The Journal of
Economic Perspectives

A journal of the
American Economic Association

Spring 2017

The Journal of
Economic Perspectives

A journal of the American Economic Association

Editorial offices:
Journal of Economic Perspectives
American Economic Association Publications
2403 Sidney St., #260
Pittsburgh, PA 15203
email: jep@jepjournal.org

The *Journal of Economic Perspectives* gratefully acknowledges the support of Macalester College.
Registered in the US Patent and Trademark Office (®).

Composed by American Economic Association Publications, Pittsburgh, Pennsylvania, USA.
Printed by LSC Communications, Owensville, Missouri, 65066, USA.
No responsibility for the views expressed by the authors in this journal is assumed by the editors or by the American Economic Association.
THE JOURNAL OF ECONOMIC PERSPECTIVES (ISSN 0895-3309), Spring 2017, Vol. 31, No. 2. The *JEP* is published quarterly (February, May, August, November) by the American Economic Association, 2014 Broadway, Suite 305, Nashville, TN 37203-2418. Annual dues for regular membership are $20.00, $30.00, or $40.00 depending on income; for an additional $15.00, you can receive this journal in print. E-reader versions are free. For details and further information on the AEA go to https://www.aeaweb.org/. Periodicals postage paid at Nashville, TN, and at additional mailing offices.
POSTMASTER: Send address changes to the *Journal of Economic Perspectives*, 2014 Broadway, Suite 305, Nashville, TN 37203.   Printed in the U.S.A.

The Journal of
Economic Perspectives

Contents

Volume 31 • Number 2 • Spring 2017

Symposia

Recent Ideas in Econometrics

Are Measures of Economic Growth Biased?

Articles

Features

Statement of Purpose

The *Journal of Economic Perspectives* attempts to fill a gap between the general interest press and most other academic economics journals. The journal aims to publish articles that will serve several goals: to synthesize and integrate lessons learned from active lines of economic research; to provide economic analysis of public policy issues; to encourage cross-fertilization of ideas among the fields of economics; to offer readers an accessible source for state-of-the-art economic thinking; to suggest directions for future research; to provide insights and readings for classroom use; and to address issues relating to the economics profession. Articles appearing in the journal are normally solicited by the editors and associate editors. Proposals for topics and authors should be directed to the journal office, at the address inside the front cover.

Policy on Data Availability

It is the policy of the *Journal of Economic Perspectives* to publish papers only if the data used in the analysis are clearly and precisely documented and are readily available to any researcher for purposes of replication. Details of the computations sufficient to permit replication must be provided. The Editor should be notified at the time of submission if the data used in a paper are proprietary or if, for some other reason, the above requirements cannot be met.

Policy on Disclosure

Authors of articles appearing in the *Journal of Economic Perspectives* are expected to disclose any potential conflicts of interest that may arise from their consulting activities, financial interests, or other nonacademic activities.

Journal of Economic Perspectives—Volume 31, Number 2—Spring 2017—Pages 3–32

The State of Applied Econometrics: Causality and Policy Evaluation

Susan Athey and Guido W. Imbens

The gold standard for drawing inferences about the effect of a policy is a randomized controlled experiment. However, in many cases, experiments remain difficult or impossible to implement, for financial, political, or ethical reasons, or because the population of interest is too small. For example, it would be unethical to prevent potential students from attending college in order to study the causal effect of college attendance on labor market experiences, and politically infeasible to study the effect of the minimum wage by randomly assigning minimum wage policies to states. Thus, a large share of the empirical work in economics about policy questions relies on observational data—that is, data where policies were determined in a way other than through random assignment. Drawing inferences about the causal effect of a policy from observational data is quite challenging. To understand the challenges, consider the example of the minimum wage. A naive analysis of the observational data might compare the average employment level of states with a high minimum wage to that of states with a low minimum wage. This difference is surely *not* a credible estimate of the causal effect of a higher minimum wage, defined as the change in employment that would occur if the low-wage states raised their minimum wage. For example, it might be the case that states with higher costs of living, as well as more price-insensitive consumers, choose higher levels of the minimum wage

■ *Susan Athey is Economics of Technology Professor and Guido W. Imbens is Applied Econometrics Professor and Professor of Economics, both at the Graduate School of Business, Stanford University, Stanford, California. Both authors are also Research Associates, National Bureau of Economic Research, Cambridge, Massachusetts. Their email addresses are athey@stanford.edu and imbens@stanford.edu.*

† For supplementary materials such as appendices, datasets, and author disclosure statements, see the article page at
https://doi.org/10.1257/jep.31.2.3 doi=10.1257/jep.31.2.3

compared to states with lower costs of living and more price-sensitive consumers. These factors, which may be unobserved, are said to be "confounders," meaning that they induce correlation between minimum wage policies and employment that is not indicative of what would happen if the minimum wage policy changed.

In economics, researchers use a wide variety of strategies for attempting to draw causal inference from observational data. These strategies are often referred to as *identification strategies* or *empirical strategies* (Angrist and Krueger 1999), because they are strategies for identifying the causal effect. We say, somewhat loosely, that a causal effect is identified if it can be learned when the dataset is sufficiently large. In the first main section of the paper, we review developments corresponding to several of these identification strategies: regression discontinuity, synthetic control and differences-in-differences methods, methods designed for networks settings, and methods that combine experimental and observational data. In the next main section, we discuss *supplementary analyses*, by which we mean analyses where the results are intended to convince the reader of the credibility of the primary analyses. These supplementary analyses have not always been systematically applied in the empirical literature, but we believe they will be of growing importance. We then briefly discuss some new developments in the machine learning literature, which focus on the combination of predictive methods and causal questions. We argue that machine learning methods hold great promise for improving the credibility of policy evaluation, and they can also be used to approach supplementary analyses more systematically.

Overall, this article focuses on recent developments in econometrics that may be useful for researchers interested in estimating the effect of policies on outcomes. Our choice of topics and examples does not seek to be an overall review. Instead it is selective and subjective, based on our reading and assessment of recent research.

