on the basis of performance and ability. Given this, it should not be surprising that an alternative source of motivation for workers is the possibility of promotion. The earliest and most important contribution here is Lazear and Rosen (1980), which argues for the role of tournaments in providing incentives. Tournaments are settings in which a group of agents compete for a set of fixed prizes rather than face a pay schedule that varies explicitly with their actions. Agents exert effort to change the probability of getting a better prize. In contrast to an individual piece rate setting, here what matters is relative performance in a particularly stark way, as only rank order performance determines pay. In simple settings, Lazear and Rosen show how tournaments can induce the first-best outcome without explicit pay for performance and offer some results on when tournaments dominate piece rates based on only individual performance. This logic has subsequently been extended by Green and Stokey (1983) and Malcomson (1984).

One difficulty that agency theory has faced has been the absence of clean empirical testing. Tournament theory is a welcome exception. There are two natural tests. First, do bigger prizes cause agents to work harder? In a number of (mostly sports) settings, this has been shown to be so. The second test—do perceived marginal probabilities of winning affect effort?—is both more subtle and more directly related to the notion that agents compete in probabilities of winning. The strongest work in that vein is Brown (2011), which showed that the presence of Tiger Woods in golfing tournaments acted as a disincentive for other golfers. They perform worse as they perceived the marginal value of effort to be lower.

Second, workers are embedded not only in firms but also in markets. Another theme that agency theory has explored successfully over the last three decades is how labor markets can act as an alternative to contracts. This logic derives from Fama (1980), which described how external audiences can constrain opportunities for moral hazard through the process

by which outside opportunities evolve. A simple example would be a professional baseball player. Despite the plethora of performance measures on his performance, pay for performance is rare. The reason, of course, is that there is an external market that sets the market pay of baseball players and a player who shirks can expect his reputation and future opportunities to worsen.

Fama does not claim that markets always offer enough discipline to solve agency problems. This issue has been elegantly formalized by Holmstrom (1999), and these two papers have now helped to develop a subfield of agency theory known as "career concerns." A more general characteristic of the orientation of the *JPE* is a focus on theoretically informed empirical testing. A particularly notable example of this in this field is the paper by Gibbons and Murphy (1992). They study the interaction between career concerns and formal pay for performance in the market for CEOs. Using a tightly specified model of career concerns interacting with formal pay for performance, they show how greater career concerns incentives reduce the need for formal pay for performance, as arises in their empirical results.

Finally, workers are located not only in hierarchies and markets but often in long-term employment relationships. This allows temporal considerations to enter optimal contracting, where long-term employment relations allow the possible use of *deferred compensation*. This is where an agent may not be rewarded for current performance today, but rather sometime in the future. There are by now many dynamic contracting papers whose optimal outcome is to hold back some component of pay until late in the agent's career. The logic is usually simple: by deferring a performance-related bonus until a worker is older, incentives for older workers improve while maintaining incentives for the younger worker (because by working hard now, she can be in line for that bonus later). One of the earliest formalizations of this logic is the study by Lazear (1979). His interest in that work is the need for mandatory retirement in optimal employment relationships. However, the reason why mandatory retirement is needed is that older workers are overpaid relative to their contemporaneous marginal productivity, which itself derives from the desire to defer compensation for the reason above.

The Assumptions Do Not Hold

Another series of extensions to the canonical model have arisen from the realization that many of its assumptions do not hold in many important settings. Here I provide a number of important examples.

The canonical model assumes that more pay for performance changes behavior only in ways that benefit the principal. However, by now it is well known that incentive pay can also induce the kind of dysfunctional

behavioral responses that cause pay for performance to backfire. Chevalier and Ellison (1997) offer a nice example. This line of research has become known as multitasking and typically mutes the use of performance pay. An elegant and tractable example of this is Baker (1992).

Other early contributions in the *JPE* on dysfunctional responses have focused on another characteristic of organizations, namely, the use of rules to allocate resources over allowing discretion. A series of papers have related this to dysfunctional behavioral responses. Consider the motivating example in Milgrom's (1988) important contribution on what he terms "influence activities." American Airlines needs to staff routes, and flight attendants have preferences over which routes they are given. Rather than allowing supervisors to assign "shifts" on the basis of the idiosyncratic preferences of employees, it uses a much simpler rule: routes are assigned by seniority, where workers with the most seniority pick first. (As a more substantive example, many firms use "last in, first out" rules for layoffs, despite the fact that some more senior workers could be less productive for the firm than their junior counterparts.) This arises in Milgrom's work as a way of deterring dysfunctional lobbying behavior by workers, where time is spent influencing superiors for resources for themselves rather than spending time on more productive activities. Said another way, while such bureaucratic rules may be inefficient at the point at which the decision is made, it may save sufficient resources at an earlier point to be worthwhile. Similar logic underlies Prendergast and Topel (1996) in a setting in which favoritism arises.

Public agencies are often accused of being unaccountable to their constituents. Foremost among these is the behavior of police forces. Another example of how ex post inefficient rules can be part of optimal oversight arises in Prendergast (2003). In that setting, legitimate consumer complaints are ignored. This is done because if public agents believe that their behavior is likely to be investigated on the basis of such complaints, they will simply capitulate to those consumers in settings in which they should not. (In the police example, they become more resistant to arresting suspects.) Once again, the response to dysfunctional behavioral actions is to ignore valuable information.

Probably the most important parallel exploration to agency contracting has been to understand governance in settings in which outcomes are noncontractible. There have been two lines of inquiry. The first has derived from the seminal work of Grossman and Hart (1986), where ownership of assets (or, in later work, the control of assets) can be used to mitigate agency concerns. This has led to the modern theory of the firm. This contribution and its extensions are described elsewhere in this issue by Rob Vishny and Luigi Zingales. The second area dealing with noncontractibility has addressed relational contracts, where repeated interactions can potentially resolve noncontractibility issues.

Work on noncontractible environments has also led to consideration of other instruments than compensation to align incentives. Contributions to the *JPE* have played a central role in this. An important example is Aghion and Tirole's (1997) work on real and formal authority. Consider a setting in which a worker cares not just about current pay and effort but also about the kinds of activities that they engage in. Furthermore, by exerting effort, she can identify which activity she prefers. Yet she will exert effort only if that preferred outcome is likely to be implemented. The problem for her is that her boss may not agree with her and instead overturns her recommendation. The principal will overturn, though, only if he is sufficiently sure of the right action. In settings in which he is not sure, the agent has real authority even though the agent may be subject to the formal authority of her boss. However, if the fear of being overturned is sufficiently salient to the agent, she will not exert any effort. In these settings, the firm may delegate formal authority to the agent. This work has been influential not just for its elegant and tractable modeling but also as the picture it paints—of organizations characterized by conflicts, with individuals vying for control—resonates with reality.

A feature of many institutions is conflict. Recent work in noncontractible settings has focused on the value of using workers who do not share the beliefs and preferences of their superiors. An early example is Che and Kartik (2009). The authors consider a setting in which an expert collects information to determine the right course of action, but she may be biased for or against that action (compared to the beliefs of her principal). Suppose that the principal could choose the bias of the agent: should she share the preferences of the principal? The intuitive suggestion that their beliefs should be aligned turns out not to be right. The reason is that an agent who shares the principal's belief realizes that if she exerts no effort, the principal is likely to do what the agent already thought was the right answer. Instead, the optimal strategy is to introduce disagreement between agent and principal, because an agent who believes that the principal's prior is wrong is more likely to work hard to dissuade him. Yet this is not costless, as she is less likely to reveal her information clearly.

Many workers, arguably most, are not rewarded on output measures. Instead their inputs are monitored, where they follow instructions provided by their superiors. Showing up on time and doing what is asked of them is the reality for most workers. Imagining input monitoring as an alternative to pay for performance has also helped to make progress on understanding one of the empirical difficulties faced by the literature. Specifically, there is little empirical support for what has become known as the trade-off between risk and incentives, where more uncertain environments would result in less pay for performance. Instead, the evidence seems more supportive of greater uncertainty leading to more

pay for performance. Prendergast (2002) addresses this by noting that for many workers, the alternative to pay for performance is a situation in which a superior tells a worker what to do. Now consider a setting with more uncertainty. In the canonical model, this only adds measurement noise to the ability to infer agent effort. This alone would attenuate pay for performance. However, in more uncertain settings, a superior may now additionally be less able to tell the agent what to do (as he knows less about what is going on). If so, the principal may need the agent to decide the right course of action in these uncertain settings, which likely leads to more pay for performance. Said another way, uncertainty may indeed render pay for performance costly; it may render the alternative even worse.

When a CEO increases the earnings of a company or a sales agent sells more, we can be pretty sure that this is a good thing. A final area of exploration in the field has been to consider settings in which it is not clear what output means. For example, when an auto mechanic tells you to have your car repaired at some expense, it is unclear if this reflects good or bad performance by the mechanic. Taylor (1995) studies this problem. A beautiful example of this inability to interpret performance measures is the work of Dewatripont and Tirole (1999) on advocates. Once again, consider a setting in which information needs to be collected on whether to carry out an action. The canonical agency model relies on a monotone likelihood ratio property (MLRP), where if it holds, pay will strictly increase in output (or profits). The difficulty here is that information can be positive or negative and can be offsetting. Specifically, an agent who has worked hard and found one piece of positive information and one piece of negative information finds herself in the same position as one who collects no information. Formally, this means that the MLRP of the canonical model fails, and it can render it impossible to have one person collect all the information. The alternative is advocacy: where one person collects only information that is positive and the other collects only the negative information, and they are rewarded if the outcome reflects the kind of information they collect. As such advocacy is so pervasive inside and outside organizations, this paper offers an insight far from the canonical model, but does so with a minimum of additional assumptions.

To conclude, it should be clear from this short essay that the literature on agency has come far from its original focus on the shape of compensation functions, and the central role played by the *JPE* in that development. It is perhaps worthwhile to conclude by noting the paucity of empirical work among the contributions above. Fields can thrive only when numbers are added to the Greek alphabet, and it is hoped that the *JPE* can play a significant role in promoting such work going forward.

References

Aghion, Philippe, and Jean Tirole. 1997. "Formal and Real Authority in Organizations." *J.P.E.* 105:1–29.

Baker, George. 1992. "Incentive Contracts and Performance Measurement." *J.P.E.* 100:598–614.

Brown, Jennifer. 2011. "Quitters Never Win: The (Adverse) Incentive Effects of Competing with Superstars." *J.P.E.* 119 (5): 982–1013.

Che, Yeon-Koo, and Navin Kartik. 2009. "Opinions as Incentives." *J.P.E.* 117 (5): 815–60.

Chevalier, Judith, and Glen Ellison. 1997. "Risk Taking by Mutual Funds as a Response to Incentives." *J.P.E.* 105:1167–1201.

Dewatripont, Matthias, and Jean Tirole. 1999. "Advocates." *J.P.E.* 107 (1): 1–39.

Fama, Eugene. 1980. "Agency Problems and the Theory of the Firm." *J.P.E.* 88:288–307.

Gibbons, Robert, and Kevin J. Murphy. 1992. "Optimal Incentive Contracts in the Presence of Career Concerns." *J.P.E.* 100:468–506.

Green, Jerry, and Nancy Stokey. 1983. "A Comparison of Tournaments and Contracts." *J.P.E.* 91:349–64.

Grossman, Sanford, and Oliver Hart. 1986. "The Costs and Benefits of Ownership: A Theory of Vertical and Lateral Organization." *J.P.E.* 94:691–719.

Holmstrom, Bengt. 1979. "Moral Hazard and Observability." *Bell J. Econ.* 10:74–91.

———. 1999. "Managerial Incentive Problems: A Dynamic Perspective." *Rev. Econ. Studies* 66 (1): 169–82.

Lazear, Edward. 1979. "Why Is There Mandatory Retirement?" *J.P.E.* 87 (6): 1261–84.

Lazear, Edward, and Sherwin Rosen. 1980. "Rank Order Tournaments as Optimal Labor Contracts." *J.P.E.* 88 (5): 841–64.

Malcomson, James. 1984. "Work Incentives, Hierarchy, and Internal Labor Markets." *J.P.E.* 92 (3): 486–507.

Milgrom, Paul. 1988. "Employment Contracts, Influence Activity, and Efficient Organization." *J.P.E.* 96:42–60.

Mirrlees, James. 1976. "The Optimal Structure of Incentives and Authority within an Organization." *Bell J. Econ.* 7 (1): 105–31.

Prendergast, Canice. 2002. "The Tenuous Trade-off between Risk and Incentives." *J.P.E.* 110 (5): 1071–1102.

———. 2003. "The Limits of Bureaucratic Efficiency." *J.P.E.* 111 (5): 929–59.

Prendergast, Canice, and Robert Topel. 1996. "Favoritism in Organizations." *J.P.E.* 104 (5): 958–78.

Taylor, Curtis. 1995. "The Economics of Breakdowns, Checkups, and Cures." *J.P.E.* 103:53–74.

Information Economics

Emir Kamenica

University of Chicago

When George Stigler published "The Economics of Information" in this journal a little over 50 years ago (1961), he was justified in his complaints about the absence of research on the topic. Yet, this complaint was soon to become obsolete. Figure 1 reports the incidence of the phrases "information economics" and "economics of information" normalized by the use of the word "economics" in books published from 1900 to 2008. The figure clearly shows that Stigler's article was a harbinger of a new field of inquiry.[1] The field grew quickly for the next 25 years, reached a plateau that lasted from the mid-1980s until the late 1990s, and then experienced another growth spurt. During the last decade, 13 percent of the articles published in the *JPE* had the word "information" in their abstract.

I. Information Acquisition

One early strand of the literature developed *search models* in which individuals incur costs to acquire information for private use. Stigler's (1961) model, in which agents decide ex ante on the number of alternatives to sample, was replaced by a sequential search formulation (McCall 1970). This formulation provided sharp and lasting intuitions about individual motives for acquiring additional information given the distribution of options but was not well suited to explain how such distributions arise in the first place (Diamond 1971). This line of research eventually grew into the matching-and-bargaining and directed-search models that are now widely used in labor macroeconomics but have little direct contact with the rest of information economics.

In other parts of macroeconomics, agents' information is now often endogenized through *rational inattention* models (Sims 2003) that postulate that the cost of information acquisition is proportional to the reduction in Shannon entropy.[2] While the rational inattention approach may seem

[1] Of course, one can almost always push intellectual origins of any field further into the past. Hayek (1945) is an important early reference in information economics. It is likely responsible for the 1-year blip in use of the phrases in 1946 visible in fig. 1.

[2] Matejka and McKay (2015) show that rational inattention can also be used to provide a microfoundation for the multinomial logit model of choice often used in empirical industrial organization.

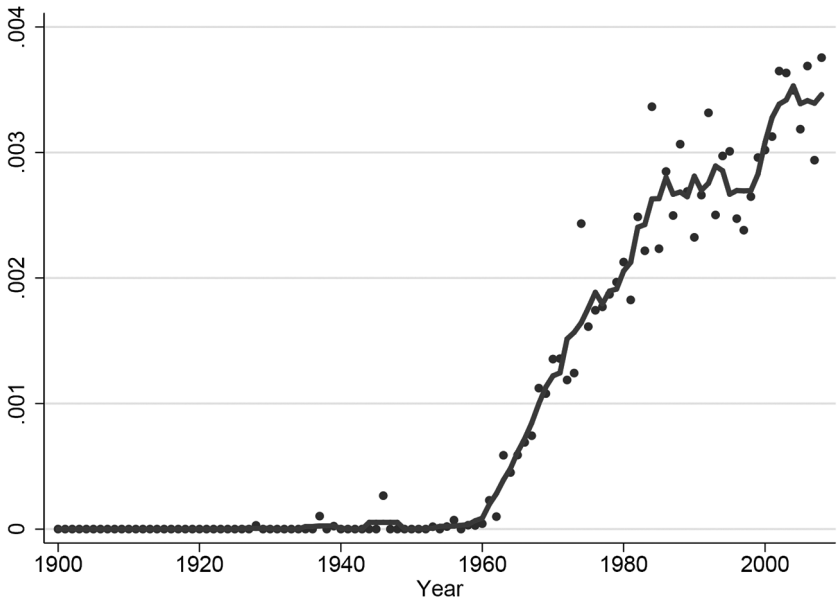

FIG. 1.—Information economics over time. Data from Google Books Ngram Viewer (Michel et al. 2011). The *y*-axis depicts the share of "information economics" among all 2-grams plus the share of "economics of information" among all 3-grams divided by the share of "economics" among all 1-grams (all phrases case insensitive) in books published that year. The dots indicate raw data by year while the solid line depicts a 5-year moving average.

like a return to the ex ante formulation of information-gathering costs, Hebert and Woodford (2016) and Morris and Strack (2017) develop sequential sampling models that generate the static cost functions employed in the rational inattention literature.

