An Approach to Incorporating Psychology into Economics*

By MATTHEW RABIN

This article proposes an approach to improving the psychological realism of economics while maintaining its conventional techniques and goals: formal theoretical and empirical analysis using tractable models, with a focus on prediction and estimation. These techniques have proven valuable for studying the behavioral and welfare implications of economic phenomena—oil shocks, higher education, job search—and policy changes—extending unemployment insurance, lowering cigarette taxes, fiscal cliffing, or regulating monopoly. Models with greater psychological realism can best improve understanding of the economy and policy when done with the mathematical and statistical rigor required by mainstream economic questions and methods. In this view, we should develop new models of humans—but use the same old model of modeling humans.

Any overly structured notion of methodology, such as what I outline in abbreviated form in this article, is necessarily too narrow. Thinking is hard.[1] So any good thinking on useful topics should be appreciated, and smart and creative people tend to do good research in whatever mode they choose. But this should not warrant agnosticism about methods, and I believe that at this moment more of our efforts to improve the psychological realism of economics should be devoted to developing (necessarily imperfect) portable, mathematical models. The claim is not that everybody should do research of this sort. Indeed, this approach is motivated by engagement with other modes of research: translating improved psychological assumptions into portable models is crucial for them to influence prevalent modes of economics, from "applied theory" to structural econometrics to theory-guided, reduced-form empirical work. There may be many motivations to identify flaws in the models economists traditionally use, but the thoughts below are written for those researchers aiming to improve mainstream economics.

* Department of Economics, University of California, Berkeley, 549 Evans Hall, #3880 Berkeley, CA 94720-3880 (e-mail: rabin@econ.berkeley.edu). I thank David Laibson, Tristan Gagnon-Bartsch, and Erik Eyster for helpful comments, but opinions expressed in this article are probably solely my own.

[1] If you don't think so, you're not doing it right.

I. Formal Theories

Two facets of economics are worth keeping in mind to see why it's important that modifications to models be broadly applicable, predictively sharp, and mathematically precise. First, most economic theory is not about developing new assumptions about people—it is about seeing the implications of fixed assumptions in different economic situations. To be integrated into economics, therefore, we must focus on portable improvements that can be useful as *input* into economic theory, not just research that will appeal to fellow flaw-finders. For good scientific reason, the heart of much psychological and behavioral research is to investigate possible flaws, caveats, and modifications to previous theories. But the heart of modifying existing economic theory is to formulate credible and systematic alternatives. Second, the core empirical exercise in economics is not to identify the existence of phenomena, but to understand their ecological significance. Predicting behavioral and welfare responses when a firm cuts wages by $1 an hour or a government raises cigarette taxes by $1 per pack requires more than the list of motivations and cognitive limitations that affect behavioral responses to such events. It requires knowing how big a role those things play, and how they interact. This generates a glaring need for new insights rendered in a way that allows meaningful measurement. Economists don't study labor markets with experiments showing the existence of tastes for more money and more leisure—they study the interplay of these tastes across settings. Economists likewise won't be able to modify labor economics simply by acknowledging aversions to loss or unfair treatment—they must study the power and interplay of these factors with other motivations. "Existence proofs" can be the right first step to convince econo-

mists that improvements might be necessary. But mathematical theory and measurable variables are needed to make those improvements.

Precision means simplification, and tolerance for the imperfections induced by simplification is a key intellectual leap needed for rigorous and credible progress. All precise theories are wrong, so all good precise theories are wrong.[2] Indeed, a core goal of precision in new theories is to help see their limits that can guide us in further improvements. In this view, maximal precision in proposing new theories comes not from confidence that the theories will be right, but rather from a desire to expose how they are wrong.

Besides the necessary evil of imperfection when embracing precision, models should aim for two crucial criteria: power and scope. By power, I mean: does the model actually tell us significant stuff *not* to expect? Does it make different or sharper predictions than competing theories? This is not just a matter of making sure a theory is "falsifiable". The goal is not just to test a theory—it is to have the theory rule enough things out to make it valuable to science. By scope, I mean: how broad is the set of situations to which it applies? We don't want separate theories for humans who are herding, for those bidding in auctions, and for those trading in financial markets.[3] Different things *matter* in different contexts; e.g., in many contexts, departures from self interest don't matter. But the basic theory of human motivations and cognitive capacities shouldn't be assumed to vary. Scope is especially important in economic theory: "comparative statics"—showing how outcomes vary as the environment changes—is the once and future key goal of economic theory. New models of humans that are undefined or uninterpretable outside of a narrow context, or whose implications in other contexts are simply neglected, are not the stuff of comparative statics.