New Developments in Program Evaluation

The econometric literature on estimating causal effects has been very active for over three decades now. Since the early 1990s, the *potential outcome* approach, sometimes referred to as the Rubin Causal Model, has gained substantial acceptance as a framework for analyzing causal problems.[1] In the potential outcome approach, there is for each unit i and each level of the treatment w, a potential outcome $Y_i(w)$, which describes the value of the outcome under treatment level w for that unit. Researchers observe which treatment a given unit received and the corresponding outcome for each unit, but because we do not observe the outcomes for other levels of the treatment that a given unit did not receive, we can never directly observe the causal effects, which is what Holland (1986) calls the "fundamental problem of causal inference." Estimates of causal effects are ultimately based on comparisons of different units with different levels of the treatment.

[1] There is a complementary approach based on graphical models (for example, Pearl 2000) that is widely used in other disciplines.

In some settings, the goal is to analyze the effect of a binary treatment, and the *unconfoundedness assumption* can be justified. This assumption requires that all "confounding factors" (that is, factors correlated with both potential outcomes and with the assignment to the treatment) are observed, which in turn implies that conditional on observed confounders, the treatment is as good as randomly assigned. Rosenbaum and Rubin (1983a) show that under this assumption, the average difference between treated and untreated groups with the same values for the confounders can be given a causal interpretation. The literature on estimating average treatment effects under unconfoundedness is very mature, with a number of competing estimators and many applications. Some estimators use matching methods (where each treated unit is compared to control units with similar covariates), some rely on reweighting observations so that the observable characteristics of the treatment and control group are similar after weighting, and some involve the propensity score (that is, the conditional probability of receiving the treatment given the covariates) (for reviews, see Imbens 2004; Abadie and Imbens 2006; Imbens and Rubin 2015; Heckman and Vytlacil 2007). Because this setting has been so well studied, we do not cover it in this article; neither do we cover the voluminous (and very influential) literature on instrumental variables.[2] Instead, we discuss issues related to a number of other identification strategies and settings.

Regression Discontinuity Designs

A regression discontinuity design enables the estimation of causal effects by exploiting discontinuities in incentives or ability to receive a discrete treatment.[3] For example, school district boundaries may imply that two children whose houses are on the same street will attend different schools, or birthdate cutoffs may limit eligibility to start kindergarten between two children born only a few days apart. Many government programs are means-tested, meaning that eligibility depends on income falling below a threshold. In these settings, it is possible to estimate the causal effect of attending a particular school or receiving a government program by comparing outcomes for children who live on either side of the boundary, or by comparing individuals on either side of an eligibility threshold.

[2] There are two recent strands of the instrumental variables literature. One focuses on heterogenous treatment effects, with a key development being the notion of the local average treatment effect (Imbens and Angrist 1994; Angrist, Imbens, and Rubin 1996). This literature has been reviewed in Imbens (2014). There is also a literature on weak instruments, focusing on settings with a possibly large number of instruments and weak correlation between the instruments and the endogenous regressor. On this topic, see Bekker (1994), Staiger and Stock (1997), and Chamberlain and Imbens (2004) for specific contributions, and Andrews and Stock (2006) for a survey. Also, we also do not discuss in detail bounds and partial identification analyses. Starting with the work by Manski (for instance, Manski 1990), these topics have received a lot of interest, with an excellent recent review in Tamer (2010).

[3] This approach has a long history, dating back to work in psychology in the 1950s by Thistlewaite and Campbell (1960), but did not become part of the mainstream economics literature until the early 2000s (with an exception being Goldberger 1972, 2008). Fairly recent reviews include Imbens and Lemieux (2008), Lee and Lemieux (2010), van der Klaauw (2008), and Skovron and Titiunik (2015).

In general, the key feature of the design is the presence of an exogenous variable, referred to as the *forcing variable*, like the student's birthday or address, where the probability of participating in the program changes discontinuously at a threshold value of the forcing variable. This design can be used to estimate causal effects under the assumption that the individuals close to the threshold but on different sides are otherwise comparable, so any difference in average outcomes between individuals just to one side or the other can be attributed to the treatment. If the jump in the conditional probability of treatment at the threshold value is from zero to one, we refer to the design as a "sharp" regression discontinuity design. In this case, a researcher can focus on the discontinuity of the conditional expectation of the outcome given the forcing variable at the threshold, interpreted as the average effect of the treatment for individuals close to the threshold. If the magnitude of the jump in probability of receiving the treatment at the threshold value is less than one, it is a "fuzzy" regression discontinuity design. For example, some means-tested government programs are also rationed, so that not all eligible people gain access. In this case, the focus is again on the discontinuity in the conditional expectation of the outcome at the threshold, but now it must be scaled by the discontinuity in the probability of receiving the treatment. The interpretation of the estimand is the average effect for "compliers" at the threshold, that is, individuals at the threshold whose treatment status would have been different had they been on the other side of the threshold (Hahn, Todd, and van der Klaauw 2001).

Let us illustrate a regression discontinuity design with data from Jacob and Lefgren (2004). They study the causal effect of attending summer school using administrative data from the Chicago Public Schools, which in 1996 instituted an accountability policy that tied summer school attendance and promotional decisions to performance on standardized tests. We use the data for 70,831 third-graders in years 1997–99. The rule was that individuals who scored below a threshold (2.75 in this case) on either reading or mathematics were required to attend summer school. Out of the 70,831 third graders, 15,846 scored below the threshold on the mathematics test, 26,833 scored below the threshold on the reading test, 12,779 scored below the threshold on both tests, and 29,900 scored below the threshold on at least one test. The outcome variable Y_i^{obs} is the math score after the summer school, normalized to have variance one. Table 1 presents some of the results. The first row presents an estimate of the effect of summer school attendance on the mathematics test, using for the forcing variable the minimum of the initial mathematics score and the initial reading score. We find that the summer school program has a substantial effect, raising the math test outcome score by 0.18 standard deviations.