The aforementioned papers focus on information acquisition by isolated decision makers. Grossman and Stiglitz (1980) show that in certain market settings, the value of acquiring information is lower when a higher fraction of other agents are informed; then full-information equilibrium outcomes are not possible. Information externalities have also been used to understand herd behavior, fashions, and customs. Banerjee (1992) and Bikhchandani, Hirshleifer, and Welch (1992) point out that if a large number of individuals all share the same state-dependent preferences, observe discrete signals about the state, and take their actions sequentially (after seeing what those before them have done), information cascades will arise: individuals start to ignore their own information and simply mimic the behavior of others.[3] Following Bikhchandani et al. in the pages

[3] Smith and Sorensen (2000) and Eyster and Rabin (2014) discuss some generalizations and limitations of this result.

of this journal, Bernheim (1994) proposed an alternative theory of conformity, one built on signaling considerations. These papers were a part of a broader shift whereby information joined the center stage, alongside preferences and technology, as a key factor for understanding socioeconomic phenomena.[4]

II. Asymmetric Information

Parts of the literature that focused on endogenizing information through information acquisition tended to consider environments with limited strategic interaction. At the same time, a largely separate literature on Bayesian games (with exogenously specified information) was coming of age. The early work by Harsanyi and others focused on "pure" game theory,[5] but then in the 1970s, a series of applied theory papers—Akerlof's (1970) market for lemons, Spence's (1973) job signaling, and Stiglitz and coauthors' analysis of screening (Stiglitz 1975; Rothschild and Stiglitz 1976; Stiglitz and Weiss 1981)—identified the crucial importance of asymmetric information. Remarkably, Rothschild and Stiglitz's (1976) analysis of insurance markets opens with the sentence, "Economic theorists traditionally banish discussions of information to footnotes" (629). Yet, with hindsight, it is clear that by the mid-1970s, information economics was flourishing.

While Akerlof, Spence, and Stiglitz played key roles in the "asymmetric information revolution," related ideas were also being explored by other scholars around the same time. For example, writing in the *JPE*, Nelson (1970, 1974) proposed, in informal terms, a Spence-like channel through which advertising could be understood as a costly signal of a firm's quality. Milgrom and Roberts (1986) subsequently formalized this view (and filled an important gap in Nelson's original argument). The core ideas about asymmetric information developed in the 1970s continue to play a key role in our analysis of health care, banking, education, and many other markets.

III. Communication

The fact that asymmetric information can have stark consequences—and is often detrimental to social welfare—poses the question of whether economic agents will share their private information so as to eliminate the informational asymmetry. The 1980s saw the development of two widely used frameworks for studying information exchange. Crawford and Sobel

[4] As an aside, one might argue that all technology is a form of information, but information-theoretic concepts have so far not been widely applied to the study of technological change.

[5] One important exception is Vickrey's (1961) analysis of auctions.

(1982) consider a *cheap talk* environment in which an informed sender can costlessly convey any message to an uninformed receiver, who then takes an action that affects the welfare of both parties. Grossman (1981) and Milgrom (1981) introduced *verifiable message* models in which the sender can choose how much of his private information to disclose but cannot tell outright lies. In cheap talk settings, the sender's problem tends to be the receiver's inertia:[6] when the player's interests are insufficiently aligned, it is not possible for the sender to convey any information even if both parties would be better off if he were able to do so. In contrast, in verifiable message models, the sender's problem is often the receiver's reaction to silence: when the sender's preferences are monotone, full information is always conveyed in equilibrium even if the sender wishes this were not so.[7] This result has played an important role in discussions of government policies regarding disclosure mandates.

Economists' analysis of strategic obstacles to information exchange complements the computer science and engineering literatures that focus on the technological constraints on information transmission (Cover and Thomas 2006). One exciting and underexplored area is the interaction between the two types of constraints. For example, Blume, Board, and Kawamura (2007) point out that the presence of technological limitations can alleviate the impact of the strategic obstacles and thus improve communication.

IV. Recent Developments

The currently most active area of research in information economics is probably *information design*, a confluence of work on Bayes correlated equilibria (Bergemann and Morris 2013) and Bayesian persuasion (Kamenica and Gentzkow 2011). Bayes correlated equilibria take as given players' state-dependent preferences and describe the set of all possible outcomes that could arise regardless of what each player knows about the state and about what others know. Bayesian persuasion models seek to identify the best outcome from this set given some objective function. Thus, research on information design seeks to identify the optimal informational environment (who should know what and when) taking as given the preferences of the players and some objective function over the players' actions. Information design can be seen as a parallel to mechanism design: in the latter, the designer can choose the game but has no control over the information structure, whereas in the former the designer can

[6] It is possible to construct a cheap talk game with an equilibrium that gives the sender a lower payoff than he would get under no communication, but such examples tend to be somewhat contrived.

[7] This result requires that the receiver is certain of what information the sender has (Dye 1985).

choose the information structure but has no control over the game (cf. Bergemann and Morris 2016; Taneva 2016). In just the few years since its inception, information design has been used to address issues in banking regulation, internet advertising, censorship, entertainment, price discrimination, traffic congestion, and so forth.[8]

In closing, it may be worthwhile to note that, over the last century, information has come to play an important role in other disciplines besides economics. In biology, we discovered that all complexity of life is encoded as information about sequences of nucleic acids. In physics, not only are informational constraints (expressed as the uncertainty principle) a central feature of quantum mechanics, a provocative "it from bit" doctrine proposes that "all things physical are information-theoretic in origin" (Wheeler 1990, 311). Perhaps in some distant future, information-theoretic approaches may reveal structures shared by biological, physical, and economic systems.

References

Akerlof, George. 1970. "The Market for 'Lemons': Quality Uncertainty and the Market Mechanism." *Q.J.E.* 84 (3): 488–500.

Banerjee, Abhijit. 1992. "A Simple Model of Herd Behavior." *Q.J.E.* 107 (3): 797–817.

Bergemann, Dirk, and Stephen Morris. 2013. "Robust Predictions in Games with Incomplete Information." *Econometrica* 81 (4): 1251–1308.

———. 2016. "Information Design, Bayesian Persuasion, and Bayes Correlated Equilibrium." *A.E.R. Papers and Proc.* 106 (5): 586–91.

Bernheim, B. Douglas. 1994. "A Theory of Conformity." *J.P.E.* 102 (5): 841–77.

Bikhchandani, Sushil, David Hirshleifer, and Ivo Welch. 1992. "A Theory of Fads, Fashion, Custom, and Cultural Change as Informational Cascades." *J.P.E.* 100 (5): 992–1026.

Blume, Andreas, Oliver Board, and Kohei Kawamura. 2007. "Noisy Talk." *Theoretical Econ.* 2 (4): 395–440.

Cover, Thomas, and Joy Thomas. 2006. *Elements of Information.* 2nd ed. Hoboken, NJ: Wiley.

Crawford, Vincent, and Joel Sobel. 1982. "Strategic Information Transmission." *Econometrica* 50 (6): 1431–51.

Diamond, Peter. 1971. "A Model of Price Adjustment." *J. Econ. Theory* 3 (2): 156–68.

Dye, Ronald. 1985. "Disclosure of Nonproprietary Information." *J. Accounting Res.* 23 (1): 123–45.

Ely, Jeffrey, Alexander Frankel, and Emir Kamenica. 2015. "Suspense and Surprise." *J.P.E.* 123 (1): 215–60.

Eyster, Erik, and Matthew Rabin. 2014. "Extensive Imitation Is Irrational and Harmful." *Q.J.E.* 129 (4): 1861–98.

Grossman, Sanford. 1981. "The Informational Role of Warranties and Private Disclosure about Product Quality." *J. Law and Econ.* 24 (3): 461–83.

[8] Contributions to information design published in the *JPE* include Rayo and Segal (2010), Kremer, Mansour, and Perry (2014), and Ely, Frankel, and Kamenica (2015).

Grossman, Sanford, and Joseph Stiglitz. 1980. "On the Impossibility of Informationally Efficient Markets." *A.E.R.* 70 (3): 393–408.
Hayek, F. A. 1945. "The Use of Knowledge in Society." *A.E.R.* 35 (4): 519–30.
Hebert, Benjamin, and Michael Woodford. 2016. "Rational Inattention with Sequential Information Sampling." Working paper, Stanford Univ.
Kamenica, Emir, and Matthew Gentzkow. 2011. "Bayesian Persuasion." *A.E.R.* 101 (6): 2590–2615.
Kremer, Ilan, Yishay Mansour, and Motty Perry. 2014. "Implementing the 'Wisdom of the Crowd.'" *J.P.E.* 122 (5): 988–1012.
Matejka, Filip, and Alisdair McKay. 2015. "Rational Inattention to Discrete Choices: A New Foundation for the Multinomial Logit Model." *A.E.R.* 105 (1): 272–98.
McCall, J. J. 1970. "Economics of Information and Job Search." *Q.J.E.* 84 (1): 113–26.
Michel, Jean-Baptiste, Yuan Kui Shen, Aviva Presser Aiden, et al. 2011. "Quantitative Analysis of Culture Using Millions of Digitized Books." *Science* 331 (6014): 176–82.
Milgrom, Paul. 1981. "Good News and Bad News: Representation Theorems and Applications." *Bell J. Econ.* 12 (2): 380–91.
Milgrom, Paul, and John Roberts. 1986. "Price and Advertising Signals of Product Quality." *J.P.E.* 94 (4): 796–821.
Morris, Stephen, and Philipp Strack. 2017. "The Wald Problem and the Equivalence of Sequential Sampling and Static Information Costs." Working paper, Princeton Univ.
Nelson, Phillip. 1970. "Information and Consumer Behavior." *J.P.E.* 78 (2): 311–29.
———. 1974. "Advertising as Information." *J.P.E.* 82 (4): 729–54.
Rayo, Luis, and Ilya Segal. 2010. "Optimal Information Disclosure." *J.P.E.* 118 (5): 949–87.
Rothschild, Michael, and Joseph Stiglitz. 1976. "Equilibrium in Competitive Insurance Markets: An Essay on the Economics of Imperfect Information." *Q.J.E.* 90 (4): 629–49.
Sims, Christopher. 2003. "Implications of Rational Inattention." *J. Monetary Econ.* 50 (3): 665–90.
Smith, Lones, and Peter Sorensen. 2000. "Pathological Outcomes of Observational Learning." *Econometrica* 68 (2): 371–98.
Spence, Michael. 1973. "Job Market Signaling." *Q.J.E.* 87 (3): 355–74.
Stigler, George. 1961. "The Economics of Information." *J.P.E.* 69 (3): 213–25.
Stiglitz, Joseph. 1975. "The Theory of 'Screening,' Education, and the Distribution of Income." *A.E.R.* 65 (3): 283–300.
Stiglitz, Joseph, and Andrew Weiss. 1981. "Credit Rationing in Markets with Imperfect Information." *A.E.R.* 71 (3): 393–410.
Taneva, Ina. 2016. "Information Design." Working paper, Univ. Edinburgh.
Vickrey, William. 1961. "Counterspeculation, Auctions, and Competitive Sealed Tenders." *J. Finance* 16 (1): 8–37.
Wheeler, John A. 1990. "Information, Physics, Quantum: The Search for Links." In *Complexity, Entropy, and the Physics of Information*, edited by W. Zurek, 309–36. Redwood City, CA: Addison-Wesley.

The Continuing Impact of Sherwin Rosen's "Hedonic Prices and Implicit Markets: Product Differentiation in Pure Competition"

Michael Greenstone

University of Chicago and National Bureau of Economic Research

Sherwin Rosen's landmark paper "Hedonic Prices and Implicit Markets: Product Differentiation in Pure Competition" (1974) fundamentally altered understanding of several fields of economics, including environmental, labor, public, and urban economics. At its most general level, the paper outlines how the market solves the problem of matching buyers and sellers of multidimensional goods. Since virtually all goods have multiple characteristics, the paper's framework has proven to be broadly applicable across a range of economic topics. It is therefore no surprise that the paper is the sixth most cited in the *Journal of Political Economy*'s history. To note just a few examples of its application, the paper has served as the foundation for inferring households' valuations of air quality (Chay and Greenstone 2005), understanding "equalizing differences" in the labor market (e.g., Brown 1980), estimating the incidence of government policies (e.g., Gruber 1994), and describing the equilibrium allocation of individuals and firms across locations (e.g., Roback 1982; Greenstone, Hornbeck, and Moretti 2010). Part of the enduring appeal of the paper is that it outlines a method for estimating relationships of extraordinary importance for the determination of optimal policy, particularly individuals' willingness to pay for goods and services for which there are not explicit markets. Canonical examples include environmental quality, school quality, crime, other amenities, and mandated government benefits.

The paper's starting point is that virtually all goods are heterogeneous (e.g., houses, jobs, and cities) and that while we can observe their overall price, this alone does not shed much light on the demand for and supply of their characteristics. To make progress on these more fundamental economic questions, Rosen's paper outlines an approach that considers goods to be a vector of their "utility-bearing attributes or characteristics." The paper's central contribution is to model how consumers' and sup-

I thank Lucas Davis for sage advice and criticisms and Michael Galperin for outstanding research assistance.

pliers' optimizing behavior governs the data-generating process that delivers the potentially observable equilibrium relationship between characteristics and their prices.

This paper briefly reviews the model that Rosen outlined, discusses the success of efforts to apply it empirically to gain understanding of several key relationships, and outlines areas for future research.

I. A Brief Review of Rosen's Hedonic Model

In Rosen's (1974) formulation, a differentiated good is described by a vector of its characteristics, $\mathbf{C} = (c_1, c_2, ..., c_n)$. In the case of a house, these characteristics may include structural attributes (e.g., number of bedrooms), neighborhood public services (e.g., local school quality), and local environmental amenities (e.g., air quality). Thus, the market price of the ith house can be written as

$$P_i = P(c_{i1}, c_{i2}, ..., c_{in}). \tag{1}$$

The partial derivative of $P(\cdot)$ with respect to the jth characteristic, $\partial P/\partial c_j$, is referred to as the marginal implicit price. It is the marginal price of the jth characteristic, holding constant all other characteristics, and is implicit in the overall price of the house.

In the hedonic model, the locus between housing prices and a given characteristic, called the hedonic price schedule (HPS), is generated by the equilibrium interactions of consumers and producers. It is assumed that markets are competitive and that all consumers rent one house at the market price. Consumers' utility depends on consumption of the numeraire X (with price equal to one) and the vector of house characteristics:

$$u = u(X, \mathbf{C}). \tag{2}$$

The budget constraint is expressed as $I - P - X = 0$, where I is income.

Maximization of (2) with respect to the budget constraint reveals that individuals choose levels of each of the characteristics to satisfy $(\partial U/\partial c_j)/(\partial U/\partial x) = \partial P/\partial c_j$. Thus, the marginal willingness to pay for c_j (e.g., air quality) must equal the marginal cost of an extra unit of c_j in the market.