The two desiradata of power and scope let us usefully deconstruct some instances where theories are marketed as "general". Low power is general and high power is specific; broad scope is general and narrow scope is specific. Clearly, theories have the most explanatory power when they are general in their applicability across contexts but *specific* in their predictions within each context. Describing something as "a general theory" elides the two notions of generality—one good, one bad—in an especially unuseful way. It would be nice to replace such obfuscatory spin with clearer statements about explanatory power.

II. Grecian Formula

The aim to have realism-improving theories be maximally useful to core economic research suggests a particular approach that I've goofily named "PEEMs" — portable extensions of existing models. One should (a) extend the existing model by formulating a modification that embeds it as parameter values with the new psychological assumptions as alternative parameter values, and (b) make it portable by defining it across domains using the same independent variables in existing research, or proposing measurable new variables.

The research program developing PEEMs involves observing phenomena that seem true, important, and *general*, and thinking about whether and how they fit with current theories. When observations seem not to fit current theories, we think about what is going on, and how to modify existing models to better match reality.[4] To embrace the derision with which some see it, the recipe is to take a current model, pick an available Greek letter, toss it in with a bunch of clear right-hand-side variables, and model away. Formally, consider a hypothetical Greek letter: "deppa", with symbol Þ. Reframe the pre-existing model as assuming some value for scalar or vector Þ (usually 0, 1, or ∞) within an explicit model $f(x_1, x_2, ...|Þ)$, where the x_i are (old or new) variables with serious potential for empirical study. They should be either observable, or not-directly-observable factors commonly used as inputs in classical economic modeling.

Examples of PEEMs abound in some domains. The many models that modify traditional expected utility to allow for non-linear probability preferences, for instance, can be seen to

[2] This of course does not imply that all imprecise theories are right. It's just harder to identify flaws when models are underspecified.

[3] And we certainly don't need additions to the 70,000 theories designed solely to explain behavior in the ultimatum game.

[4] Because extant theories are rarely devoid of truth and insight, often this involves "adding" a component while preserving many of the extant theory's predictions.

embed the same risk-free cardinal preferences as expected utility, differing only in how the probabilities enter the risk preferences. Within behavioral game theory, several models have been proposed as alternatives to Nash equilibrium that use precisely the same definitions of games and no more information. Once a parameter value is chosen, there are no degrees of freedom in the literal formulations of McKelvey and Palfrey's **(**1996**)** "quantal-response equilibrium", with parameter $\lambda \geq 0$, or Eyster and Rabin's (2005) cursed equilibrium, with parameter $\chi \in [0, 1]$. Recent models of non-Bayesian probability judgments parameterize errors in updating beliefs about hypotheses by 1 or 2 parameters. Kőszegi and Rabin (2006) model a variant of prospect theory with parameters $\eta \geq 0$, $\lambda \geq 1$, and $\alpha \in (0, 1]$. The most successful PEEM to date is surely the simple model developed by Strotz **(**1956**)** and rejuvenated by Laibson (1997), on what O'Donoghue and Rabin (1999) refer to as present bias: parameter β captures short-term impatience, and parameter $\widehat{\beta}$ captures misprediction of future impatience.

There is a catch. Although the models in McKelvey and Palfrey (1996**)** and Eyster and Rabin (2005) are (*when literally applied*) fully formed, most PEEMs require interpretations that go beyond the original model, even when the independent variables are the regular stuff of economic theories and empirics. The model of present bias, for instance, requires interpretations of the timing of utility flows that need not *per se* be considered a primitive of models without present bias. Fortunately, the timing of utility is obvious, observable, and well-understood in prevailing contexts.[5] Models of probability judgment that depart from Bayesian reasoning, on the other hand, often require even more ancillary assumptions specifying how decisionmakers frame hypotheses.