Researchers who are implementing a regression discontinuity approach might usefully bear four pointers in mind. First, we recommend using *local linear* methods for the estimation process, rather than *local constant* methods that simply attempt to estimate average outcomes on either side of the boundary using a standard kernel regression. A kernel regression predicts the average outcome at a point by taking a weighted average of outcomes for nearby observations, where closer observations are weighted more highly. The problem is that when applying such a method near a

Table 1

Regression Discontinuity Designs: The Jacob–Lefgren Data

Outcome	*Sample*	*Estimator*	*Estimate*	*Standard error*	*IK Bandwidth*
Math	All	Local Linear	0.18	(0.02)	0.57
Math	Reading > 3.32	Local Linear	0.15	(0.02)	0.57
Math	Math > 3.32	Local Linear	0.17	(0.03)	0.57
Math	Math and Reading < 3.32	Local Linear	0.19	(0.02)	0.57
Math	All	Local Constant	−0.15	(0.02)	0.57

Note and Source: This table illustrates a regression discontinuity design with data from Jacob and Lefgren (2004). They study the causal effect of attending summer school, using use administrative data from the Chicago Public Schools, which in 1996 instituted an accountability policy that tied summer school attendance and promotional decisions to performance on standardized tests. We use the data for 70,831 third-graders in years 1997–99. The rule was that individuals who scored below a threshold (2.75 in this case) on either a reading or mathematics were required to attend summer school. Out of the 70,831 third graders, 15,846 scored below the threshold on the mathematics test, 26,833 scored below the threshold on the reading test, 12,779 score below the threshold on both tests, and 29,900 scored below the threshold on at least one test. The outcome variable Y_i^{obs} is the math score after the summer school, normalized to have variance one. The first row presents an estimate of the effect of summer school attendance on the mathematics test, using for the forcing variable the minimum of the initial mathematics score and the initial reading score. We find that the summer school program has a substantial effect, raising the math test outcome score by 0.18 standard deviations. Rows 2–4 in Table 1 present estimates for separate subsamples. In this case, we find relatively little evidence of heterogeneity in the estimates.

boundary, all of the observations lie on one side of the boundary, creating a bias in the estimates (Porter 2003). As an alternative Porter suggested *local linear regression*, which involves estimating linear regressions of outcomes on the forcing variable separately on the left and the right of the threshold, and then taking the difference between the predicted values at the threshold. This approach works better if the outcomes change systematically near the boundary because the model accounts for this and corrects the bias that arises due to truncating data at the boundary. The local linear estimator has substantially better finite sample properties than nonparametric methods that do not account for threshold effects, and it has become the standard in the empirical literature. For details on implementation, see Hahn, Todd, and van der Klaauw (2001), Porter (2003), and Calonico, Cattaneo, and Titiunik (2014a).[4]

A second key element in carrying out regression discontinuity analysis, given a local linear estimation method, is the choice of the bandwidth—that is, how to weight nearby versus more distant observations. Conventional methods for choosing optimal bandwidths in nonparametric regressions look for bandwidths that are optimal for estimating an entire regression function, but here the interest is solely in the value of the regression function at a particular point. The current literature

[4] There are some suggestions that using local quadratic methods may work well given the current technology for choosing bandwidths. Some empirical studies use global high-order polynomial approximations to the regression function, but Gelman and Imbens (2014) argue that such methods have poor properties.

suggests choosing the bandwidth for the local linear regression using asymptotic expansions of the estimators around small values for the bandwidth (Imbens and Kalyanaraman 2012; Calonico, Cattaneo, and Titiunik 2014a).

This example of summer school attendance also illustrates a situation in which the discontinuity involves multiple exogenous variables: in this case, students who score below a threshold on either a language or a mathematics test are required to attend summer school. Although not all the students who are required to attend summer school do so (a fuzzy regression discontinuity design), the fact that the forcing variable is a known function of two observed exogenous variables makes it possible to estimate the effect of summer school at different margins. For example, one can estimate the effect of summer school for individuals who are required to attend because of failure to pass the language test, and compare this with the estimate for those who are required because of failure to pass the mathematics test. The dependence of the threshold on multiple exogenous variables improves the ability to detect and analyze heterogeneity in the causal effects. Rows 2–4 in Table 1 present estimates for separate subsamples. In this case, we find relatively little evidence of heterogeneity in the estimates.

A third concern for regression discontinuity analysis is how to assess the validity of the assumptions required for interpreting the estimates as causal effects. We recommend carrying out supplementary analyses to assess the credibility of the design, and in particular to test for evidence of manipulation of the forcing variable, as well as to test for discontinuities in average covariate values at the threshold. We will discuss examples later.

Fourth, we recommend that researchers investigate the external validity of the regression discontinuity estimates by assessing the credibility of extrapolations to other subpopulations (Bertanha and Imbens 2014; Angrist and Rokkanen 2015; Angrist and Fernandez-Val 2010; Dong and Lewbel 2015). Again, we return to this topic later in the paper.

An interesting recent development in the area of regression discontinuity designs involves the generalization to discontinuities in derivatives, rather than levels, of conditional expectations. The basic idea is that at a threshold for the forcing variable, the slope of the outcome function (as a function of the forcing variable) changes, and the goal is to estimate this change in slope. The first discussions of these regression kink designs appear in Nielsen, Sorensen, and Taber (2010), Card, Lee, Pei, and Weber (2015), and Dong (2014). For example, in Card, Lee, Pei, and Weber (2015), the goal of the analysis is to estimate the causal effect of an increase in the unemployment benefits on the duration of unemployment spells, where earnings are the forcing variable. The analysis exploits the fact that, at the threshold, the relationship between benefit levels and the forcing variable changes. If we are willing to assume that in the absence of the kink in the benefit system, the derivative of the expected duration of unemployment would be smooth in lagged earnings, then the change in the derivative of the expected duration with respect to lagged earnings is informative about the relation between the expected duration and the benefit schedule.

Synthetic Control Methods and Difference-In-Differences

Difference-in-differences methods have been an important tool for empirical researchers since the early 1990s. These methods are typically used when some groups, like cities or states, experience a treatment, such as a policy change, while others do not. In this situation, the selection of which groups experience the treatment is not necessarily random, and outcomes are not necessarily the same across groups in the absence of the treatment. The groups are observed before and after the treatment. The challenge for causal inference is to come up with a credible estimate of what the outcomes would have been for the treatment group in the absence of the treatment. This requires estimating a (counterfactual) change over time for the treatment group if the treatment had not occurred. The assumption underlying difference-in-differences strategies is that the change in outcomes over time for the control group is informative about what the change would have been for the treatment group in the absence of the treatment. In general, this requires functional form assumptions. If researchers make a linearity assumption, they can estimate the average treatment effect as the difference between the change in average outcomes over time for the treatment group, minus the change in average outcomes over time for the control group.