It is convenient to substitute the budget constraint into (2), which gives $u = u(I - P, c_1, c_2, ..., c_n)$. Inverting this equation and holding all characteristics but j constant results in an expression for willingness to pay for c_j:

$$B_j = B_j(I - P, c_j; \mathbf{C}^*_{-j}, u^*). \tag{3}$$

Here, u^* is the highest level of utility attainable given the budget constraint and $\mathbf{C}^*_{-j}$ is the vector of the optimal quantities of other character-

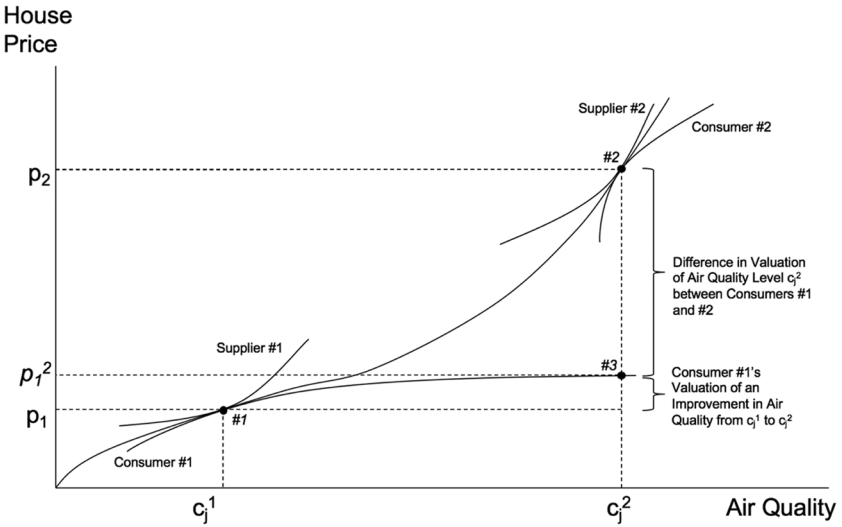

FIG. 1.—Bid curves, offer curves, and the equilibrium HPS in a hedonic market for air quality.

istics. This is referred to as a bid (or indifference) curve, because it reveals the maximum amount that an individual would pay for different values of c_j, holding utility constant.

Heterogeneity in individuals' bid functions due to differences in preferences and/or incomes leads to differences in the chosen quantities of a characteristic. This is depicted in figure 1, which plots the HPS and bid curves for c_j of two consumer types. The consumers are denoted as types #1 and #2; there is potentially an unlimited number of consumer types, each of which has a bid curve that is tangent to the HPS. Each bid function reveals the standard declining marginal rate of substitution between c_j and X (because $X = I - P$). The two types choose houses in locations where their marginal willingness to pay for c_j is equal to the market-determined marginal implicit price, which occurs at c_j^1 and c_j^2, respectively. Given market prices, these consumers' utilities would be lower at sites with any other level of local environmental quality.

The other side of the market comprises suppliers of housing services. It is assumed that suppliers are heterogeneous because of differences in their cost functions. This heterogeneity may result from differences in the land they own. For example, it may be very expensive to provide a high level of air quality on a plot of land located near a steel factory. By inverting a supplier's profit function, we can derive the supplier's offer curve for the characteristic c_j:

$$O_j = O_j(c_j; \mathbf{C}^*_{-j}, \pi^*), \tag{4}$$

where π^* is the maximum available profit given the supplier's cost function and the HPS. Figure 1 depicts offer curves for two types of suppliers. With this setup, individuals who live in a house that they own would be both consumers and suppliers, and their supplier self would rent to their consumer self.

The HPS is formed by tangencies between consumers' bid and suppliers' offer functions. At each point on the HPS, the marginal price of a housing characteristic is equal to an individual consumer's marginal willingness to pay (MWTP) for that characteristic and an individual supplier's marginal cost of producing it. From the consumer's perspective, the gradient of the HPS with respect to air quality gives the equilibrium differential that compensates consumers for accepting the disutility (e.g., increased health risk, aesthetic disamenities) associated with poorer air quality. Put another way, areas with high levels of air pollution must have lower housing prices to attract potential homeowners, and the HPS reveals the price that allocates consumers across locations (and pollution levels). From the suppliers' perspective, the gradient of the HPS reveals the equilibrium marginal cost of supplying a cleaner local environment.

The HPS itself is useful for a limited range of welfare analysis. The gradient at each point along the HPS reveals the MWTP for the set of consumers that have sorted themselves to the pair of prices and quantities of the relevant characteristic. Thus, it is possible to infer the welfare gain associated with a marginal change for different segments of the population. The overall HPS can be used to determine the average MWTP in the relevant population.

Figure 1 illustrates that knowledge of the HPS is not sufficient to conduct welfare analysis for nonmarginal changes, at least in partial equilibrium. Consider an improvement in air quality from c_j^1 to c_j^2. Consumer 1's valuation or willingness to pay for this change is equal to the difference between p_1^2 and p_1; yet the HPS would suggest that the gain is much larger, equal to the difference between p_2 and p_1. The difficulty is that we observe only one point on each person's bid function (i.e., the pair of prices and quantities that they choose). Other points are observed only for other individuals who presumably have different tastes or income levels. Thus, the HPS cannot be used to determine the welfare consequences of nonmarginal changes in key characteristics.

Rosen's solution to this problem is a two-step econometric procedure that, in principle, delivers the HPS and consumer's bid functions.[1] Consider again the case of houses and air quality. The procedure's first step is to regress house prices against all housing characteristics, including air quality, allowing for their effects to be nonlinear:

[1] An analogous procedure delivers suppliers' offer functions.

$$p = \alpha + f(c_1, c_2, \ldots, c_n) + \varepsilon. \tag{5}$$

The marginal implicit price of air quality is the derivative of housing prices with respect to air quality. This quantity is used in the following second-stage equation:

$$\partial p(\mathbf{C})/\partial c_{\text{air quality}} = \alpha' + g\left(c_{\text{air quality}}\right) + \varepsilon', \tag{6}$$

where ε' includes all demand shifters, such as tastes and income, that in principle are observable and can be included as covariates. The Rosen method then calls for evaluating $g(c_{\text{air quality}})$ at different values of $c_{\text{air quality}}$ to trace out the bid function.

Successful implementation of this two-step procedure would have tremendous practical value, because it would reveal consumers' demand primitives. Thus, it would be possible to obtain measures of the welfare effects of nonmarginal changes in the characteristics of goods. This helps to explain the enduring influence of Rosen's model.

II. The Rosen Model in Action

Rosen developed this method with the aim of improving understanding of the world. Indeed, he wrote, "We anticipate that the basic conceptual framework outlined above will have a variety of applications to many practical problems" (Rosen 1974, 54). However, at the time of Rosen's paper, the economics profession did not adequately appreciate the difficulty in obtaining causal estimates of empirical relationships, particularly cross-sectional ones. Thus, the paper paid little attention to the challenges of consistent estimation of the HPS. Furthermore, the paper refers to the second-stage estimation as a "garden variety identification problem," which, with the benefit of more than 40 years of continued research, also appears quite optimistic.

For the first 30 years after its publication, the hedonic approach was largely unsuccessful empirically, and indeed, my judgment is that its practical value was in question. However, the last 10–15 years have seen great advances in the empirical credibility focused on the estimation of the first stage, or HPS. As Rosen originally conceived it, there are few, if any, instances of credible estimation of the second stage.

This section outlines the challenges with estimation, where there have been successes, and where more work is needed.

A. Estimation of the HPS

The consistent estimation of equation (1) is the foundation on which any welfare calculation rests. The reason is that the welfare effects of a

marginal change in a characteristic are obtained directly from the hedonic price schedule. Furthermore, inconsistent estimation of the HPS will result in an inconsistent MWTP function, invalidating any welfare analysis of nonmarginal changes regardless of the method used to recover preference or technology parameters. These challenges were apparent to at least one researcher just 1 year after the publication of the Rosen paper: "I have entirely avoided . . . the important question of whether the empirical difficulties, especially correlation between pollution and unmeasured neighborhood characteristics, are so overwhelming as to render the entire method useless. I hope that . . . future work can proceed to solving these practical problems. . . . The degree of attention devoted to this [problem] . . . is what will really determine whether the method stands or falls" (Small 1975, 107).

For roughly 30 years, these words proved prophetic as researchers found consistent estimation of the HPS as in equation (1) to be extraordinarily challenging. For example, the cross-sectional estimation of the HPS exhibited signs of misspecification (e.g., great sensitivity to the exact set of controls and frequent findings of perversely signed estimates) in a number of settings, including the relationships between land and/or house prices and air quality (Smith and Huang 1995), school quality (Black 1999), proximity to hazardous waste sites (Greenstone and Gallagher 2008), and climate variables (Deschenes and Greenstone 2007). Similar problems characterized the estimation of compensating wage differentials for job characteristics, such as the risk of injury or death: in a wide range of studies, the regression-adjusted association between wages and many job amenities was found to be weak and often had a counterintuitive sign (e.g., Smith 1979; Brown 1980; Black and Kneisner 2003).

One attempted solution to the challenges of estimating cross-sectional hedonic equations was to move to a panel setting. An especially important example of this method is Brown's (1980) paper on equalizing differences in the labor market, which included person-specific fixed effects and year fixed effects. However, the results were disappointing in that they remained sensitive to small changes in specification and often had the counterintuitive sign. Indeed, Brown concluded, "The impacts of the intercepts on the coefficients of job characteristics vary considerably, and there is no marked improvement in the correspondence between these coefficients and a priori predictions" (130).

Table 1, taken from Chay and Greenstone (2005), clearly illustrates the problems the literature encountered with the cross-sectional and panel data approaches. It presents "conventional" estimates of the capitalization of total suspended particulates (TSPs) air pollution into housing values based on fitting regressions with county-level census data. Across different panels, the model is estimated using cross-sectional data from 1970

TABLE 1
ESTIMATES OF THE EFFECT OF TSPs POLLUTION ON LOG HOUSING VALUES

	(1)	(2)	(3)	(4)
	A. 1970 Cross Section			
Mean TSPs (1/100)	.032	−.062	−.040	−.024
	(.038)	(.018)	(.017)	(.017)
R^2	.00	.79	.84	.85
Sample size	988	987	987	987
	B. 1980 Cross Section			
Mean TSPs (1/100)	.093	.096	.076	.027
	(.066)	(.031)	(.030)	(.028)
R^2	.00	.82	.89	.89
Sample size	988	984	984	984
	C. 1970–80 (First Differences)			
Mean TSPs (1/100)	.102	.024	.004	−.006
	(.032)	(.020)	(.016)	(.014)
R^2	.02	.55	.65	.73
Sample size	988	983	983	983
County Data Book covariates	no	yes	yes	yes
Flexible form of county covariates	no	no	yes	yes
Region fixed effects	no	no	no	yes

NOTE.—The housing and overall consumer price index series are used to deflate all housing values to 1982–84 dollars. The TSPs data are derived from the Environmental Protection Agency's network of pollution monitors. The 1970 (1980) mean TSPs concentration is the average across all counties' mean TSPs concentration from 1969 to 1972 (1977 to 1980). Each county's annual mean TSPs concentration is calculated as the weighted average of the geometric mean concentrations of each monitor in the county, using the number of observations per monitor as weights. The county-level mean across multiple years (e.g., 1969–72) is the average of the annual means. The flexible functional form includes quadratics, cubics, and interactions of the variables as controls. The mean of the natural log of 1970 housing prices is 10.55. The means of the dependent variables in panels B and C are 10.82 and 0.27, respectively. Standard errors (in parentheses) are estimated using the Eicker-White formula to correct for heteroskedasticity. This table appears as table 3 in Chay and Greenstone (2005, 408).

(panel A) and 1980 (panel B) and by first-differencing the 1970 and 1980 data (panel C) to remove the influence of time-invariant unobservables. Across columns 1–4, an increasing number of covariates are used to adjust the effect of TSPs on housing values.

The instability of the estimates across specifications and within a specification across panels is striking and suggests that the conventional approach to estimating the HPS is prone to misspecification. The column 2 results illustrate this point powerfully, because they use a specification typical of the three decades of research following the publication of Rosen's paper. With this specification, the 1970 data suggest that a one-unit decline in TSPs is associated with a 0.06 percent increase in housing values, the 1980 data suggest that it is associated with a 0.10 percent decrease, and

the first-difference data find no meaningful relationship. The point is that even when holding the specification constant, it is possible to find virtually any effect that one desires; not even the sign is constant across the different data sets. Unfortunately, this pattern of results is not specific to the relationship between housing prices and TSPs; for example, Black and Kneisner's (2003) paper on the value of a statistical life convincingly demonstrates a similar variability when estimating the relationship between wages and on-the-job fatality risks. The result of this uncertainty was that despite the publication of perhaps hundreds of empirical papers based on cross-sectional and panel data estimation of the hedonic model, the prevailing sense was that the resulting estimates were plagued by omitted variable bias. The hedonic method's practical relevance was in doubt.

The late 1990s and early 2000s marked a turning point in the field as the "credibility revolution" sparked renewed interest in estimation of Rosen-style hedonic equations. Specifically, researchers began to identify quasi-experimental variation in the variables of interest that was plausibly unrelated to unobserved determinants of the studied outcome (e.g., housing prices, wages). In a quasi-experiment, variation in the variable of interest is determined by nature, politics, an accident, or some other action beyond the researcher's control. The identifying assumption is that this variation is exogenous, and in high-quality quasi experiments, researchers understand well the source of this variation and characterize it clearly for the reader.

This turning point manifested itself with great advances in understanding of a wide range of topics. As part of my dissertation, ultimately published in the *JPE*, Kenneth Chay and I exploited spatial and temporal variation in the introduction of the Clean Air Act, along with knowledge of the exact rule that determined that variation, to study the relationship between TSPs concentration and housing values. We found that the elasticity of housing values with respect to TSPs concentrations ranges from -0.20 to -0.35 (Chay and Greenstone 2005). In a compelling study, Davis (2004) found that an increase in the incidence of pediatric leukemia in a Nevada county reduced housing prices by about 15 percent. Building on the important Black (1999) paper, Bayer, Ferreira, and McMillan (2007) estimate households' willingness to pay for school quality using variation across boundaries for catchment zones.[2] There are many compelling examples, and any effort to be comprehensive will surely fall short; but a few other quasi-experimental findings include a negative MWTP for proximity to a convicted sex offender (Linden and Rockoff 2008), willing-

[2] This paper's contribution goes beyond credible estimation of the HPS. It provides great insight into the relationship between estimates of willingness to pay from hedonic price regressions vs. those from random utility model discrete choice approaches to demand estimation.

ness to pay for Superfund cleanups that is smaller than their costs (Greenstone and Gallagher 2008), and an 11 percent decline in housing values within 0.5 mile of newly opened industrial plants that emit toxic air pollutants (Currie et al. 2015).

Overall, the application of the quasi-experimental approach has breathed new life into the Rosen model. The proliferation of papers in recent years has generated tremendous insights across a wide range of areas where there was previously little credible empirical work that could guide people's understanding. There is currently tremendous interest in randomized control trial experiments in economics, but I am not aware of any field experiment applications of Rosen's hedonic model to date (although they would be an incredible addition both substantively and methodologically).

B. What about the Bid Functions?

The greatest promise of the Rosen model is not just to consistently estimate the HPS but to recover individuals' bid functions (and suppliers' offer functions). Estimation of these primitives allows for assessing the welfare consequences of nonmarginal changes and counterfactual policies; this is the difference between estimating what has been and what might be. The great challenge, of course, is that estimating individuals' bid functions requires the quite difficult task of observing the same individual or taste type facing two sets of prices.

While the estimation of the HPS faces a formidable enemy (i.e., omitted variables), the estimation of bid functions has found itself facing seemingly even more formidable foes. Following Rosen's paper, there were some initial efforts to recover these bid functions (e.g., Palmquist 1984). However, Brown and Rosen (1982) cautioned researchers by demonstrating the hedonic method's reliance on potentially strong functional form assumptions. A pair of later papers showed that efforts to infer consumers' bid functions from the HPS are further undermined by taste-based sorting and that the difficulty of addressing this problem using standard exclusion restrictions means that quite strong assumptions are necessary (Bartik 1987; Epple 1987). Thus by the late 1980s, there was a sense that credible estimation of bid functions with the Rosen approach was not possible, and efforts to even attempt their estimation began to disappear.