Other extensions of existing theories that one might call FEEMs—frameworks extending existing models—go much further than needing interpretation: they introduce new factors that can influence outcomes without specifying restrictions on those factors across contexts, or deriving those factors from traditional variables. Consider Jehiel's (2005) elegant notion of analogy-based expectations equilibrium (ABEE), in which players bundle the results of other players' behavior into analogy classes and best-respond to beliefs corresponding to the average behavior within each analogy class. Jehiel (2005) and subsequent papers have embedded a wide variety of different analogy classes. And it is noteworthy that ABEE is frequently mentioned in published papers as a literal or near analog of specific new limited-rationality theories, indicating that ABEE can be made to match or approximate the insights of a wide array of independently developed models. Relaxing unrealistic restrictions or highlighting new factors is excellent science, and FEEMs can help organize and inspire the development of new PEEMs.[6] ABEE, for instance, may accelerate progress on new theories of failures of contingent thinking of the sort researchers are exploring. But when the restrictions outlined in specific applications are not pinned down *a priori*, FEEMs are better described as *accommodating* phenomena than as *predicting* them.[7]

[5] As such, in fact, the theory pins down predictions on a priori grounds much more than many old and new notions of self control, since temptation, taste for commitment, etc., are all determined without degrees of freedom by the timing of utility flow.

[6] Or a framework may introduce new variables or inputs into economic models. One might imagine doing game theory by building analogy classes into a game's structure. A model for this is Geanakoplos, Pearce, and Stacchetti's (1989) elegant "psychological-games" framework, which modifies games by putting higher-order beliefs into the preference structure. (Indeed, this example illustrates how FEEMs can inspire PEEMs: Rabin (1993) builds a PEEMish model by inflexibly transforming all cardinal-payoff games into psychological games in a way that captures concerns for reciprocal fairness.)

[7] Likewise, PHEEMs—post hoc extensions of existing models—that appear in many papers to describe specific data sets should be treated differently than the *a priori* restrictions of PEEMs. Perhaps we can find ways to apply techniques analogous to the statistical adjustments used in empirical work to make comparisons among models with differing degrees of freedom more meaningful. Just as empirical research communities increase standards on "specification mining" and post-hoc hypotheses, we could promote corresponding standards for improvements in psychological realism. Researchers selecting a particular implementation of their framework could be encouraged to characterize fully the set of outcomes accommodated by their general FEEM, and researchers fitting models to specific data sets could report all the context-specific theories they considered before settling on the particular PHEEM reported in the paper. Perhaps a useful technique for those researchers prolific in generating new theories for new findings is to combine a current experimental data set with their other recent papers and formulate a single theory to explain all the data from this meta-experiment. Being able to observe the resulting tightness of fit and degrees of freedom will make comparisons to PEEMs that are pinned down across experiments more meaningful.

III. Advantages of This Approach

The development of PEEMs that are conducive to integration within economics will place an intellectual burden on economists to consider seriously insights from psychology. Even when assuming irrationality, even when assuming components of the utility function not learned in graduate school, and even when insights are inspired by the genius of Kahneman, Tversky, and Thaler rather than the genius of Samuelson, Arrow, and Becker, economists will have a scientific obligation to treat new theories with the same standards as more familiar theories. Researchers can use PEEMs to commit radical acts of normal science: by turning all empirical research into tests of the mean and confidence interval of a parameter Þ, we can insist new-to-paradigm assumptions be put on equal footing with more familiar assumptions.