Here we discuss two recent developments to the difference-in-differences approach: the synthetic control approach and the nonlinear changes-in-changes method. The synthetic control approach developed by Abadie, Diamond, and Hainmueller (2010, 2014) and Abadie and Gardeazabal (2003) is arguably the most important innovation in the policy evaluation literature in the last 15 years. This method builds on difference-in-differences estimation, but uses systematically more attractive comparisons. To gain some intuition about these methods, consider the classic difference-in-differences study by Card (1990; see also Peri and Yasenov 2015). Card is interested in the effect of the Mariel boatlift, which brought low-skilled Cuban workers to Miami. The question is how the boatlift affected the Miami labor market, and specifically the wages of low-skilled workers. He compares the change in the outcome of interest for the treatment city (Miami) to the corresponding change in a control city. He considers various possible control cities, including Houston, Petersburg, and Atlanta.

In contrast, the synthetic control approach moves away from using a single control unit or a simple average of control units, and instead uses a weighted average of the set of controls. In other words, instead of choosing between Houston, Petersburg, or Atlanta, or taking a simple average of outcomes in those cities, the synthetic control approach chooses weights for each of the three cities so that the weighted average is more similar to Miami than any single city would be. If pre-boatlift wages are higher in Houston than in Miami, but lower in Atlanta than Miami, it would make sense to compare Miami to the average of Houston and Atlanta rather than to either Houston or Atlanta. The simplicity of the idea, and the obvious improvement over the standard methods, have made this a widely used method in the short period of time since its inception.

The implementation of the synthetic control method requires a specific choice for the weights. The original paper, Abadie, Diamond, and Hainmueller (2010), uses a minimum distance approach, combined with the restriction that the resulting weights are nonnegative and sum to one. This approach often leads to a unique set of weights. However, if a certain unit is on the extreme end of the distribution of units, then allowing for weights that sum up to a number different from one or allowing for negative weights may improve the fit. Doudchenko and Imbens (2016) explore alternative methods for calculating appropriate weights for a synthetic control approach, such as best subset regression or LASSO (the least absolute shrinkage and selection operator) and elastic nets methods, which perform better in settings with a large number of potential control units.

Functional form assumptions can play an important role in difference-in-differences methods. For example, in the extreme case with only two groups and two periods, it is not clear whether we should assume that the percentage change over time in average outcomes would have been the same in the treatment and control groups in the absence of the treatment, or whether we should assume that the level of the change over time would have been the same. In general, a treatment might affect both the mean and the variance of outcomes, and the impact of the treatment might vary across individuals.

For the case where the data includes repeated cross-sections of individuals (that is, the data include individual observations about many units within each group in two different time periods, but the individuals cannot be linked across time periods or may come from a distinct sample such as a survey), in Athey and Imbens (2006), we propose a nonlinear version of the difference-in-differences model. This approach, which we call changes-in-changes, does not rely on functional form assumptions, while still allowing the effects of time and treatment to vary systematically across individuals. For example, one can imagine a situation in which the returns to skill are increasing over time, or in which a new medical treatment holds greater benefit for sicker individuals. The distribution of outcomes that emerges from the nonlinear difference-in-differences model is of direct interest for policy implications, beyond the average effect of the treatment itself. Further, a number of authors have used this approach as a robustness check, or what we will call in the next main section a supplementary analysis, for the results from a linear model.

Estimating Average Treatment Effects in Settings with Multivalued Treatments

Much of the earlier econometric literature on treatment effects focused on the case with binary treatments, but a more recent literature discusses the issues posed by multivalued treatment, which is of great relevance as, in practice, many treatments have multiple versions. For example, a get-out-the-vote campaign (or any advertising campaign) might consider a variety of possible messages; or a firm might consider several different price levels. In the case of a binary treatment, there are a variety of methods for estimating treatment effects under the unconfoundedness assumption, which requires that the treatment assignment is as good as random conditional on covariates. One method that works well when the number of covariates is small is to model average outcomes as a function of observed covariates, and then use the model to adjust for the extent to which differences in the treatment and control group are accounted for by observables.

However, this type of modeling performs less well if there are many covariates, or if the differences between the treatment and control group in terms of covariates are large, because errors in estimating the impact of covariates lead to large biases. An alternative set of approaches relies on the concept of a *propensity score* (Rosenbaum and Rubin 1983a), which is the probability that an individual gets a treatment, conditional on the individual's observable characteristics. In environments where unconfoundedness holds, it is sufficient to control for the propensity score (a single-dimensional variable that summarizes how observables affect the treatment probability), and it is not necessary to model outcomes as a function of all observables. That is, a comparison of two people with the same propensity score, one of whom received the treatment and one who did not, should in principle adjust for confounding variables. In practice, some of the most effective causal estimation methods in nonexperimental studies using observable data appear to be those that combine some modeling of the conditional mean of outcomes (for example, using regression adjustments) with a covariate balancing method such as subclassification, matching, or weighting based on the propensity score (Imbens and Rubin 2015), making them doubly robust (Bang and Robins 2005).

Substantially less attention has been paid to extensions of these methods to the case where the treatment takes on multiple values (exceptions include Imbens 2000; Lechner 2001; Imai and Van Dyk 2004; Cattaneo 2010; Hirano and Imbens 2004; Yang et al. 2016). However, the recent literature shows that the dimension-reducing properties of a generalized version of the propensity score, and by extension the doubly robust properties, can be maintained in the multivalued treatment setting, but the role of the propensity score is subtly different, opening up the area for empirical research in this setting. Imbens (2000) introduced the concept of a generalized propensity score, which is based on an assumption of weak unconfoundedness, requiring only that the indicator for receiving a particular level of the treatment and the potential outcome for that treatment level are conditionally independent. Weak unconfoundedness implies similar dimension-reduction properties as are available in the binary treatment case. This approach can be used to develop matching or propensity score subclassification strategies (where groups of individuals whose propensity scores lie in an interval are compared as if treatment assignment was random within the band) (for example, Yang et al. 2016). The main insight is that it is not necessary to look for subsets of the covariate space where one can interpret the difference in average outcomes by all treatment levels as estimates of causal effects. Instead, subsets of the covariate space are constructed where one can estimate the marginal average outcome for a particular treatment level as the conditional average for units with that treatment level, one treatment level at a time.