In response to this decline, several papers tried to revive Rosen's method's aim of recovering demand primitives through structure or by invoking alternative assumptions. Epple and Sieg (1999) outlined a "locational equilibrium" model that can be used to develop estimates of the demand primitives. Returning to the Rosen model's roots, Ekeland, Heckman, and Nesheim (2004) outline the assumptions necessary to identify the de-

mand (and supply) functions in an additive version of the hedonic model with data from a single market.[3] Bajari and Benkard (2005) similarly return to the Rosen model and, through alternative approaches to the first and second stages, outline an approach to construct bounds on individuals' utility parameters and other economic objects. Each of these approaches is promising, but they all require potentially strong assumptions and to date their influence on applied work has not been tremendous (although researchers continue to experiment with them).

III. Conclusions

Sherwin Rosen's hedonic method is a great achievement of economic theory. Taking as its starting point an observable relationship that by itself does not shed any light on the economic behavior underlying it, the paper outlines a model of buyer and seller optimizing behavior to explain the process that generates what is observed in the data. In outlining this framework, Rosen fundamentally altered how we understand the world.

On the applied side, the application of quasi-experimental techniques to the estimation of the HPS has reinstated the hedonic method as a workhorse in environmental, labor, public, urban, and other parts of economics. Although consistent estimation of the HPS cannot be used for counterfactual policy analysis of nonmarginal changes, the decades of empirical research that have been guided by Rosen's paper demonstrate that there are many marginal changes worth analyzing, and the last 10–15 years illustrate that it is possible to produce credible evidence on their welfare consequences.

With respect to nonmarginal changes, the picture is not quite as bright when it comes to using the Rosen method. However, there are already some promising approaches that merit greater investigation, application, and exploration. If history is any guide, the coming years will see the development of new methods that build on Rosen's method to recover bid functions as the questions that can be answered remain vital and urgent.

Although consistent estimation of the HPS does not recover the underlying bid functions, it can be used to estimate the welfare impacts of nonmarginal changes of past changes in amenities. Specifically, Greenstone and Gallagher (2008) demonstrate that knowledge of the HPS can be used to infer the historical welfare consequences of a nonmarginal

[3] Heckman, Matzkin, and Nesheim (2005, 2010) examine identification and estimation of nonadditive hedonic models and the performance of estimation techniques for additive and nonadditive models.

amenity change for landowners in a particular location (e.g., a neighborhood, city, or county) before the change occurs.[4] In their empirical application, Greenstone and Gallagher use the HPS to learn about the welfare consequences of Superfund clean-ups of industrial sites, but the historical record is filled with countless other consequential nonmarginal amenity changes. Use of the HPS in this manner has potentially tremendous practical value, because the benefits of previous policies are an important guide about the benefits of future or alternative policies; this is especially so in the all too frequent case in which reliable estimates of the underlying demand primitives or bid functions are unavailable.

Perhaps most notably, more than four decades after its publication, Rosen's model is being used to generate insights about the world, and researchers are still actively engaged with its methods. It is evident that Rosen's hedonic method has fundamentally altered our understanding of the world and in so doing has passed the acid test of time.

References

Bajari, P., and C. Benkard. 2005. "Demand Estimation with Heterogeneous Consumers and Unobserved Product Characteristics: A Hedonic Approach." *J.P.E.* 113 (6): 1239–76.

Bartik, Alexander W., Janet Currie, Michael Greenstone, and Christopher R. Knittel. 2018. "The Local Economic and Welfare Consequences of Hydraulic Fracturing." Manuscript, Soc. Sci. Res. Network. https://ssrn.com/abstract=2692197.

Bartik, T. J. 1987. "The Estimation of Demand Parameters in Hedonic Price Models." *J.P.E.* 95 (1): 81–88.

Bayer, P., F. Ferreira, and R. McMillan. 2007. "A Unified Framework for Measuring Preferences for Schools and Neighborhoods." *J.P.E.* 115 (4): 588–638.

Black, D. A., and T. J. Kniesner. 2003. "On the Measurement of Job Risk in Hedonic Wage Models." *J. Risk and Uncertainty* 27 (3): 205–20.

Black, S. E. 1999. "Do Better Schools Matter? Parental Valuation of Elementary Education." *Q.J.E.* 114 (2): 577–99.

Brown, C. 1980. "Equalizing Differences in the Labor Market." *Q.J.E.* 94 (1): 113–34.

Brown, J. N., and H. S. Rosen. 1982. "On the Estimation of Structural Hedonic Price Models." *Econometrica* 50 (3): 765–68.

Chay, K., and M. Greenstone. 2005. "Does Air Quality Matter? Evidence from the Housing Market." *J.P.E.* 113 (2): 376–424.

Currie, J., L. Davis, M. Greenstone, and R. Walker. 2015. "Environmental Health Risks and Housing Values: Evidence from 1,600 Toxic Plant Openings and Closings." *A.E.R.* 105 (2): 678–709.

[4] This is a key population in their own right. Further, the welfare impacts on them are equal to the social welfare impacts with the admittedly strong assumption of zero moving costs, as well as no change in the overall price level. See Roback (1982) and more recent papers (e.g., Moretti 2011; Bartik et al. 2018) that consider the cases of nonzero moving costs and elastic housing supply.

Davis, L. W. 2004. "The Effect of Health Risk on Housing Values: Evidence from a Cancer Cluster." *A.E.R.* 94 (5): 1693–1704.

Deschenes, O., and M. Greenstone. 2007. "The Economic Impacts of Climate Change: Evidence from Agricultural Output and Random Fluctuations in Weather." *A.E.R.* 97 (1): 354–85.

Ekeland, I., J. Heckman, and L. Nesheim. 2004. "Identification and Estimation of Hedonic Models." *J.P.E.* 112 (S1): S60–S109.

Epple, D. 1987. "Hedonic Prices and Implicit Markets: Estimating Demand and Supply Functions for Differentiated Products." *J.P.E.* 95 (1): 59–80.

Epple, D., and H. Sieg. 1999. "Estimating Equilibrium Models of Local Jurisdictions." *J.P.E.* 107 (4): 645–81.

Greenstone, M., and J. Gallagher. 2008. "Does Hazardous Waste Matter? Evidence from the Housing Market and the Superfund Program." *Q.J.E.* 123 (3): 951–1003.

Greenstone, M., R. Hornbeck, and E. Moretti. 2010. "Identifying Agglomeration Spillovers: Evidence from Winners and Losers of Large Plant Openings." *J.P.E.* 118 (3): 536–98.

Gruber, J. 1994. "The Incidence of Mandated Maternity Benefits." *A.E.R.* 84 (3): 622–41.

Heckman, J. J., R. L. Matzkin, and L. P. Nesheim. 2005. "Simulation and Estimation of Hedonic Models." In *Frontiers in Applied General Equilibrium Modeling*, edited by T. J. Kehoe, T. N. Srinivasan, and J. Whalley. New York: Cambridge Univ. Press.

———. 2010. "Nonparametric Identification and Estimation of Nonadditive Hedonic Models." *Econometrica* 78 (5): 1569–91.

Linden, L., and J. E. Rockoff. 2008. "Estimates of the Impact of Crime Risk on Property Values from Megan's Laws." *A.E.R.* 98 (3): 1103–27.

Moretti, Enrico. 2011. "Local Labor Markets." In *Handbook of Labor Economics*, vol. 4b, edited by Orley Ashenfelter and David Card, 1237–1312. Amsterdam: Elsevier.

Palmquist, R. B. 1984. "Estimating the Demand for the Characteristics of Housing." *Rev. Econ. and Statis.* 66 (3): 394–404.

Roback, J. 1982. "Wages, Rents, and the Quality of Life." *J.P.E.* 90 (6): 1257–78.

Rosen, S. 1974. "Hedonic Prices and Implicit Markets: Product Differentiation in Pure Competition." *J.P.E.* 82 (1): 34–55.

Small, K. A. 1975. "Air Pollution and Property Values: Further Comment." *Rev. Econ. and Statis.* 57 (1): 105–7.

Smith, R. S. 1979. "Compensating Wage Differentials and Public Policy: A Review." *Indus. and Labor Relations Rev.* 32 (3): 339–52.

Smith, V. K., and J. Huang. 1995. "Can Markets Value Air Quality? A Meta-Analysis of Hedonic Property Value Models." *J.P.E.* 103 (1): 209–27.

Assignment Problems

Philip J. Reny

University of Chicago

I. Introduction

An assignment problem is one in which a number of goods, each in some fixed quantity, must be assigned to a number of individuals. The class of assignment problems that will concern us here are those in which no monetary transfers are possible.[1] Assigning committee positions to members of Congress or dormitories to students are but two of many such examples.

When the individuals' tastes are known, it is not difficult in principle to achieve an assignment of goods to individuals that is Pareto efficient. But this becomes considerably more difficult when preferences are private information because one must then ensure that no individual has any incentive to misreport his or her preferences.

In a seminal *JPE* paper, Hylland and Zeckhauser (1979) consider assignment problems in which each individual can receive at most one good and at most one unit of it (as in the two examples above). They showed that if individuals are endowed with fiat money and participate in a market that sets nominal prices for the probabilities with which goods can be obtained, then competitive equilibrium prices (for probabilities) exist and yield ex ante efficient lotteries that can be resolved to produce ex post efficient outcomes.[2] Consequently, when there are sufficiently many individuals so that no single individual has any significant impact on prices, each individual would be willing to report his preferences truthfully in a mechanism that computes and implements the competitive equilibrium outcome for those preferences. Such is the mechanism proposed by Hylland and Zeckhauser.

An even more challenging class of assignment problems are the so-called *combinatorial* assignment problems. In such a problem, there are

I thank Eric Budish for helpful comments and the National Science Foundation (SES-1227506 and SES-1724747) for financial support.

[1] In contrast, e.g., to Koopmans and Beckman (1957).

[2] That the difficulties created by indivisibilities might be circumvented by introducing probabilities was first recognized by von Neumann (1953), whose work influenced Koopmans and Beckman (1957), who interpret probabilities as fractions of perfectly divisible surrogate goods. Birkhoff's (1946) theorem, that doubly stochastic matrices are convex combinations of permutation matrices, is the central mathematical result that is at the heart of this approach.

again many units of many goods to be allocated, but there are no a priori restrictions on the bundles of goods that individuals can receive. Especially challenging are cases in which individual preferences over bundles of goods exhibit complementarities.

Recently, an important combinatorial assignment problem has been considered by Budish (2011). He considers the challenging problem of assigning classes to students, a problem in which complementarities arise naturally from course scheduling constraints even if student preferences over classes, without those constraints, are additively separable.[3]

As in the Hylland-Zeckhauser model, the goods in Budish's (2011) model, namely classes, are indivisible. Hylland and Zeckhauser (1979) circumvent the indivisibility problem by creating a market for probabilities. Unfortunately, as Budish observes, in the presence of complementarities there may be no prices for the probabilities with which individual classes can be obtained that lead students to choose lotteries over bundles of classes that efficiently exhaust the total available probability and that are feasible to carry out. In short, the combinatorial assignment problem cannot, in general, be solved by using the lottery technique of Hylland and Zeckhauser. One must deal with the indivisibilities and complementarities head on.[4]

Like Hylland and Zeckhauser (1979), Budish (2011) uses a market mechanism with fiat money to attack the problem. Students are given "income" in the form of fiat money that can be used to purchase classes at prices that are determined in market equilibrium. Importantly, Budish allows the students' preferences to be almost completely general and does not assume that the fiat money has any intrinsic value to them (unlike other mechanisms in use for this problem).

Because issues such as "fairness" are particularly important in this and other assignment contexts, it would be natural for each student to receive the same amount of fiat money. However, as Budish (2011) shows, this can lead to the nonexistence of the type of market equilibrium that he considers. Budish's striking result is that, with arbitrarily small departures from equal relative incomes, existence is restored, and several attractive efficiency and fairness criteria can be obtained. Furthermore, with large numbers of individuals, Budish's market mechanism is approximately incentive compatible.

The objective in this short note marking the 125th anniversary of the *Journal of Political Economy* is modest. It is shown here that Budish's (2011) result can be generalized to allow arbitrary preferences and both

[3] I am grateful to Eric Budish for pointing this out.

[4] But see Budish et al. (2013) for particular conditions under which the lottery technique can be made to work.

divisible and indivisible goods, as would exist, for example, in the committee assignment problem when the workload on some committees can be divided up in any way among committee members.[5] Thus, while the absence of divisible goods is important in Budish's proof, it is inessential for his results.

On the technical side, the proof offered here is rather simple. Indivisibilities create discontinuities in demand as prices vary, because strictly preferred bundles can suddenly become affordable. These discontinuities are the source of most of the complications that arise in Budish's (2011) clever proof. The main technical contribution here is to note that one can avoid discontinuities altogether by considering a surrogate economy in which agents, instead of maximizing utility subject to their budget constraint, maximize a Lagrangian in which violations of their budget constraint yield a suitably high utility cost per unit of overexpenditure. Equilibria of this surrogate economy are shown to yield equilibria in the sense of Budish's paper. It is entirely possible that this surrogate economy, because it is continuous, might lead to more efficient and/or stable algorithms for computing the requisite equilibria. But these computational issues have not been explored here in any detail whatsoever.[6]

II. Assignment Problems

An *assignment problem*, (I, L, X, u, ω), consists of the following items:

1. I and L are positive integers, where I is the number of agents and L is the number of commodities;
2. $X = \times_{i=1}^{I} X_i$, where each *consumption set* X_i is a compact set of nonnegative vectors in $\mathbb{R}^L$ with $0 \in X_i$;
3. ω is an *aggregate endowment vector* in $\mathbb{R}^L$ such that $\omega_l > 0$ for every $l = 1, \ldots, L$;
4. $u = (u_1, \ldots, u_I)$, where each *utility function* $u_i : X_i \to [0, 1]$ is continuous.[7]

[5] Budish (2011) allows arbitrary preferences except for the assumption that no agent is indifferent between any two bundles in his finite consumption set. Because divisible goods are allowed here, my consumption sets can be uncountably infinite. Therefore, to accommodate continuous preferences, indifference must be permitted, and this is done whether consumption sets are finite or infinite.

[6] The proof technique employed here can also provide a generalization of the results of Dierker (1971) to include both indivisible and divisible goods, while at the same time simplifying his proof.

[7] All the results can be derived under the slightly more general assumption that for each consumer i there is a reflexive and transitive (but possibly incomplete) binary relation, $\succsim_i$, on X_i that is continuous; i.e., for every $x_i \in X_i$, the sets $\{y_i \in X_i : y_i \succsim_i x_i\}$ and $\{y_i \in X_i : x_i \succsim_i y_i\}$ are closed.

REMARK 1. Consumer i's consumption set X_i need not be convex or even connected. In particular, an assignment problem here can accommodate the simultaneous presence of divisible and indivisible goods.

For any $p \in [0, \infty)^L$, for any $c_i > 0$, and for any agent i, let $D_i(p, c_i) \subseteq X_i$ be the set of solutions to the maximization problem,

$$\max_{x_i \in X_i} u_i(x_i) \quad \text{subject to } px_i \leq c_i.$$

Let $\|\cdot\|$ denote the Euclidean norm. For any $\varepsilon > 0$ and for any $c_1, \ldots, c_I > 0$, let $c = (c_1, \ldots, c_I)$ and define

$$\delta_\varepsilon(p, c) := \sup \| x_i - y_i \|,$$

where the supremum is over all agents $i \in \{1, \ldots, I\}$ and all $x_i, y_i \in \cup_{\varepsilon' \in [0,\varepsilon]} D_i(p, c_i + \varepsilon')$.

Throughout the paper, for any $l \in \{1, \ldots, L\}$, $e_l = (0, \ldots, 0, 1, 0, \ldots, 0)$ denotes the lth unit vector in $\mathbb{R}^L$.