PEEMs can rectify all sorts of traditional reactions by economists to alternative theories. Loose claims that some unfamiliar assumption won't matter for economics can be replaced with tight and careful investigations. Here's the credit-card industry when β = 1. Here's the credit-card industry when β = .7 and $\widehat{\beta}$ = .9. It matters. And it is truer. More generally, PEEMs can replace a certain type of market mysticism that markets will wipe out "behavioral" phenomena with the same sort of serious market analysis that predominates economics based on more familiar assumptions. Vague intuitions that (say) cigarette markets will shut down if demand for cigarettes derives from systematic underestimation of their addictiveness or from self-control problems rather than rational consumption can be shown to be false. PEEMs can also help put to rest habitual dismissiveness that "anything can happen" once one modifies familiar assumptions. Although we are necessarily still at the stage of estimating parameters, it is useful to be concrete that improvements are not coming through degrees of freedom. Eyster and Rabin (2005) propose that modeling a type of failure of inference using $\chi = .4$ fits evidence across games better than the fully rational $\chi = 0$. Models of present bias and naivety specifying $\beta = .7$, $\widehat{\beta} = .8$ fit most data better than the classical parameters $\beta = \widehat{\beta} = 1$. Indeed, in this domain a good guess is that improved assumptions will fit behavior better while *removing* a degree of freedom: instead of trying to fit data with wildly varying values for the traditional discount factor, δ, once we use better values of β and $\widehat{\beta}$ we can also restrict our models to the more reasonable yearly $\delta = .95$.

Turning empirical puzzles and controversies into debates over parameters can neutralize an array of other ways new-to-paradigm theories are treated differently than traditional theories. It can eliminate a common miscounting of "the number of assumptions", treating the assumption $\beta = \widehat{\beta} = 1$ built into most economic models is not an assumption, but β = .7, $\widehat{\beta}$ = .9 as an assumption. Making them explicit is not adding assumptions, it is just unburying them. And the "parameterization" of empirical debates can expose a canard on how to think about "unstable" factors. When parameter estimates are highly variable across contexts, it is important to make this salient so that researchers know further improvements are needed. But instability of estimates should not be used to support bad precision over good precision. If measured present bias varies between 0 and 1, then always assuming one particular $\beta < 1$ is an imperfect model—but always assuming $\beta = 1$ is a worse one.

In addition, the act of parameterizing debates can help clarify some ways that tests of new theories fail to achieve a fundamental criterion of mainstream empirical economics: identification. The most traditional form of such *non-identification tests* is "sufficiency bias," a genre of argument familiar to behavioral economists from the early days. Every time a within-paradigm theory *could* explain a phenomenon, it was coded as evidence in favor of the theory, without worrying whether proposed alternatives could also explain it. Examples where what looked like concern for fairness could also be consistent with pure self interest were greeted with great excitement. But since it was clearly possible that what looked like a concern for fairness was in fact a concern for fairness, greater enthusiasm for instead seeking out examples where the two *were* distinguished would have been a healthier response.

A newer genre of non-identification test could be called "insufficiency bias": a theory is rejected based on data inconsistent with it, without due concern for whether *any* theory explains the data. Theory-A-can't-explain-phenomenon-

X tests look like good science. But when, upon closer inspection, it is clear that the theory is rejected because it omits factors that all other theories omit, they look less so. If the *distinguishing* features of a theory are not the *cause* of the poor fit, it comes closer to scientific mischief than scientific progress to single out that theory's failure. Such approaches would be deemed silly if turned into PEEMish empirical analysis: if a new theory differs from a previous one solely by claiming the coefficient on some variable is $Þ = .2$ rather than $Þ = 0$, it would not be part of coherent empirical analysis to state "$Þ = .2$ cannot explain our data". The appeal of such non-identification tests is clear in historical context: because many improvements to psychological realism have been motivated by the desire to explain important "anomalies", examples where the anomaly arises in the absence of the explanation may tempt researchers to doubt the new theories. Examples where the types of self-control problems or time inconsistency often explained by present bias *can't* be explained by present bias, or the types of overbidding in auctions that is often explained by boundedly rational strategic thinking *can't* be explained by any existing models are invoked to criticize these explanations. This has some appeal if one fully tethers a new theory with an anomaly it is purported to address. But framed as normal-science empirical testing, the evidence would only be compelling when it is part of a well-specified model that identifies $Þ = 0$ rather than $Þ = .2$ once a confound is controlled for. In normal-science economics, it is almost never of interest to seek a *sole* independent variable that explains outcomes of interest, and science-friendly interpretations of new theories should focus on whether the theories are improvements over older explanations.

A four-word premise of the entire PEEM approach is worth repeating, because it captures the above concerns and is an abiding theme throughout. *All theories are wrong*. It is normal for research in some domains of empirical research to "test a theory", and either reject it because it is imperfect or accept it because it does "surprisingly well". It is *not* normal in other literatures. When researchers believe all theories are simplifications with the goal of improving upon previous theories, theories are to be judged by the degree to which they improve. Empirical evidence showing old and new theories are missing something becomes fodder for further iterations of improvement—not tests of whether the previous increment is perfect. PEEMs help re-direct empirical and theoretical work towards such improvements.