Causal Effects in Networks and Social Interactions

Peer effects, and more generally causal effects of various treatments, in networks is an important area. For example, individuals in a social network may receive information, or may gain access to a product or service, and we wish to understand the impact of that treatment both on the treated individuals, but also their peers. This area has seen much novel work in recent years, ranging from econometrics (Manski 1993) to economic theory (Jackson 2010). Here, we discuss some of the progress that has been made in econometrics. In general, this literature focuses on causal effects in settings where units, often individuals, interact in a way that violates the no-interference assumptions (more precisely, the SUTVA or Stable Unit Treatment Value Assumption as in Rosenbaum and Rubin 1983a; Imbens and Rubin 2015) that are routinely made in the treatment effects literature. In some cases, the way in which individuals interact is simply a nuisance, and the main interest continues to be on the direct causal effects of own treatments. In other cases, the magnitude of the interactions, or peer effects, is itself the subject of interest.

Networks and peer effects can operate through many scenarios, which has led to the literature becoming somewhat fractured and unwieldy. For example, there is a distinction between, on the one hand, settings where the population can be partitioned into subpopulations with all units within a subpopulation connected, as, for example, in classrooms (for example, Manski 1993; Carrell, Sacerdote, and West 2013), workers in a labor market (Crépon et al. 2013), or roommates in college (Sacerdote 2001). One can also consider settings with general networks, in which friends of friends are not necessarily friends themselves (Christakis and Fowler 2007). Another important distinction is between settings with many disconnected networks, where asymptotic arguments for consistency rely on the number of networks getting large, and settings with a single connected network. It may be reasonable in some cases to think of the links as symmetric, and in others of links operating only in one direction. Links can be binary, with links either present or not, or a network may contain links of different strengths.

A seminal paper in the econometric literature in this area focuses on Manski's linear-in-means model (Manski 1993; Bramoullé, Djebbari, and Fortin 2009; Goldsmith-Pinkham and Imbens 2013). Manski's original paper focuses on the setting where the population is partioned into groups (like classrooms), and peer effects are constant within the groups. The basic model specification is

$$Y_i = \beta_0 + \beta_{\overline{Y}} \cdot \overline{Y}_i + \beta'_X X_i + \beta'_{\overline{X}} \overline{X}_i + \beta'_Z Z_i + \varepsilon_i,$$

where i indexes the individual. Here Y_i is the outcome for individual i, say educational achievement; $\overline{Y}_i$ is the average outcome for individuals in the peer group for individual i; X_i is a set of exogenous characteristics of individual i, like prior test scores in an educational setting; $\overline{X}_i$ is the average value of the characteristics in individual i's peer group; and Z_i is a vector of group characteristics that is constant for all individuals in the same peer group, like quality of teachers in a classroom setting. Manski considers three types of peer effects that lead to correlations in outcomes between individuals. Outcomes for individuals in the same group may be correlated because of a shared environment. These effects are called correlated peer effects, and captured by the coefficient on Z_i. Next are the exogenous peer effects, captured by the coefficient on the group average $\overline{X}_i$ of the exogenous variables. The third type is the endogenous peer effect, captured by the coefficient on the group average outcomes $\overline{Y}_i$.

Manski (1993) concludes that separate identification of these three effects, even in the linear model setting with constant coefficients, relies on very strong assumptions and is unrealistic in many settings. In subsequent empirical work, researchers have often put additional structure on the effects (for example, by ruling out some of the effects) or brought in additional information (for example, by using richer network structures) to obtain identification. Graham (2008) focuses on a setting very similar to that of Manski's linear-in-means model. He considers restrictions on the within-group covariance matrix of the ε_i assuming homoskedasticity at the individual level. In that case, a key insight is that variation in group size implies restrictions on the within and between group variances that can be used to identify peer effects. Bramoullé, Djebbari, and Fortin (2009) allow for a more general network configuration than Manski, one in which friends of friends are not necessarily connected, and demonstrate the benefits of such configurations for identification of peer effects. Hudgens and Halloran (2008) start closer to the Rubin Causal Model or potential outcome setup. They focus primarily on the case with a binary treatment, and consider how the vector of treatments for the peer group affects the individual. They suggest various structures on these treatment effects that can aid in identification. Aronow and Samii (2013) allow for general networks and peer effects, investigating the identifying power from randomization of the treatments at the individual level.

Two other branches of the literature on estimation of causal effects in a context of network and peer effects are worth mentioning. One part focuses on developing models for network formation. Such approximations require the researcher to specify in what way the expanding sample would be similar to or different from the current sample, which in turn is important for deriving asymptotic approximations based on large samples. Recent examples of such work in economics include Jackson and Wolinsky (1996), Jackson (2010), Goldsmith-Pinkham and Imbens (2013), Christakis, Fowler, Imbens, and Kalyanaraman (2010), and Mele (2013). Chandrasekhar and Jackson (2016) develop a model for network formation and a corresponding central limit theorem in the presence of correlation induced by network links. Chandrasekhar (2016) surveys the general econometrics literature on network formation.