The main result below replicates the main result in Budish (2011), but does not require the sets X_i to be finite, requiring instead only that they be compact. Also included is the minor modification that the positive incomes $c_1, \ldots, c_I$ of the agents can be arbitrary, whereas Budish focuses on the most central case in which $c_1 = \cdots = c_I = 1$.[8]

THEOREM 1. For any assignment problem (I, L, X, u, ω), for any $\varepsilon > 0$ and for any positive real numbers $c_1, \ldots, c_I$, there exist $p^* \in [0, \infty)^L$ and $x^* \in X$ such that

a. for every agent i,

i. $p^* x_i^* \leq c_i + \varepsilon$, and
ii. x_i^* solves $\max_{y_i \in X_i} u_i(y_i)$ subject to $p^* y_i \leq \max(p^* x_i^*, c_i)$;

b. $\| z^* \| \leq \delta_\varepsilon(p^*, c)\sqrt{L}/2$, where $c = (c_1, \ldots, c_I)$ and where for each $l = 1, \ldots, L$,

$$z_l^* := \begin{cases} \sum_i x_{il}^* - \omega_l, & \text{if } p_l^* > 0 \\ \max\left(\sum_i x_{il}^* - \omega_l, 0\right) & \text{if } p_l^* = 0; \end{cases}$$

and

c. $p^* \omega \leq \Sigma_i (c_i + \varepsilon)$.

[8] The present result is not an exact replication of the result in Budish (2011) because the bound on the market-clearing error here (specifically, the coefficient of $\sqrt{L}/2$ in part *b* of theorem 1) can be smaller or larger than the bound that he describes in his n. 15. However, the important feature of both bounds is that they are of the same order of magnitude and, most importantly, that they are independent of the number of agents, I.

REMARK 2. To obtain the result in Budish (2011), restrict each X_i to be a finite set of nonnegative vectors with integer coordinates, set $c_1 = \cdots = c_I = 1$ above, and define his $b_i^* := \max(p^* x_i^*, 1)$ for every $i = 1, \ldots, I$. Note that theorem 1 cannot be obtained from Budish's result by discretizing the compact X_i with a finite grid and then taking the limit as the grid size shrinks to zero. Such a limiting argument can ensure in part a(ii) only that x_i^* solves $\max_{y_i \in X_i} u_i(y_i)$ subject to $p^* y_i < \max(p^* x_i^*, c_i)$, because strictly better bundles that are unaffordable along the sequence might become exactly affordable at the limit.

REMARK 3. Endowing the agents with even slightly different amounts of fiat money, instead of with real goods, and allowing the vector of goods prices to be any nonnegative vector, including the zero vector, means that differences in real incomes can be arbitrarily large and will be determined by the goods prices in equilibrium.[9] This can be advantageous since efficiency might sometimes require such real income differences when preferences are not strictly monotone (e.g., as in the problem of assigning classes to students).

Budish (2011) introduces two fairness-related concepts that can be usefully applied to problems that include indivisible goods.[10] The first of these is an agent's "maxmin utility."[11] Before defining this, first define, for any positive integer n, agent i's *n-maxmin utility* to be the utility number $\max \min(u_i(y_1), \ldots, u_i(y_n))$, where the maximum is over all $y_1, \ldots, y_n \in X_i$ such that $\sum_{j=1}^{n} y_j \leq \omega$. Then, define agent i's maxmin utility to be his I-maxmin utility. Maxmin utility is one way to generalize the "I cut you choose" method of fair division to many agents. Budish's second fairness concept presumes that X_i contains only vectors with integer coordinates and is as follows. An allocation $x \in X$ is *envy-free up to a single unit of a single good* iff for every pair of agents i and j, if $u_i(x_j) > u_i(x_i)$, that is, if i envies j, then there is a commodity l such that either $u_i(x_i) \geq u_i(x_j - e_l)$ or $x_j - e_l \notin X_i$.

Budish (2011) shows that the allocations that he obtains are envy-free up to a single unit of a single good and that, while they might not yield each agent his maxmin utility, they do yield each agent his $(I + 1)$-maxmin utility. Budish also shows that his allocations cannot be Pareto improved on if the agents are allowed to trade among themselves after the assignment is made. Budish does not rule out the possibility that if there is ex-

[9] Relative incomes can be made arbitrarily similar. With $c_1 = \cdots = c_I = 1$, the ratio of any pair of the incomes $b_1^*, \ldots, b_I^*$ defined in the previous remark is between 1 and $1 + \varepsilon$.

[10] Indivisibilities can lead to nonexistence of efficient and envy-free allocations. See Budish (2011).

[11] Budish (2011) uses the term maxmin *share* since he focuses on the bundle that achieves the maxmin utility. I find it more convenient to define the utility level, even though this number obviously depends on the particular utility representation.

cess supply, then that excess supply could be used to achieve a Pareto improvement.

The remarks to follow, in part, describe the sense in which Budish's important fairness and efficiency results can be maintained and sometimes slightly improved on. In these remarks, p^* and x^* are as in theorem 1.

REMARK 4 (η-envy-free). If $c_1 = \cdots = c_I = 1$, then for any $\eta > 0$, there is $\varepsilon > 0$ sufficiently small so that the allocation x^* is "η-envy-free" in the following sense. For any agents i and j, if $x_j^* \in X_i$ and $u_i(x_j^*) > u_i(x_i^*)$, that is, if i envies j, then there is a commodity l such that $u_i(x_i^*) \geq u_i(y_i)$ for every $y_i \in X_i$ such that $y_i \leq x_j^* - \eta e_l$.[12] In particular, if $X_i = \{0, 1, \ldots, k\}^L$ as in Budish (2011), then we can choose $y_i = x_j^* - e_l$ and x^* is 1-envy-free, that is, envy-free up to a single unit of a single good.

REMARK 5 ($(I+1)$–maxmin utility). If $c_1 = \cdots = c_I = 1$ and $\varepsilon < 1/I$, then $u_i(x_i^*) \geq \max(\min(u_i(y_1), u_i(y_2), \ldots, u_i(y_{I+1})))$ for each agent i, where the maximum is over all $y_1, \ldots, y_{I+1} \in X_i$ such that $\Sigma_{j=1}^{I+1} y_j \leq \omega$. That is, x_i^* is at least as good for i as his $(I+1)$–maxmin share (Budish 2011).[13]

REMARK 6 (Weak stability and efficiency). There do not exist $S \subseteq \{1, \ldots, I\}$ and $\hat{x} \in X$ such that $\hat{x}_i \neq x_i^*$ for at least one $i \in S$, $\Sigma_{i \in S} \hat{x}_{il} \leq \Sigma_{i \in S} x_{il}^*$ for every l with $p_l^* > 0$, and $u_i(\hat{x}_i) > u_i(x_i^*)$ for every $i \in S$ such that $\hat{x}_i \neq x_i^*$.[14] In particular, setting $S = \{1, \ldots, I\}$ shows that x^* is weakly Pareto efficient in this economy for any aggregate endowment vector ω^* that satisfies $\omega_l^* = \Sigma_i x_{il}^*$ if $p_l^* > 0$, and $\omega_l^* \geq \Sigma_i x_{il}^*$ if $p_l^* = 0$. If, as in Budish (2011), all X_i are finite and preferences exhibit no indifference, then weak stability and efficiency are equivalent to standard stability and efficiency, wherein blocking requires only one individual in the coalition to be made strictly better off.

The next two remarks indicate that the efficiency and maxmin results described in the previous remarks can be improved on with an arbitrarily small degradation of the bound in part *b* of theorem 1. Fix any arbitrarily small $\eta > 0$. Choose $\bar{\varepsilon} > 0$ small enough so that $\bar{\varepsilon} I \sqrt{L} < (\Sigma_{l=1}^{L} p_l)\eta$ for every

[12] Given $\eta > 0$, choose $\varepsilon > 0$ so that $\max_l p_l > \varepsilon/\eta$ for every price vector $p \in [0, \infty)^L$ such that $px_j \geq 1$ for some consumer j and some $x_j \in X_j$. Such an ε exists by the compactness of the consumption sets. With this choice of ε, suppose that $u_i(x_j^*) > u_i(x_i^*)$. Then $p^* x_j^* > 1$, and so there is l such that $p_l^* \eta > \varepsilon$. Hence, because also $p^* x_j^* \leq 1 + \varepsilon$ by part *a*(i), we have $p^*(x_j^* - \eta e_l) < 1$. So any $y_i \leq x_j^* - \eta e_l$ that is in X_i satisfies $p^* y_i < 1 = c_i$, and so $u_i(x_i^*) \geq u_i(y_i)$ by part *a*(ii).

[13] Otherwise, if $y_1, \ldots, y_{I+1}$ is a solution to the maxmin problem, then $u_i(y_j) > u_i(x_i^*)$ for every j. But then by part *a*(ii), $p^* y_j > 1$ for every j, and so $p^* \Sigma_{j=1}^{I+1} y_j > I + 1 > (1 + \varepsilon) I \geq p^* \omega$, where the final inequality is by part *c*. But the outer strict inequality contradicts the feasibility of $y_1, \ldots, y_{I+1}$ for the maxmin problem.

[14] If such an $\hat{x}$ were to exist, then part *a*(ii) implies that $p^* \hat{x}_i > p^* x_i^*$ for every $i \in S$ such that $\hat{x}_i \neq x_i^*$. Since there is at least one $i \in S$ such that $\hat{x}_i \neq x_i^*$, $p^* \Sigma_{i \in S} \hat{x}_i > p^* \Sigma_{i \in S} x_i^*$. Hence, there is an l such that $p_l^* > 0$ and $\Sigma_{i \in S} \hat{x}_{il} > \Sigma_{i \in S} x_{il}^*$, contradicting the feasibility of $\hat{x}$ for the coalition S.

price vector $p \in [0,\infty)^L$ satisfying $\max_{x\in X} p\Sigma_i x_i \geq I$. Such an $\bar{\varepsilon} > 0$ exists by the compactness of X.

REMARK 7 (Maxmin utility). If $c_1 = \cdots = c_I = 1$, if $\varepsilon < \bar{\varepsilon}$, and if the bound in part *b* is weakened to $\| z^* \| \leq \eta + \delta_\varepsilon(p^*, c)\sqrt{L}/2$, then we can ensure that $u_i(x_i^*) \geq \max(\min(u_i(y_1), u_i(y_2), \ldots, u_i(y_I))$, where the maximum is over all $y_1, \ldots, y_I \in X_i$ such that $\Sigma_{j=1}^I y_j \leq \omega$. That is, x_i^* yields agent i at least his maxmin utility.[15]

REMARK 8 (Weak Pareto efficiency). If the bound in part *b* is weakened to $\| z^* \| \leq \eta + \delta_\varepsilon(p^*, c)\sqrt{L}/2$, then we can ensure that there does not exist $\hat{x} \in X$ distinct from x^* such that $\Sigma_{i=1}^I \hat{x}_{il} \leq \omega_l$ for every l, and $u_i(\hat{x}_i) > u_i(x_i^*)$ for every i such that $\hat{x}_i \neq x_i^*$.[16]

REMARK 9. The relevance of the efficiency and fairness results in the previous remarks is called into question by the possibility that x^* might not be feasible. This difficulty can be mitigated as follows. Define $\bar{\delta}_\varepsilon(c) = \sup \delta_\varepsilon(p, c)$, where the supremum is over all $p \in [0,\infty)^L$.[17] If $\omega_l \geq \bar{\delta}_\varepsilon(c)\sqrt{L}/2$ for every l, then an application of theorem 1 using the endowment vector $\tilde{\omega} = \omega - (\bar{\delta}_\varepsilon(c)\sqrt{L}/2)\mathbf{1}$ instead of ω yields p^* and x^* satisfying part *a* of theorem 1 and, by part *b*, satisfying $\Sigma_i x_{il}^* \leq \omega_l$ for every l and $\omega_l - \bar{\delta}_\varepsilon(c)\sqrt{L} \leq \Sigma_i x_{il}^* \leq \omega_l$ for every l with $p_l^* > 0$, and, by part *c*, satisfying $p^*(\omega - (\bar{\delta}_\varepsilon(c)\sqrt{L}/2)\mathbf{1}) \leq \Sigma_i(c_i + \varepsilon)$. All of the efficiency and fairness results in the remarks above then go through, but with respect to the economy with aggregate endowment $\tilde{\omega}$ instead of ω. In particular, x^* is feasible, η-envy-free, and weakly stable (because the η-envy-free and weak stability properties do not depend on the aggregate endowment). However, while x^* is weakly Pareto efficient using only the aggregate endowment $\tilde{\omega}$, x^* need not be Pareto efficient for the actual economy with aggregate endowment ω. A second way to handle infeasibility is to note that, except for at most L agents, every agent i can actually receive his bundle x_i^*.[18] The L exceptional agents can then be assigned any bundles that are fea-

[15] Let $y_1, \ldots, y_{I+1}$ be a solution to the maxmin problem. Replace ω with $\omega + (\eta/\sqrt{L})\mathbf{1}$ in the statement of theorem 1. Now suppose, by way of contradiction, that $u_i(y_j) > u_i(x_i^*)$ for every j. Then by part *a*(ii), $p^* y_j > 1$ for every j. Therefore, $p^*\Sigma_{j=1}^I y_j \geq p^*(\omega + (\eta/\sqrt{L})\mathbf{1}) - \varepsilon I$, where the final inequality is by part *c*. But $p^*\Sigma_{j=1}^I y_j > I$ implies by our choice of ε that $p^*((\eta/\sqrt{L})\mathbf{1}) - \varepsilon I > 0$ and so $p^*\Sigma_{j=1}^I y_j > p^*\omega$, contradicting the feasibility of $y_1, \ldots, y_I$ for the maxmin problem.

[16] Suppose that such an $\hat{x}$ exists. Replace ω with $\omega + (\eta/\sqrt{L})\mathbf{1}$ in the statement of theorem 1. Then for every i, either $u_i(\hat{x}_i) > u_i(x_i^*)$ or $\hat{x}_i = x_i^*$, and so part *a*(ii) implies that $p^*\hat{x}_i \geq 1$ for every i with strict inequality whenever $\hat{x}_i \neq x_i^*$. Therefore, since $\hat{x} \neq x^*$, $p^*\Sigma_{i=1}^I \hat{x}_i > I \geq p^*(\omega + (\eta/\sqrt{L})\mathbf{1}) - \varepsilon I$, where the second inequality is by part *c*. But our choice of $\bar{\varepsilon}$ implies that $p^*((\eta/\sqrt{L})\mathbf{1}) - \varepsilon I > 0$ and so $p^*\Sigma_{i=1}^I \hat{x}_i > p^*\omega$, contradicting the feasibility of $\hat{x}$.

[17] The function $\bar{\delta}_\varepsilon(c)$ exists and is finite because $\delta_\varepsilon(p, c)$ is bounded above by $\max \| x_i - y_i \|$, where the maximum is over all i and all $x_i, y_i \in X_i$.

[18] Observe that $j \leq L$ in the proof below because $\Sigma_{i=1}^j (\#S_i - 1) \leq L$ and, in the summand, each $(\#S_i - 1) \geq 1$. Hence, $\omega \geq y^* = y_1^* + \cdots + y_j^* + x_{j+1}^* + \cdots + x_I^*$ for some y_i^* in the convex hull of S_i, $i = 1, \ldots, j$.

sible given the remaining goods. It then follows as in remarks 6 and 7 that all but the L exceptional agents receive at least their maxmin utility and that no coalition involving only agents outside the exceptional set can trade among themselves following the assignment so that each is better off.

REMARK 10. For any closed subset S of $\mathbb{R}^L$ we may follow Starr (1969) and define

$$\rho(S) = \sup_{y \in co S} \inf_{A \subseteq S; y \in co A} \sup_{a, a' \in A} \| a - a' \|.$$

The function ρ is a measure of the convexity of the set S, with $\rho(S) = 0$ if and only if S is convex. Defining $S_i(p, c_i, \varepsilon) := \cup_{\varepsilon' \in [0, \varepsilon]} D_i(p, c_i + \varepsilon')$, the proof shows that the bound in part b can be reduced to $\sup_i \rho(S_i(p^*, c_i, \varepsilon))\sqrt{L}/2$.

REMARK 11. If there is a continuum of each type of agent $i = 1, 2, \ldots, I$, then no single agent would have an incentive to misreport his utility function because he could not affect the price or the bundle that he receives. Furthermore, the per capita market-clearing error, being bounded above by the bound in part b divided by the (infinite) number of agents, is zero. Hence, as in Hylland and Zeckhauser (1979) and Budish (2011), the solution here can be the basis for an ε-incentive-compatible mechanism with small per capita market-clearing error when aggregate endowments are proportional to the large but finite number of agents.