Finally, PEEMs lend themselves to integration into mainstream economics by their use in two types of comparative statics. The first is to look within chosen environments at how predictions change with Þ. This allows us to test a theory's empirical validity in comparison to existing theories by estimating Þ, and to assess their potential value added in comparison to existing theories by observing whether Þ matters in important economic contexts.[8] The second type of comparative statics is the more traditional one: fixing a value of Þ that accords to the improved assumptions, how does changing the environment affect economic outcomes? Applying this form of comparative statics indicates that the improved psychological realism is ready for economic primetime.

IV. Conclusion

I conclude with three comments on the limits to methodology. The first is a reminder: most great research improving psychological realism is outside the framework I propose. Although some of what I discuss above touches on the one methodological rule that should be universal—honest and accurate statistical analysis of a model's fit—for the most part the discussion is a rule of thumb for what style of research might be most productive.

The second is specific to PEEMs. The aim of having portable models is to assure the testability of their accuracy and of their usefulness to economics. The act of formalizing ideas in precise models does not achieve these goals if these goals are not taken seriously, beginning with researchers. With explanatory power comes explanatory responsibility: researchers developing PEEMs should ask themselves what the theories imply in basic settings outside the theoretical or empirical contexts in which they are developed. One of the most active areas in developing al-

[8] As such, this approach will surely uncover cases where seemingly true alternative models turn out, upon further analysis, to be worse fits than extant models, or to have far less economic significance than conjectured.

ternative models of people is also among the most PEEMish: social preferences. Seeded by game forms specifying cardinal selfish payoffs to each player, these models capture departures from purely self-interested preferences with formulaic, universally defined transformations of the selfish payoffs. But once one steps outside of very circumscribed settings, facets of these models clearly correspond to parameter values that fit worse than the baseline self-interested model they aim to replace. Because cases where these models make worse predictions than the classical model are pervasive, and because attempts to derive plausible economic implications from these models are so few, one does not gain confidence that this literature is aiming to leverage these PEEMs to improve economic theory. In other domains, some models seem likely to destroy huge swathes of realistic economic predictions for the sake of explaining behavior in a particular domain or data set. Perhaps more efforts to work out implications of proposed PEEMs across settings will allow us to improve the quality of theories we propose.

The final limit on methodology brings us back to the core of this research program. The two most important ingredients of any PEEM are its psychological realism and its economic relevance. This article emphasizes the value of power and scope in proposing modifications to classical economic assumptions. These modifications are only helpful when the resulting precise predictions are, in ways that matter in important economic situations, more realistic than the old precise predictions.

V. References

Eyster, Erik, and Matthew Rabin. 2005. "Cursed Equilibrium." Econometrica, 73(5): 1623–72.

Geanakoplos, John, David Pearce, Ennio Stacchetti. 1989. "Psychological games and sequential rationality." Games and Economic Behavior, 1(1): 60–79.

Jehiel, Philippe. 2005. "Analogy-Based Expectation Equilibrium." Journal of Economic Theory, 123(2): 81–104.

Kőszegi, Botond and Matthew Rabin. 2006. "A Model of Reference-Dependent Preferences." Quarterly Journal of Economics, 1221(4): 1133–65.

Laibson, David. 1997. "Golden Eggs and Hyperbolic Discounting." Quarterly Journal of Economics, 112(2): 443–77.

McKelvey, Richard D., and Thomas R. Palfrey. 1996. "A Statistical Theory of Equilibrium in Games." Japanese Economic Review, 47(2): 186–209.

O'Donoghue, Ted and Matthew Rabin. 1999. "Doing It Now or Later." American Economic Review, 89(1): 103–24.

Rabin, Matthew. 1993. "Incorporating Fairness Into Game Theory and Economics." The American Economic Review, 83: 1281-302.

Strotz, Robert H. 1956. "Myopia and Inconsistency in Dynamic Utility Maximization." Review of Economic Studies, 23(3): 165–80.