The other branch worth a mention is the use of randomization inference in the context of causal regressions involving networks, as a way of generating exact *p*-values. As an example of randomization inference, consider the null hypothesis that a treatment has no effect. Because the null of no effects is sharp (that is, if the null hypothesis is true, we know exactly what the outcomes would be in alternative treatment regimes after observing the individual in one treatment regime), it allows for the calculation of exact *p*-values. The approach works by simulating alternative (counterfactual) treatment assignment vectors and then calculating what the test statistic (for example, difference in means between treated and control units) would have been if that assignment had been the real one. This approach relies heavily on the fact that the null hypothesis is sharp, but many interesting null hypotheses are not sharp. In Athey, Eckles, and Imbens (forthcoming), we discuss a large class of alternative null hypotheses: for example, hypotheses restricting higher order peer effects (peer effects from friends-of-friends) while allowing for the presence of peer effects from friends; hypotheses about whether a dense network can be represented by a simplified or *sparsified* set of rules; and hypotheses about whether peers are exchangeable, or whether some peers have larger or different effects. To test such hypotheses, in Athey, Eckles, and Imbens (forthcoming), we introduce the notion of an artificial experiment, in which some units have their treatment assignments held fixed, and we randomize over the remaining units. The artificial experiment starts by designating an arbitrary set of units to be focal. The test statistics considered depend only on outcomes for these focal units. Given the focal units, one derives the set of assignments that does not change the outcomes for the focal units. The exact distribution of the test statistic can then be inferred despite the original null hypothesis not being sharp. This approach allows us to test hypotheses about, for example, the effect of friends-of-friends, without making additional assumptions about the network structure and without resorting to asymptotics in the size of the network.

External Validity

Even when a causal study is done carefully, both in analysis and design, there is often little assurance that the causal effects are valid for populations or settings other than those studied. This concern has been raised particularly forcefully in experimental studies (for examples, see the discussions in Deaton 2010; Imbens 2010; Manski 2013). Some have emphasized that without internal validity, little can be learned from a study (Shadish, Cook, and Cambell 2002; Imbens 2013). However, Deaton (2010), Manski (2013), and Banerjee, Chassang, and Snowberg (2016) have argued that external validity should receive more emphasis.

In some recent work, approaches have been proposed that allow researchers to directly assess the external validity of estimators for causal effects. A leading example concerns settings with instrumental variables (for example, Angrist 2004; Angrist and Fernandez-Val 2010; Dong and Lewbel 2015; Angrist and Rokkanen 2015; Bertanha and Imbens 2014; Kowalski 2016; Brinch, Mogstad, and Wiswall 2015). An instrumental variables estimator is often interpreted as an estimator of the local average treatment effect, that is, the average effect of the treatment for individuals whose treatment status is affected by the instrument. So under what conditions can these estimates be considered representative for the entire sample? In this context, one can partition the sample into several groups, depending on the effect of the instrumental variable on the receipt of the treatment. There are two groups that are unaffected by the instrumental variable: *always-taker*s, who always receive the treatment, and *never-takers,* who never receive the treatment, no matter the value of the instrumental variable. *Compliers* are those whose treatment status is affected by the instrumental variable. In that context, Angrist (2004) suggests testing whether the difference in average outcomes for always-takers and never-takers is equal to the average effect for compliers. Bertanha and Imbens (2014) suggest testing a combination of two equalities: whether the average outcome for untreated compliers is equal to the average outcome for never-takers; and whether the average outcome for treated compliers is equal to the average outcome for always-takers. Angrist and Fernandez-Val (2010) seek to exploit the presence of other exogenous covariates using *conditional effect ignorability*, which is that, conditional on these additional covariates, the average effect for compliers is identical to the average effect for never-takers and always-takers.

In the context of regression discontinuity designs, concerns about external validity are especially salient. In that setting, the estimates are in principle valid only for individuals with values of the forcing variable near the threshold. There have been a number of approaches to assess the plausibility of generalizing those local estimates to other parts of the population. Some of them apply to both sharp and fuzzy regression discontinuity designs, and some apply only to fuzzy designs. Some require the presence of additional exogenous covariates, and others rely only on the presence of the forcing variable. For example, Dong and Lewbel (2015) observe that in general, in regression discontinuity designs with a continuous forcing variable, one can estimate the magnitude of the discontinuity as well as the magnitude of the change in the first derivative of the regression function, or even higher-order derivatives, which allows one to extrapolate away from values of the forcing variable close to the threshold. In another approach, Angrist and Rokkanen (2015) suggest testing whether conditional on additional covariates, the correlation between the forcing variable and the outcome vanishes. Such a finding would imply that the treatment assignment can be thought of as unconfounded conditional on the additional covariates, which again allows for extrapolation away from the threshold. Finally, Bertanha and Imbens (2014) propose an approach based on a fuzzy regression discontinuity design. They suggest testing for continuity of the conditional expectation of the outcome conditional on the treatment and the forcing variable at the threshold, adjusted for differences in the covariates.

Leveraging Experiments

In some cases, we wish to exploit the benefits of the experimental results, in particular the high degree of internal validity, in combination with the external validity and precision from large-scale representative observational studies. Here we discuss three settings in which experimental studies can be leveraged in combination with observational studies to provide richer answers than either design could provide on its own. In the first example, the surrogate variables case, the primary outcome was not observed in the experiment, but an intermediate outcome was observed. In a second case, both the intermediate outcome and the primary outcome were observed. In the third case, multiple experiments bear on a common outcome. These examples do not exhaust the settings in which researchers can leverage experimental data more effectively, and more research in this area is likely to be fruitful.

In the case of surrogate variables, studied in Athey, Chetty, Imbens, and Kang (2016), the researcher uses an intermediate variable as a surrogate for the treatment variable. For example, in medical trials there is a long history of attempts to use intermediate health measures as surrogates (Prentice 1989). The key condition for an intermediate variable to be a valid surrogate is that, in the experimental sample, conditional on the surrogate and observed covariates, the (primary) outcomes and the treatment are independent (Prentice 1989; Begg and Leung 2000; Frangakis and Rubin 2002). In medical settings, where researchers often used single surrogates, this condition was often not satisfied in settings where it could be tested. But it may be more plausible in other settings. For example, suppose an internet company is considering a change to the user experience on the company's website. It is interested in the effect of that change on the user's purchases over a year-long period. The firm carries out a randomized experiment over a month, during which it measures details concerning the customer's engagement like the number of visits, webpages visited, and the length of time spent on the various webpages. The firm may also have historical records on user characteristics, including past engagement. The combination of the pretreatment variables and the surrogates may be sufficiently rich so that, conditional on the combination, the primary outcome is independent of the treatment.