III. Proof of Theorem 1

Fix any $\varepsilon > 0$. Because every coordinate of ω is strictly positive, we may choose a positive real number B such that

$$B > \sum_{i=1}^{I} (c_i + \varepsilon)/\omega_l \quad \text{for every } l = 1, \ldots, L. \tag{1}$$

For any $\alpha \in \mathbb{R}$, let $\alpha^+ = \max(\alpha, 0)$.

Define an $I + 1$–person game between players $i = 0, 1, \ldots, I$ as follows. Player $i = 0$'s set of pure strategies is $[0, B]^L$. For $i \in \{1, \ldots, I\}$, player i's set of pure strategies is X_i. For any $(p, x) \in [0, B]^L \times X$, the players' payoffs are as follows:

$$U_i(p, x) := \varepsilon u_i(x_i) - (p x_i - c_i)^+ \quad \text{for } i = 1, \ldots, I$$

and

$$U_0(p, x) := p\left(\sum_i x_i - \omega\right).$$

All of the payoff functions $U_i : [0, B]^L \times X \to \mathbb{R}$ are continuous.

Let us allow each player $i > 0$ to use a mixed strategy, that is, a probability measure, m_i on X_i. For $i > 0$, let M_i denote player i's space of mixed strategies, henceforth simply strategies, and let $M = \times_{i=1}^{I} M_i$. Each M_i is endowed with the weak* topology. Like $[0, B]^L$, each M_i is nonempty, compact, and convex.

For any $m = (m_1, \ldots, m_I) \in M$, let $\bar{m}$ denote the product measure $m_1 \times \cdots \times m_I$.

Extend the players' payoff functions to $[0, B]^L \times M$, by an expected utility calculation. Then, all of the payoff functions $U_i : [0, B]^L \times M \to \mathbb{R}$ are continuous. Since each player's payoff function is quasi-concave (linear in fact) in his own strategy for any fixed strategies of the others, the players' best reply correspondences satisfy all of the hypotheses of Glicksberg's (1952) fixed-point theorem. Hence, this game possesses a Nash equilibrium (p^*, m^*).

For each player $i > 0$, equilibrium implies that the support of m_i^* contains only elements x_i of X_i that maximize $U_i(p^*, x_i)$.[19] Hence, if x_i is in the support of m_i^*, then x_i solves

$$\max_{y_i \in X_i} (\varepsilon u_i(y_i) - (p^* y_i - c_i)^+). \tag{2}$$

Since u_i is nonnegative and $0 \in X_i$, the maximum value in (2) is nonnegative. Consequently, since u_i is bounded above by 1, it must be the case that

$$p^* x_i \leq c_i + \varepsilon \quad \text{for every } x_i \text{ in the support of } m_i^*. \tag{3}$$

Furthermore, any solution x_i to the maximization problem (2), and hence any x_i in the support of m_i^*, must solve

$$\max_{y_i \in X_i} u_i(y_i) \quad \text{subject to } p^* y_i \leq \max(p^* x_i, c_i). \tag{4}$$

To see this claim, let x_i solve (2). Then for every $y_i \in X_i$,

$$\begin{aligned} \varepsilon u_i(x_i) &\geq \varepsilon u_i(y_i) - (p^* y_i - c_i)^+ + (p^* x_i - c_i)^+ \\ &\geq \varepsilon u_i(y_i) \quad \text{if } p^* y_i \leq \max(p^* x_i, c_i), \end{aligned}$$

as claimed.

Define

$$y^* := \int_X \sum_i x_i \bar{m}^*(dx).$$

[19] Recall that the support of a probability measure in a separable metric space is the smallest closed subset having probability one.

Then player 0's equilibrium payoff, $U_0(p^*, m^*)$, satisfies

$$U_0(p^*, m^*) = p^*(y^* - \omega) \geq p(y^* - \omega) \quad \forall p \in [0, B]^L.$$

Consequently, for any $l \in \{1, \ldots, L\}$, $y_l^* < \omega_l \Rightarrow p_l^* = 0$ and $y_l^* > \omega_l \Rightarrow p_l^* = B$. But then $y_l^* > \omega_l$ is impossible since $p_l^* = B$ implies, by (3), that $y_{il}^* \leq (c_i + \varepsilon)/B$ for every agent i and so $y_l^* \leq \Sigma_i(c_i + \varepsilon)/B < \omega_l$ by (1). We may conclude that

$$y_l^* \leq \omega_l \quad \text{for every } l = 1, \ldots, L \tag{5}$$

and that $p_l^* = 0$ for any l for which the inequality is strict.

For each $i = 1, \ldots, I$, let e_i denote the ith unit vector $(0, \ldots, 0, 1, 0, \ldots, 0)$. Then

$$\int_X \left[\frac{1}{I}\sum_{i=1}^{I}(e_i, x_i)\right] \bar{m}^*(dx) = \frac{1}{I}(1, \ldots, 1, y^*) \in \Delta_I \times \mathbb{R}^L, \tag{6}$$

where (e_i, x_i) denotes the concatenation of e_i and x_i, and Δ_I denotes the $I-1$–dimensional unit simplex.

The equality in (6) says that $(1, \ldots, 1, y^*)/I$ is in the convex hull of the closed subset C of $\Delta_I \times \mathbb{R}^L$ that consists of all points of the form (e_i, x_i), where x_i is in the support of m_i^* for each i. By Caratheodory's theorem (Rockafellar 1970), $(1, \ldots, 1, y^*)/I$ can therefore be written as a convex combination of $I + L$ or fewer points belonging to C. Thus, for some positive integer K we may write

$$\frac{1}{I}(1, \ldots, 1, y^*) = \sum_{i=1}^{I}\sum_{k=1}^{K}\lambda_{ik}(e_i, x_i^k), \tag{7}$$

where the λ_{ik}'s are nonnegative and sum to one, and at most $I + L$ of the λ_{ik} are strictly positive and $\lambda_{ik} > 0$ implies that x_i^k is in the support of m_i^*.

For each $i = 1, \ldots, I$, let $S_i = \{x_i^k : \lambda_{ik} > 0\}$. Since the first I coordinates of the vector on the left-hand side of (7) are positive, each S_i contains at least one element. Reindexing if necessary, let $S_1, \ldots, S_j$ denote those S_i that contain two or more elements. So $S_{j+1}, \ldots, S_I$ are singletons, and since at most $I + L$ of the λ_{ik} are strictly positive, the union of $S_1, \ldots, S_j$ contains no more than $L + j$ elements.

For every $i = 1, \ldots, j$, every x_i in S_i is in the support of m_i^* and therefore satisfies (3) and solves (4). In particular, for any $x_i \in S_i$ letting $\varepsilon' = (p^* x_i - c_i)^+$, we have $\varepsilon' \in [0, \varepsilon]$ and $x_i \in D_i(p^*, c_i + \varepsilon')$. Consequently, the distance between any two points in S_i is no greater than $\delta_\varepsilon(p^*, c)$, where $c = (c_1, \ldots, c_I)$. Therefore, the distance between any point in S_i and the simple average of all of the points in S_i is no greater than $\delta_\varepsilon(p^*, c)(\#S_i - 1)/(\#S_i)$.

The equality in (7) for the first I coordinates implies that $\Sigma_{k=1}^{K} I\lambda_{ik} = 1$ for each i, and the equality for the last L coordinates then implies that y^* is contained in the sum of the convex hulls of the sets $S_1, \ldots, S_I$. Hence, y^* is contained in the convex hull of $S_1 + \cdots + S_I$.[20] Consequently, by the Shapley-Folkman theorem (see Starr 1969) we may select $x_i^* \in S_i$ for each $i = 1, \ldots, I$, so that

$$\left\| y^* - \sum_{i=1}^{I} x_i^* \right\|^2 \leq \sum_{i=1}^{j} \left[\frac{(\#S_i - 1)\delta_\varepsilon(p^*, c)}{\#S_i} \right]^2.$$

Since $\#S_i \geq 2$ for every $i = 1, \ldots, j$, we have $[(\#S_i - 1)/(\#S_i)]^2 \leq (\#S_i - 1)/4$ for every $i = 1, \ldots, j$. Hence,

$$\left\| y^* - \sum_{i=1}^{I} x_i^* \right\|^2 \leq \sum_{i=1}^{J} \frac{(\#S_i - 1)\delta_\varepsilon^2(p^*, c)}{4} \leq \delta_\varepsilon^2(p^*, c)L/4,$$

where the second inequality follows because the union of the sets $S_1, \ldots, S_j$ contains no more than $L + j$ elements, and so $\Sigma_{i=1}^{j}(\#S_i - 1) \leq L$. Hence, we may conclude that

$$\left\| y^* - \sum_{i=1}^{I} x_i^* \right\| \leq \delta_\varepsilon(p^*, c)\sqrt{L}/2. \tag{8}$$

For every l, either $\Sigma_i x_{il}^* \leq \omega_l$ or $y_l^* \leq \omega_l < \Sigma_i x_{il}^*$, by (5). Consequently, for every $l = 1, \ldots, L$,

$$\max\left(\sum_i x_{il}^* - \omega_l, 0 \right) \leq \left| y_l^* - \sum_i x_{il}^* \right|.$$

Hence, by (8) and the fact that $p_l^* > 0$ implies that $y_l^* = \omega_l$ (by [5]), we have

$$\| z^* \| \leq \delta_\varepsilon(p^*, c)\sqrt{L}/2,$$

where for each $l = 1, \ldots, L$,

$$z_l^* := \begin{cases} \sum_i x_{il}^* - \omega_l & \text{if } p_l^* > 0 \\ \max\left(\sum_i x_{il}^* - \omega_l, 0 \right) & \text{if } p_l^* = 0, \end{cases}$$

which establishes part *b* of theorem 1.

[20] Because the sum of the convex hulls of any finite number of sets is equal to the convex hull of their sum.

Part *a* is established by noting that for each $i > 0$, $x_i^* \in S_i$ implies that x_i^* is in the support of m_i^*, and so parts i and ii follow by (3) and (4).

Finally, part *c* is established by noting that $p^*\omega = p^*y^* \leq \Sigma_i(c_i + \varepsilon)$, where the equality follows from (5) and the inequality follows from (3). QED

REMARK 12. If one were to use the game in the proof of theorem 1 as the basis for an algorithm to compute the prices and allocations p^*, x^*, then one would need to ask agents to report their ordinal preferences (e.g., their indifference maps), not their utility functions. The mechanism would then use some canonical procedure to generate a utility representation. The reason is that if the algorithm maximizes $\varepsilon u_i(y_i) - (py_i - c_i)^+$ for each agent *i*, then, for any fixed ε, any agent who cannot affect the price would have an incentive to scale up his reported utility numbers $u_i(y_i)$.

References

Birkhoff, Garrett. 1946. "Three Observations on Linear Algebra." *Univ. Nacional Tucumán Revista A* 5:147–51.

Budish, Eric. 2011. "The Combinatorial Assignment Problem: Approximate Competitive Equilibrium from Equal Incomes." *J.P.E.* 119 (6): 1061–1103.

Budish, Eric, Yeon-Koo Che, Fuhito Kojima, and Paul Milgrom. 2013. "Designing Random Allocation Mechanisms: Theory and Applications." *A.E.R.* 103 (2): 585–623.

Dierker, Egbert. 1971. "Equilibrium Analysis of Exchange Economies with Indivisible Commodities." *Econometrica* 39 (6): 997–1008.

Glicksberg, Irving L. 1952. "A Further Generalization of the Kakutani Fixed Point Theorem, with Applications to Nash Equilibrium Points." *Proc. American Math. Soc.* 3 (1): 170–74.

Hylland, Aanund, and Richard Zeckhauser. 1979. "The Efficient Allocation of Individuals to Positions." *J.P.E.* 87 (2): 293–314.

Koopmans, Tjalling, and Martin Beckmann. 1957. "Assignment Problems and the Location of Economic Activities." *Econometrica* 25 (1): 53–76.

Rockafellar, R. Tyrrell. 1970. *Convex Analysis*. Princeton, NJ: Princeton Univ. Press.

Starr, Ross M. 1969. "Quasi-Equilibria in Markets with Non-convex Preferences." *Econometrica* 37 (1): 25–38.

von Neumann, John. 1953. "A Certain Zero-Sum Two-Person Game Equivalent to the Optimal Assignment Problem." In *Contributions to the Theory of Games*, vol. 2, edited by H. W. Kuhn and A. W. Tucker, 5–12. Princeton, NJ: Princeton Univ. Press.

Auctions in the *Journal of Political Economy*, 1894–2017

Ali Hortaçsu

University of Chicago

A quick JSTOR query for the keyword "auction" in the *JPE* returns 215 articles. The very first article in the list is "The Assignats: A Study in the Finances of the French Revolution," by Emile Levasseur, published in 1894 in the second issue of the second volume of the journal. In this engaging account of the monetary trials and tribulation of the French Revolution, there is a single reference to an auction: assignats were claims to public lands (mostly confiscated from the clergy) that were also decreed to become the currency of the revolutionary government, and auctions were used to sell the public lands to interested parties. Levasseur asserts that very few assignat holders actually used them to claim land, and as a paper money, the assignat's credibility was very much shaken by the lack of tax revenue for the new government and the many new additional assignat issues. Indeed, by the time the assignats were taken out of circulation in 1794 (in quite spectacular fashion: "at nine o'clock in the morning, with a great crowd of people looking on, all the tools which had been used in printing assignats were brought to the Place Vendome; the plates and stamps were broken, and reams of paper and 1167 millions of assignats burned" [191]), they had depreciated by 98 percent.

The topic of assignats in the French Revolution is taken up a century later by Thomas Sargent and François Velde, in their "Macroeconomic Features of the French Revolution" (1995), which appears in the 103rd volume of the *JPE*. Sargent and Velde reexamine the data on the price level and government expenditures during the French Revolution through the lens of modern macroeconomic theory. Their reading of the data is more favorable toward the assignat than Levasseur's: they write that "the tax-backed money scheme functioned adequately until a war broke out in 1792, which initially went badly for France. The government wanted more resources, so it divorced note issues from the land sales. The tax-backed money plan devolved into a fiat money scheme, causing real balances to drop and prices to rise quickly in early 1793 and threatening

I would like to thank John Asker, Alex Wolitzky, and Robert Porter, who generously commented on an early version of this note.

the base of the tax (the inflation tax) that was the government's lifeline" (475).

Perhaps more interestingly from an auction's perspective, Sargent and Velde's appendix provides a detailed description of the auction format used to sell clerical lands. The format is that of a two-stage English auction. In the first auction, a highest bidder is declared, but the highest bid is advertised as the starting value for a second auction. We should note an interesting rule used in the second auction that we will revisit below: "At the second auction, one-minute candles were lit in sequence and bidding started at the highest bid established during the first auction. If two candles expired without any new bids, the estate was awarded to the highest bidder from the first auction" (1995, 514).

The use of auctions in the sale or lease of government lands is also mentioned in G. O. Virtue's "Public Ownership of Mineral Lands in the United States," published in 1895 in the third volume of the journal. What we learn from this article is that, in its early history, the US government refrained from selling land with valuable mineral deposits, especially lead and salt. "No doubt the chief motive which induced Congress to reserve the salines from sale was the fear that they might become a monopoly" (187), says Virtue. Only shorter-term leases were allowed on reserved government mineral land. However, in 1847, the policy was reversed, and the government began to sell saline lands at public auctions.

The sale of the public's assets remains an important problem to this day, and auctions remain a very popular method of sale. Fast-forwarding a century from Virtue, we find several very interesting papers regarding the sale of government property through auctions. The first, "How Does Privatization Work? Evidence from the Russian Shops," by Barberis et al. (1996), is an empirical study of 452 shops in Russia, of which 413 were privatized, through either auctions or another method of privatization. This is a small but empirically insightful cross section of one of the most important privatization episodes in history, that of Soviet Russia, in which auctions played an important role.