In administrative and survey research databases used in economics, a large number of intermediate variables are often recorded that lie on or close to the causal path between the treatment and the primary outcome. In such cases, it may be plausible that the full set of surrogate variables satisfies at least approximately the independence condition. In this setting, in Athey, Chetty, Imbens, and Kang (2016), we develop multiple methods for estimating the average effect. One method corresponds to estimating the relation between the outcome and the surrogates in the observational data and using that to impute the missing outcomes in the experimental sample. Another corresponds to estimating the relation between the treatment and the surrogates in the experimental sample and using that to impute the treatment indicator in the observational sample. Yet another exploits both methods, using the efficient influence function. In the same paper, we also derive the biases from violations of the surrogacy assumption.

In the second setting for leveraging experiments, studied in Athey, Chetty, and Imbens (2016), the researcher has data from a randomized experiment, in this case containing information on the treatment and the intermediate variables, as well as pretreatment variables. In an observational study, the researcher observes the same variables plus the primary outcome. One can then compare the estimates of the average effect on the intermediate outcomes based on the observational sample, after adjusting for pretreatment variables, with those from the experimental sample. The latter are known to be consistent, and so if one finds substantial and statistically significant differences, then unconfoundedness need not hold. For that case, in Athey, Chetty, and Imbens (2016), we develop methods for adjusting for selection on unobservables, exploiting the observations on the intermediate variables.

The third setting, involving the use of multiple experiments, has not received as much attention, but provides fertile ground for future work. Consider a setting in which a number of experiments were conducted that vary in terms of the population from which the sample is drawn or in the exact nature of the treatments included. The researcher may be interested in combining these experiments to obtain more efficient estimates, perhaps for predicting the effect of a treatment in another population or estimating the effect of a treatment with different characteristics. These issues are related to external validity concerns but include more general efforts to decompose the effects from experiments into components that can inform decisions on related treatments. In the treatment effects literature, aspects of these problems have been studied in Hotz, Imbens, and Mortimer (2005), Imbens (2010), and Allcott (2015). They have also received some attention in the literature on structural modeling, where experimental data are used to anchor aspects of the structural model (for example, Todd and Wolpin 2006).

Supplementary Analyses

Primary analyses focus on point estimates of the primary estimands along with standard errors. In contrast, supplementary analyses seek to shed light on the credibility of the primary analyses. These supplementary analyses do not seek a better estimate of the effect of primary interest, nor do they (necessarily) assist in selecting among competing statistical models. Instead, the analyses exploit the fact that the assumptions behind the identification strategy often have implications for the data beyond those exploited in the primary analyses. Supplementary analyses can take on a variety of forms, and we are not aware of a comprehensive survey to date. This literature is very active, both in theoretical and empirical studies and likely to be growing in importance in the future. Here, we discuss some examples from the empirical and theoretical literatures, which we hope provide some guidance for future work.

We will discuss four forms of supplementary analysis: 1) placebo analysis, where pseudo-causal effects are estimated that are known to be equal to zero based on a priori knowledge; 2) sensitivity and robustness analyses that assess how much estimates of the primary estimands can change if we weaken the critical assumptions underlying the primary analyses; 3) identification and sensitivity analyses that highlight what features of the data identify the parameters of interest; and 4) a supplementary analysis that is specific to regression discontinuity analyses, in which the focus is on whether the density of the forcing variable is discontinuous at the threshold, which would suggest that the forcing variable is being manipulated.

Placebo Analyses

In a placebo analysis, the most widely used of the supplementary analyses, the researcher replicates the primary analysis with the outcome replaced by a pseudo-outcome that is known not to be affected by the treatment. Thus, the true value of the estimand for this pseudo-outcome is zero, and the goal of the supplementary analysis is to assess whether the adjustment methods employed in the primary analysis, when applied to the pseudo-outcome, lead to estimates that are close to zero. These are not standard specification tests that suggest alternative specifications when the null hypothesis is rejected. The implication of rejection here is that it is possible the original analysis was not credible at all.

One type of placebo test relies on treating lagged outcomes as pseudo-outcomes. Consider, for example, the dataset assembled by Imbens, Rubin, and Sacerdote

(2001), which studies participants in the Massachusetts state lottery. The treatment of interest is an indicator for winning a big prize in the lottery (with these prizes paid out over a 20-year period), with the control group consisting of individuals who won one small, one-time prize. The estimates of the average treatment effect rely on an unconfoundedness assumption, namely that the lottery prize is as good as randomly assigned after taking out associations with some pre-lottery variables: for example, these variables include six years of lagged earnings, education measures, gender, and other individual characteristics. Unconfoundedness is certainly a plausible assumption here, given that the winning lottery ticket is randomly drawn. But there is no guarantee that unconfoundedness holds. The two primary reasons are: 1) there is only a 50 percent response rate for the survey; and 2) there may be differences in the rate at which individuals buy lottery tickets. To assess unconfoundedness, it is useful to estimate the average causal effect with pre-lottery earnings as the outcome. Using the actual outcome, we estimate that winning the lottery (with on average a $20,000 yearly prize), reduces average post-lottery earnings by $5,740, with a standard error of $1,400. Using the pseudo-outcome we obtain an estimate of minus $530, with a standard error of $780. This finding, along with additional analyses, strongly suggests that nonconfoundedness holds.

However, using the same placebo analysis approach with the LaLonde (1986) data on job market training that are widely used in the econometric evaluation literature (for example, Heckman and Hotz 1989; Dehejia and Wahba 1999; Imbens 2015), the results are quite different. Imbens (2015) uses 1975 (pretreatment) earnings as the pseudo-outcome, leaving only a single pretreatment year of earnings to adjust for the substantial difference between the trainees and comparison group from the Current Population Survey. Imbens first tests whether the simple average difference in adjusted 1975 earnings is zero. Then he tests whether both the level of 1975 earnings and the indicator for positive 1975 earnings are different in the trainees and the control groups, using separate tests for individuals with zero and positive 1974 earnings. The null is clearly rejected, casting doubt on the unconfoundedness assumption.