Two other very important privatization efforts in recent memory in which auctions played an important role were in the auctioning of the wireless spectrum, which essentially led to the birth of the wireless communications sector, and the privatization of the electricity generation sector, which, once again, was an important application of auctions. The *JPE* is very fortunate to have featured two seminal papers on both topics: "Putting Auction Theory to Work: The Simultaneous Ascending Auction" by Paul Milgrom, published in 2000, and, in 1992, "Competition in the British Electricity Spot Market," by Richard Green and David Newbery.

Paul Milgrom is one of the co-inventors (along with Robert Wilson and Preston McAfee) of the simultaneous ascending auction (SAA) mecha-

nism used to auction off wireless spectrum licenses in many countries around the world. This is a package auction, where a large number of regional licenses are sold simultaneously, and where bidders may wish to purchase multiple licenses. The SAA proceeds in multiple rounds: in each round, the auctioneer sets the minimum bid as a price increment above the latest high bid on a license. Bidders on each round have to decide whether to not bid or meet/exceed the minimum bid requirement to keep the auction going. If no new bids are submitted, the auction ends. Milgrom first analyzes an essentially nonstrategic version of the auction, where bidders evaluate, at each round, what their preferred bundle of licenses are given the current minimum bid prices and their standing high bids. In the case in which the licenses are substitutes, Milgrom shows that straightforward bidding, in which bidders keep bidding for their preferred bundle at the current prices, leads to an efficient, competitive allocation.

What is key to this result is the fact that bidders never regret having become the standing high bidder on a license when the substitutes condition is met. In the case in which some licenses are complements, however, an "exposure" problem emerges, where some bidders may indeed regret having bid so high for a particular license, whose complement may have become unaffordable. Milgrom provides the example of a spectrum auction in the Netherlands in which this exposure problem became an important issue.

An important design feature that Milgrom discusses is the need for "activity rules" that will prevent bidders from playing "wait-and-see" tactics that can slow down progress in the auction. Note that in the clerical land auctions used in the French Revolution, the 1-minute candles played a similar role—to speed up the auction. Milgrom gives the example of an auction in which a laxer activity rule was utilized, which led to a much slower progression of bids and convergence to the final allocation.

In their 1992 piece studying electricity deregulation in England and Wales, Green and Newbery discuss the design for the spot market for electricity. What is used is a uniform price multi-unit auction in which generators submit supply schedules comprising multiple price-quantity bids. Green and Newbery assume, for the sake of modeling, that the supply schedules are continuous and thus invoke the "supply function equilibrium" model of Klemperer and Meyer (1989). There are multiple potential equilibria in this game, though Green and Newbery argue that capacity constraints and/or the threat of entry will reduce the range of potential equilibria. They then utilize calibrations of the supply function equilibrium model using data on generation costs to demonstrate the potential for considerable market power exercise and efficiency loss in the duopoly market structure of England and Wales. Beyond its specific

findings, the supply function equilibrium approach to modeling electricity markets has been very influential in the literature on privatized electricity generation markets.

Another market that, like the electricity market, operates through a multi-unit auction mechanism is the Treasury securities auction market. Treasury securities are typically sold by one of two mechanisms. The first is the discriminatory or "pay-as-bid" auction in which bidders submit multiple price-quantity bids and are charged the area under their revealed "demand" curve (which, of course, is not their true demand curve). The other mechanism is the "uniform price" auction, in which the aggregate "demand" curve of the bidders is intersected with the Treasury's supply to determine a single market-clearing price at which all inframarginal bids are fulfilled.

Which method is better? Discussed in Henry Goldstein's 1962 piece "The Friedman Proposal for Auctioning Treasury Bills," Milton Friedman had strong objections to the pay-as-bid mechanism in favor of the uniform price auction. Friedman argued that the uniform price auction would lead to competitive outcomes if no bidder had the power to affect the market-clearing price, which would lead to straightforward bidding (in the Milgrom sense), where each bidder would reveal his true willingness to pay. Friedman further argued that not needing to strategize in the auction would give incentive to small bidders to participate in the auction and thus lead to higher participation and potentially better rates for the Treasury. Friedman also thought that the pay-as-bid system would create more incentive for communication and, potentially, collusion among the bidders, as it required guessing where the market-clearing price would be. Goldstein's article provides some argument as to why collusion in the Treasury auction market may be difficult to sustain, a point that is also echoed by Michael Rieber in the 1964 piece "Collusion in the Auction Market for Treasury Bills." In his "Comment on 'Collusion in the Auction Market for Treasury Bills'" (1964), Friedman concedes that he may have overstated his case for the presence of potential collusion in the Treasury market but reiterates that his other points regarding the desirability of the uniform price auction remain unassailed.

Interestingly, the Green and Newbery article on England and Wales's uniform price auctions for electricity points out a case in which Friedman's assumption of "no market power" does not hold true and in which bidders can strategically withhold supply (or demand) to affect the market-clearing price, with important allocational and revenue/cost implications. Whether bidders have significant market power in (multi-unit) auctions is an empirical question that depends on the elasticity of the residual supply (or demand) function that each bidder faces. Empirically estimating the distribution of residual supply/demand functions that a bidder may ex-

pect to face in auction, and whether a given bidder is able to strategically "shade" her bid in response, is the topic of Hortaçsu and McAdams (2010). The empirical strategy also allows us to compute counterfactual revenues in the uniform price auction, under the assumption of equilibrium bidding. We find, in the context of Turkish Treasury auctions, that switching to the uniform price auction from a discriminatory format would not have significantly increased revenues. That said, our analysis does not take into account the differential participation incentives that Friedman hypothesized, which, I am sure, will be a topic for further discussion in the next decades to come.

On the topic of collusion in auctions, it is difficult to eschew a mention to George Stigler's classic "A Theory of Oligopoly" (1964), which appeared in volume 72 of the journal. Aside from laying out a theory of collusive behavior that is embodied in the vast literature on repeated games, Stigler had strong opinions regarding the effectiveness of public procurement auctions. In particular, Stigler thought that publishing bids after the close of each auction was an invitation to collusive behavior. The idea that price transparency facilitates collusion is a proposition that has been taken for granted for many decades, though a forthcoming paper by Takuo Sugaya and Alex Wolitzky points out interesting counterexamples to Stigler's intuition.

The topic of empirically detecting collusion in auctions, a topic of discussion between Friedman and his respondents in the context of Treasury auctions, is well represented in the *JPE* through two pioneering papers: Porter and Zona (1993) and Baldwin, Marshall, and Richard (1997). Porter and Zona, especially, argue how difficult it is to formulate rigorous empirical strategies to tell collusion apart from competition. To this date, much of the literature on collusion in auctions has followed an ex post strategy, where one analyzes documented bid rigging cases to gain insight into what the ring members were doing.[1]

Because of lack of space, with sincere apologies, I have omitted mention of many other well-known papers on auctions published by the *JPE*. Still, I hope that this brief survey has managed to convey its simple point: a lot has been written about auctions in the last 125 years, but it appears that many questions, some more than a century old, lie open for further inquiry. I hope that future volumes of the *JPE* will continue to capture this lively discussion.

[1] They could be implementing very sophisticated mechanisms to allocate rents. See, e.g., Asker (2010).

References

Asker, John. 2010. "A Study of the Internal Organization of a Bidding Cartel." *A.E.R.* 100 (3): 724–62.
Baldwin, Laura H., Robert C. Marshall, and Jean-Francois Richard. 1997. "Bidder Collusion at Forest Service Timber Sales." *J.P.E.* 105 (4): 657–99.
Barberis, Nicholas, Maxim Boycko, Andrei Shleifer, and Natalia Tsukanova. 1996. "How Does Privatization Work? Evidence from the Russian Shops." *J.P.E.* 104 (4): 764–90.
Friedman, Milton. 1964. "Comment on 'Collusion in the Auction Market for Treasury Bills.'" *J.P.E.* 72 (5): 513–14.
Goldstein, Henry. 1962. "The Friedman Proposal for Auctioning Treasury Bills." *J.P.E.* 70 (4): 386–92.
Green, Richard J., and David M. Newbery. 1992. "Competition in the British Electricity Spot Market." *J.P.E.* 100 (5): 929–53.
Hortaçsu, Ali, and David McAdams. 2010. "Mechanism Choice and Strategic Bidding in Divisible Good Auctions: An Empirical Analysis of the Turkish Treasury Auction Market." *J.P.E.* 118 (5): 833–65.
Klemperer, Paul D., and Margaret A. Meyer. 1989. "Supply Function Equilibria in Oligopoly under Uncertainty." *Econometrica* 57:1243–77.
Levasseur, Emile. 1894. "The Assignats: A Study in the Finances of the French Revolution." *J.P.E.* 2 (2): 179–202.
Milgrom, Paul. 2000. "Putting Auction Theory to Work: The Simultaneous Ascending Auction." *J.P.E.* 108 (2): 245–72.
Porter, Robert H., and J. Douglas Zona. 1993. "Detection of Bid Rigging in Procurement Auctions." *J.P.E.* 101 (3): 518–38.
Rieber, Michael. 1964. "Collusion in the Auction Market for Treasury Bills." *J.P.E.* 72 (5): 509–12.
Sargent, Thomas J., and François R. Velde. 1995. "Macroeconomic Features of the French Revolution." *J.P.E.* 103 (3): 474–18.
Stigler, George J. 1964. "A Theory of Oligopoly." *J.P.E.* 72 (1): 44–61.
Sugaya, Takuo, and Alexander Wolitzky. Forthcoming. "Maintaining Privacy in Cartels." *J.P.E.*
Virtue, G. O. 1895. "Public Ownership of Mineral Lands in the United States." *J.P.E.* 3 (2): 185–202.

The Economics of Crime

Steven D. Levitt

University of Chicago

The *Journal of Political Economy* has played an absolutely central role in the birth and the evolution of the literature devoted to the economics

of crime. By my calculations, eight of the 10 most highly cited works, including Becker's (1968) seminal article, either were published in the journal or were coauthored by an editor of the journal. It could be credibly argued that no single journal has had such a profound influence on a particular field of economics as the *JPE* has had on the economics of crime.

The modern literature on the economics of crime traces its roots to Becker's remarkable publication in this journal in 1968.[1] Filling 48 journal pages in a time when the typical article was fewer than 20 pages in length, Becker's article was a tour de force. At the heart of the paper was a relatively straightforward price theory model of the supply of and demand for offenses—a model of deterrence that has come to be known as "the economic model of crime." Interestingly, many of the predictions of Becker's model turn out to be quite at odds with what we observe in the real world. For instance, in his framework, fines are a more efficient means of punishment than imprisonment, which should make fines preferred to incarceration as a punishment for crime, but is not what we observe in practice. His model also argues that the combination of a low probability of punishment accompanied by an extremely severe penalty when a criminal is caught is the most efficient way to deter crime (because detection is costly but extracting wealth once a criminal is caught is "cheap"). Again, this is not how modern criminal justice systems function. It is not, however, the particulars of the predictions that make Becker's paper so remarkable, but rather the amazing breadth of facts, insights into human nature, and theoretical conjectures contained in the article. The basic idea underlying almost every theoretical paper in the voluminous literature that subsequently emerged can be traced back to Becker (1968).[2] This paper would become Becker's fourth most cited work, with roughly 10 times as many cites as any other article in the economics of crime literature.

One of the few theoretical insights that Becker missed in his 1968 paper was the idea of "marginal deterrence," a point developed by Stigler (1970) in an influential piece in this journal. In Becker's paper, all crimes would be punished with extremely severe sanctions. As Stigler notes, however, that leads to distorted incentives on the margin. If, for example, the punishment for robbery is the same as for murder, then a cornered robber might be willing to kill the police officers who apprehend him. More generally, to the extent that crimes are substitutes for one another, there

[1] Of course, as Becker (1968) notes, a number of great economic thinkers had written about some of the same issues more than 100 years earlier.

[2] I have often joked that Becker did a severe disservice to the economics of crime field by writing a paper that was too good and left too little for those who followed him to improve on.

is a role for distorting expected punishments in a manner that shifts crimes toward those that are least socially costly.

Interestingly, although a great deal of academic effort was devoted generally to exploring the economic model of crime in the 1970s and 1980s, these contributions were largely incremental, and the *JPE* published very little work on crime in those decades.[3] It was not until the influential work of Sah (1991) that the journal once again began to put its imprimatur on the field. Sah's paper extends Becker (1968) by endogenizing individual perceptions of punishment and adding dynamics to the economic model of crime. Because individual perceptions depend critically on local conditions, large differences in criminal propensities can emerge and persist across groups or communities that are observably similar in Sah's study. This work was important not only within the economics of crime but also more generally as an early example of endogenizing individual beliefs.

The late 1990s marked a turning point in the field. Up until then, virtually all of the most-cited works in the area were theory papers; since that time, highly cited papers have almost exclusively had a large empirical component.[4] An early example of data-driven crime studies in *JPE* is Levitt (1998), which exploits the natural experiment associated with the sharp discontinuity in expected punishment as an offender transitions from the juvenile justice system to the adult system. The juvenile justice system emerged in the 1800s as a response to the perception that mixing juveniles and adults in a prison system constituted cruel and unusual punishment for juveniles and had a corrupting influence. To this date, all states handle crimes committed by juveniles and adults very differently. Conveniently for the economist wishing to study the impact of punishment on crime, there is state-level variation in the age of majority (the age at which jurisdiction switches from the juvenile to the adult system) and in the relative severity of the juvenile and adult justice systems. Levitt finds strong evidence that criminal behavior is highly responsive to changes in expected punishment that arise with the transition to the adult system. In a field bedeviled by reverse causality (Fisher and Nagin 1978), this study was one of the first to provide empirical support for the economic model of crime using plausibly exogenous variation.

Duggan's (2001) article "More Guns, More Crime" proved influential in shaping an important public policy debate surrounding the impact of gun availability on crime. A few years earlier, in a highly controversial and frequently challenged paper published in the *Journal of Legal Studies*, Lott and Mustard (1997) argued empirically that laws making it easier for

[3] This is particularly notable given that George Stigler edited the journal for almost the entirety of those two decades.

[4] Becker, Murphy, and Grossman (2006) is an exception to this trend. This paper is an "old-school" application of price theory to the question of illegal drugs. While disarmingly simple, this paper sheds a plethora of important and nonintuitive policy-relevant insights.

people to carry concealed weapons sharply reduced crime. A number of papers provided specific critiques of Lott and Mustard's paper; see, for instance, Ludwig (1998). Duggan (2001) took a very different approach to the problem. Rather than focusing narrowly on concealed weapons, he tackled the issue of guns and crime in a more holistic manner. An enormous obstacle to such a study was the absence of data; somewhat amazingly, there are no reliable data about gun ownership at the state or local level, making it extremely difficult to address the issue with standard micro-empirical techniques. Duggan's first insight was to recognize that circulation counts of magazines devoted to gun issues—which were available at the local level—could be used as a credible proxy for gun ownership.[5] Duggan's most important conclusion is that increases in gun ownership (as proxied by gun magazine sales) strongly predict gun homicides but are essentially unrelated to nongun homicides and other crimes.

Fisman and Miguel (2007) provide another example of creative empirical work. They address the question of corruption and, in particular, the extent to which observed differences in corruption across countries can be tied to either cultural norms or enforcement of law. Corruption is a notoriously difficult question to address empirically, because it typically requires cross-country comparisons in a setting in which data are rare and of poor quality. Fisman and Miguel exploit a unique natural experiment tied to diplomatic immunity with respect to parking violations for United Nations diplomats. Prior to 2002, there is complete diplomatic immunity. Interestingly, Fisman and Miguel find a strong correlation between widely used measures of corruption in a country (based on surveys of people conducting business in that country) and the number of parking violations generated by the country's UN diplomats. This suggests that the sort of cultural norms that lead to corruption spill over into other behaviors, even among diplomats stationed thousands of miles away. After 2002, New York police were given the power to punish parking violations. In one of the clearest examples ever of the power of deterrence, violations fall by 98 percent when legal enforcement becomes possible.

Another influential empirical contribution published in *JPE* is Drago, Galbiati, and Vertove (2009), which exploits a remarkable natural experiment in Italy. At Pope John Paul II's urging, Italian lawmakers passed a law that led to the release on August 1, 2006, of any inmate with less than 3 years remaining on his or her sentence. Forty percent of all Italian prisoners

[5] One might think it would be an impossible task to convince the reader of the validity of a proxy when the reason a proxy is needed in the first place is the absence of direct data on guns. Duggan, however, manages to do this methodically and artfully by demonstrating a high degree of correlation between his proxy, which is available both at a geographically disaggregated level and with variation over time, and other gun-related outcomes that are available only at high levels of geographic aggregation or in the cross section.

were released that day! Notably, if a released prisoner was convicted of a new crime, an amount of time equal to the commuted sentence was added to the new sentence. Therefore, some criminals (i.e., those who had 3 years left to serve when they received early release) faced the specter of having an extra 3-year penalty if they got caught doing another crime. Those who were only a month away from completing their sentences on August 1, 2006, faced only an extra month. This policy thus induced plausibly exogenous variation in the severity of punishment, allowing Drago et al. to estimate how responsive former prisoners were to deterrence. They find that the Italian ex-convicts are extremely responsive to these variations in expected punishment, providing some of the most compelling evidence there is in support of deterrence.

After nearly 50 years of theoretical and empirical work, much is now understood about the economics of crime. Nonetheless, huge questions remain unanswered. It remains a mystery, for instance, why crime rose so much in the 1960s. We continue to have relatively little insight into the question of why some individuals become criminals and others do not. It is also difficult to explain why crime varies so much both spatially and temporally (although for a start on this question, see Glaeser, Sacerdote, and Scheinkman [1996] and Glaeser and Sacerdote [1999]). In some sense, however, public policies to reduce crime (many of them informed by economic thinking) have proven too successful from the perspective of the academic interested in studying crime. With the crime rate at less than half the level it was two decades ago in the United States and lower almost everywhere else in the world as well, the demand for crime research has no doubt also been diminished.

References

Becker, Gary S. 1968. "Crime and Punishment: An Economic Approach." *J.P.E.* 76 (2): 169–217.

Becker, Gary S., Kevin Murphy, and Michael Grossman. 2006. "The Market for Illegal Goods: The Case of Drugs." *J.P.E.* 114 (1): 38–60.

Drago, Francesco, Roberto Galbiati, and Pietro Vertova. 2009. "The Deterrent Effects of Prison: Evidence from a Natural Experiment." *J.P.E.* 117 (2): 257–80.

Duggan, Mark. 2001. "More Guns, More Crime." *J.P.E.* 109 (5): 1086–1114.

Fisher, Franklin M., and Daniel S. Nagin. 1978. "On the Feasibility of Identifying the Crime Function in a Simultaneous Model of Crime Rates and Sanction Levels." In *Deterrence and Incapacitation: Estimating the Effects of Criminal Sanctions on Crime Rates*. Washington, DC: Nat. Acad. Sci.

Fisman, Raymond, and Edward Miguel. 2007. "Corruption, Norms, and Legal Enforcement: Evidence from Diplomatic Parking Tickets." *J.P.E.* 115 (6): 1020–48.

Glaeser, Edward, and Bruce Sacerdote. 1999. "Why Is There More Crime in Cities?" *J.P.E.* 107, no. 6, pt. 2 (December): S225–S258.

Glaeser, Edward, Bruce Sacerdote, and Jose Scheinkman. 1996. "Crime and Social Interactions." *Q.J.E.* 111 (2): 507–48.

Levitt, Steven D. 1998. "Juvenile Crime and Punishment." *J.P.E.* 106 (6): 1156–85.
Lott, John, and David Mustard. 1997. "Crime, Deterrence, and Right-to-Carry Concealed Handguns." *J. Legal Studies* 26 (1): 1–68.
Ludwig, Jens. 1998. "Concealed-Gun-Carrying Laws and Violent Crime: Evidence from State Panel Data." *Internat. Rev. Law and Econ.* 18:239–54.
Sah, Raj. 1991. "Social Osmosis and Patterns of Crime." *J.P.E.* 99 (6): 1272–95.
Stigler, George J. 1970. "The Optimum Enforcement of Laws." *J.P.E.* 78 (3): 526–36.

Experimental Economics in the *Journal of Political Economy*

John List

University of Chicago

Similar to the spirit in which astronomy draws on the data from mechanics and physics to make deeper insights, experiments can help to provide the necessary behavioral principles to permit sharper inference from naturally occurring data in economics. Indeed, experiments have helped to uncover the key causes and underlying conditions necessary to produce data patterns observed in the field. At the same time, the experimental method has gone beyond merely a complementary role, as today experiments generate data that provide crisper tests of economic theory than previously achieved.

In contrast to other sciences, the experimental approach has not progressed to the point of being the cornerstone of the scientific method in economics just yet, but it has progressed sufficiently to find itself in the center of key debates and is well represented in every major economics journal. This was not always the case. Indeed, the *Journal of Political Economy* has played a central role in the general acceptance of the experimental approach.

To showcase this fact, I focus narrowly on two areas of experimental inquiry: market institutions and individual choice. Given that two standard assumptions that underlie standard economic theory are that (i) markets are cleared via the institution of Walrasian tâtonnement and (ii) agents aim to maximize utility, it is fitting that the *JPE* has contributed to both experimental research agendas.

I. Market Institutions

Empirical attempts to explore the effects of institutions have been plentiful, but traditional tools have had limited success. In what is broadly believed to be the first market experiment, Chamberlin (1948) used Harvard students participating in decentralized one-shot bargaining markets to explore whether price and quantity converged to the intersection of supply and demand. Chamberlin observed that volume was typically higher and prices typically lower than predicted by competitive models of equilibrium. Efficiency was also frustrated in these bilateral negotiating markets.

Vernon Smith, a Harvard student at the time and one of Chamberlin's experimental subjects, later refined Chamberlin's work (Smith 1962) by varying two key aspects of the experimental design: (i) centrally occurring open outcry of bids and offers (commonly termed "double auction markets") and (ii) multiple market periods (allowing agents to learn). Empirical results from Smith's experiments were staggering—quantity and price levels were very near competitive levels—and served to present the first evidence that Walrasian tâtonnement, conducted by a central auctioneer, was not necessary for market outcomes to approach neoclassical expectations. It is fair to say that this general result remains one of the most robust findings in experimental economics today.

Smith complemented this early work with another *JPE* study in 1965 that presented the "excess rent" model, which is defined as the total rent that would be obtained if all agents who want to trade at the prevailing price were to trade minus the total rent at the competitive equilibrium. In his *experimentum crucis*, Smith (1965) reports that predictions of the excess rent model are more consonant with his data than the Walrasian model using a series of double auction markets.

Several research areas have arisen from these seminal contributions of Chamberlin and Smith. List (2004) represents a field experiment that moves the Chamberlin bilateral bargaining institution from the lab to the natural setting in which the actors actually undertake decisions. List's field experiment therefore represents an early empirical test in an actual marketplace in which agents engage in face-to-face continuous bilateral bargaining in a multilateral market context. Much as in Smith's (1962, 1965) setup, the market mechanics in these bilateral bargaining markets are not Walrasian. In contrast to Smith's work, however, List's design shifts the task of adaptation from the auctioneer to the agents; in doing so, the market structure transforms the problem of stability of equilibria as a question about the behavior of actual people as a psychological question—as opposed to a question about an impersonal market. Two key results of List's study are the high efficiency obtained in his markets and the strong

tendency for exchange prices to approach the neoclassical competitive model predictions, especially in symmetric markets.

In a dramatic extension of Smith's (1962, 1965) double auction market, Gode and Sunder (1993) design markets that compare results from human traders with those from "zero-intelligence traders." These traders are irrational in the sense that they are not maximizing; rather they are submitting random bids and offers. The sole constraint is that buyers cannot pay more than their willingness to pay and sellers cannot sell for less than their marginal cost. Their amazing results suggest that even with such traders, the market approaches 100 percent efficiency. Their conclusion is that the efficiency of the double auction market arises from the institution itself, not from market experience, learning, or the like (this is not to imply that there is no role for human traders to significantly improve their ability to gain surplus when learning about the state of the world in double auction markets; see Plott and Sunder 1982).

Of course, there are hundreds of other excellent examples of noteworthy experiments in this area that could fill more than five tomes, but I need to move on to discuss individual choice, however unjust and risky that might appear.

II. Individual Choice

A second important strand of experimental work that the *JPE* has contributed to advancing is in the area of individual choice. An early choice experiment that set into motion several branches of subsequent work was due to Thurstone (1931). In a series of hypothetical choice exercises, Thurstone attempted to measure individuals' indifference curves by asking individuals to make choices between bundles of hats and coats, hats and shoes, and shoes and coats. In an exciting early test of ordinal preference theory, Thurstone concluded that indifference curves could adequately represent his choice data.

Thurstone's work drew the ire of Wallis and Friedman (1942), who in a provocative study argued that the choice set was ill-specified and that the choices themselves were hypothetical. For example, they argued that "it is questionable whether a subject in so artificial an experimental situation could know what choices he would make in an economic situation; not knowing, it is almost inevitable that he would, in entire good faith, systematize his answers in such a way as to produce plausible but spurious results" (179–80).

The Wallis and Friedman (1942) critique led immediately to a follow-up study due to Rousseas and Hart (1951), who constructed (arguably) a more realistic choice scenario whereby experimental subjects in the lab

made choices over breakfast items: number of eggs and strips of bacon, for example. In a key innovation, they had individuals make a single choice repeatedly. This avoided certain complications, but the aggregation of choices across individuals to test the relevant theory became an issue. Related to this inquiry was the early lab experimental work of Mosteller and Nogee (1951), which explored utility over additional money income. This overarching line of research even induced a mini debate about the merits, and what had been accomplished, within the area of experimental economics that was published in the *JPE* (Castro and Weingarten 1970; Naylor 1972).

This early work anticipated several lines of interesting research. For example, within the area of hypothetical choice, Cummings et al. (1997) explored whether referenda for nonmarketed goods and services were influenced by hypothetical bias. They found that considerably more people voted in favor of the referendum when it was hypothetical versus when it was real. In an early field experiment, List and Shogren (1998) executed hypothetical and real auctions and reported similar results: the average bid in the hypothetical auction was roughly three times higher than in the real auction.

An area of individual choice following in the spirit of Wallis and Friedman (1942) is on the "demand side" of charitable fund-raising. An early field experiment on the demand side is summarized by List and Lucking-Reiley (2002), who raised money for the Center for Environmental Policy Analysis at the University of Central Florida in a field experiment. They found key success for the theoretical prediction of Andreoni (1998): that seed money increases the amount of public-good provision in a charitable fund-raiser, from zero to some positive equilibrium level G^* (greater than or equal to the threshold level). This research led to many related field experiments, with the recent work of Andreoni, Rao, and Trachtman (2017) and Perez-Truglia and Cruces (2017) representing two excellent examples.

Relatedly, to explore the importance of social preferences, List (2006) carries out various field experiments analyzing gift exchange. The games have buyers making price offers to sellers, and in return sellers select the quality level of the good provided to the buyer. The artefactual field experimental (lab-like) results mirror the typical lab findings with other subject pools: strong evidence for social preferences was observed through a positive price and quality relationship. Yet, when the environment is moved to the marketplace via a natural field experiment, where dealers are unaware that their behavior is being recorded as part of an experiment, little statistical relationship between price and quality emerges. Similar insights on the effect of the situation in an entirely different setting and game form can be found in List (2007).

Much as in the brief summary of market institutions above, there are several other examples of noteworthy individual choice experiments. Camerer (1998) is one neat example that explores betting choices at a racetrack. Cason and Plott (2014) explore recent challenges of revealed preference theory by using lab experiments with interesting framings. They conclude that "mistakes in choices obscured by a possible error at the foundation of the theory of framing can masquerade as having been produced by nonstandard preferences" (1235). Kahneman, Knetsch, and Thaler (1990), in seminal work, test how well predictions of reference-dependent theory explain trading decisions of lab subjects. They find evidence consonant with reference dependence, and this work has touched off a line of research that continues to thrive today.

References

Andreoni, James. 1998. "Toward a Theory of Charitable Fund-Raising." *J.P.E.* 106 (6): 1186–1213.

Andreoni, James, Justin M. Rao, and Hannah Trachtman. 2017. "Avoiding the Ask: A Field Experiment on Altruism, Empathy, and Charitable Giving." *J.P.E.* 125 (3): 625–53.

Camerer, Colin F. 1998. "Can Asset Markets Be Manipulated? A Field Experiment with Racetrack Betting." *J.P.E.* 106 (3): 457–82.

Cason, Timothy N., and Charles R. Plott. 2014. "Misconceptions and Game Form Recognition: Challenges to Theories of Revealed Preference and Framing." *J.P.E.* 122 (6): 1235–70.

Castro, Barry, and Kenneth Weingarten. 1970. "Toward Experimental Economics." *J.P.E.* 78 (3): 598–607.

Chamberlin, Edward. 1948. "An Experimental Imperfect Market." *J.P.E.* 56 (2): 95–108.

Cummings, Ronald G., Steven Elliott, Glenn W. Harrison, and James Murphy. 1997. "Are Hypothetical Referenda Incentive Compatible?" *J.P.E.* 105 (3): 609–21.

Gode, Dhananjay K., and Shyam Sunder. 1993. "Allocative Efficiency of Markets with Zero-Intelligence Traders: Market as a Partial Substitute for Individual Rationality." *J.P.E.* 101 (1): 119–37.

Kahneman, Daniel, Jack L. Knetsch, and Richard H. Thaler. 1990. "Experimental Tests of the Endowment Effect and the Coase Theorem." *J.P.E.* 98 (6): 1325–48.

List, John A. 2004. "Testing Neoclassical Competitive Theory in Multilateral Decentralized Markets." *J.P.E.* 112 (5): 1131–56.

———. 2006. "The Behavioralist Meets the Market: Measuring Social Preferences and Reputation Effects in Actual Transactions." *J.P.E.* 114 (1): 1–37.

———. 2007. "On the Interpretation of Giving in Dictator Games." *J.P.E.* 115 (3): 482–93.

List, John A., and David Lucking-Reiley. 2002. "The Effects of Seed Money and Refunds on Charitable Giving: Experimental Evidence from a University Capital Campaign." *J.P.E.* 110 (1): 215–33.

List, John A., and Jason F. Shogren. 1998. "Calibration of the Difference between Actual and Hypothetical Valuations in a Field Experiment." *J. Econ. Behavior and Org.* 37 (2): 193–205.

Mosteller, Frederick, and Philip Nogee. 1951. "An Experimental Measurement of Utility." *J.P.E.* 59 (5): 371–404.

Naylor, Thomas H. 1972. "Experimental Economics Revisited." *J.P.E.* 80 (2): 347–52.

Perez-Truglia, Ricardo, and Guillermo Cruces. 2017. "Partisan Interactions: Evidence from a Field Experiment in the United States." *J.P.E.* 125 (4): 1208–43.

Plott, Charles R., and Shyam Sunder. 1982. "Efficiency of Experimental Security Markets with Insider Information: An Application of Rational-Expectations Models." *J.P.E.* 90 (4): 663–98.

Rousseas, Stephen W., and Albert G. Hart. 1951. "Experimental Verification of a Composite Indifference Map." *J.P.E.* 59 (4): 288–318.

Smith, Vernon L. 1962. "An Experimental Study of Competitive Market Behavior." *J.P.E.* 70 (2): 111–37.

———. 1965. "Experimental Auction Markets and the Walrasian Hypothesis." *J.P.E.* 73 (4): 387–93.

Thurstone, Louis L. 1931. "The Indifference Function." *J. Soc. Psychology* 2 (2): 139–67.

Wallis, W. Allen, and Milton Friedman. 1942. "The Empirical Derivation of Indifference Functions." In *Studies in Mathematical Economics and Econometrics, in Memory of Henry Schultz*, edited by Oscar Lange, Francis McIntyre, and Theodore O. Yntema, 175–89. Chicago: Univ. Chicago Press.