Placebo approaches can also be used in other contexts, like regression discontinuity design. Covariates typically play only a minor role in the primary analyses there, although they can improve precision (Imbens and Lemieux 2008; Calonico, Cattaneo, and Titiunik 2014a, b). However, these exogenous covariates can play an important role in assessing the plausibility of the regression discontinuity design. According to the identification strategy, they should be uncorrelated with the treatment when the forcing variable is close to the threshold. We can test this assumption, for example by using a covariate as the pseudo-outcome in a regression discontinuity analysis. If we were to find that the conditional expectation of one of the covariates is discontinuous at the threshold, such a discontinuity might be interpreted as evidence for an unobserved confounder whose distribution changes at the boundary, one which might also be correlated with the outcome of interest. We can illustrate this application with the election data from Lee (2008), who is interested in estimating the effect of incumbency on electoral outcomes. The treatment is a Democrat winning a congressional election, and the forcing variable is the

Democratic vote share minus the Republication vote share in the current election, and so the threshold is zero. We look at an indicator for winning the next election as the outcome. As a pretreatment variable, we consider an indicator for winning the previous election to the one that defines the forcing variable. Our estimates for the actual outcome (winning the next election) are substantially larger than those for the pseudo-outcome (winning the previous election), where we cannot reject the null hypothesis that the effect on the pseudo-outcome is zero.

One final example of the use of placebo regressions is Rosenbaum (1987), who is interested in the causal effect of a binary treatment and focuses on a setting with multiple comparison groups (see also Heckman and Hotz 1989; Imbens and Rubin 2015). In Rosenbaum's case, there is no strong reason to believe that one of the comparison groups is superior to another. Rosenbaum proposes testing equality of the average outcomes in the two comparison groups after adjusting for pretreatment variables. If one finds that there are substantial differences left after such adjustments, it shows that at least one of the comparison groups is not valid, which makes the use of either of them less credible. In applications to evaluations of labor market programs, one might implement such methods by comparing a control group of individuals who are eligible but choose not to participate with another control group of individuals who are not eligible, as in Heckman and Hotz (1989). The biases from evaluations based on the first control group might correspond to differences in motivation, whereas evaluations based on the second control group could be biased because of direct associations between eligibility criteria and outcomes.

Robustness and Sensitivity

The classical frequentist statistical paradigm suggests that a researcher specifies a single statistical model, estimates this model on the data, and reports estimates and standard errors. This is of course far from common practice, as pointed out, for example, in Leamer (1978, 1983). In practice, researchers consider many specifications and perform various specification tests before settling on a preferred model. Standard practice in modern empirical work is to present in the final paper estimates of the preferred specification of the model in combination with assessments of the robustness of the findings from this preferred specification. These alternative specifications are intended to convey that the substantive results of the preferred specification are not sensitive to some of the choices in that specification, like using different functional forms of the regression function or alternative ways of controlling for differences in subpopulations.

Some recent work has sought to make these efforts at assessing robustness more systematic. In Athey and Imbens (2015), we propose one approach to this problem, which we illustrate here in the context of regression analyses, although it can also be applied to more complex nonlinear or structural models. In the regression context, suppose that the object of interest is a particular regression coefficient that has an interpretation as a causal effect. We suggest considering a set of different specifications based on splitting the sample into two subsamples, and estimating them separately. (Specifically, we suggest splitting the original sample once for each of the elements of the original covariate vector Z_i, and splitting at a threshold that optimizes fit by minimizing the sum of squared residuals.) The original causal effect is then estimated as a weighted average of the estimates from the two split specifications. If the original model is correct, the augmented model still leads to a consistent estimator for the estimand. Notice that the focus is *not* on finding an alternative specification that may provide a better fit; rather, it is on assessing whether the estimate in the original specification is robust to a range of alternative specifications. This approach has some weaknesses. For example, adding irrelevant covariates to the procedure might decrease the standard deviation of estimates. If there are many covariates, some form of dimensionality reduction may be appropriate prior to estimating the robustness measure. Refining and improving this approach is an interesting direction for future work. For example, the theoretical literature has developed many estimators in the setting with unconfoundedness. Some rely on estimating the conditional mean, others rely on estimating the propensity score, and still others rely on matching on the covariates or the propensity score (for a review of this literature, see Imbens and Wooldridge 2009). We recommend that researchers should report estimates based on a variety of methods to assess robustness, rather than estimates based on a single preferred method.

In combination with reporting estimates based on the preferred specification, it may be useful to report ranges of estimates based on substantially weaker assumptions. For example, Rosenbaum and Rubin (1983b, see also Rosenbam 2002) suggest starting with a restrictive specification, and then assessing the changes in the estimates that result from small to modest relaxations of the key identifying assumptions such as unconfoundedness. In the context Rosenbaum and Rubin consider, that of estimating average treatment effects under selection on observables, they allow for the presence of an unobserved covariate that should have been adjusted for in order to estimate the average effect of interest. They explore how strong the correlation between this unobserved covariate and the treatment, and the correlation between the unobserved covariate and the potential outcomes, would have to be in order to substantially change the estimate for the average effect of interest. Imbens (2003) builds on the Rosenbaum and Rubin approach by developing a data-driven way to obtain a set of correlations between the unobserved covariates and treatment and outcome.

In other work along these lines, Arkhangelskiy and Drynkin (2016) study sensitivity of the estimates of the parameters of interest to misspecification of the model governing the nuisance parameters. Tamer (2010) reviews how to assess robustness based on the partial indentification or bounds literature originating with Manski (1990).

Altonji, Elder, and Taber (2008) and Oster (2015) focus on the correlation between the unobserved component in the relation between the outcome and the treatment and observed covariates, and the unobserved component in the relation between the treatment and the observed covariates. In the absence of functional form assumptions, this correlation is not identified. These papers therefore explore the sensitivity to fixed values for this correlation, ranging from the case where the correlation is zero (and the treatment is exogenous), to an upper limit chosen to match the correlation found between the observed covariates in the two regression functions. Oster takes this further by developing estimators based on this equality. This useful approach provides the researcher with a systematic way of doing the sensitivity analyses that are routinely done in empirical work, but often in an unsystematic way.

Identification and Sensitivity

Gentzkow and Shapiro (2015) take a different approach to sensitivity. They propose a method for highlighting what statistical relationships in a dataset are most closely related to parameters of interest. Intuitively, the idea is that simple correlations between particular combinations of variables identify particular parameters. To operationalize this, they investigate, in the context of a given model, how the key parameters of interest relate to a set of summary statistics. These summary statistics would typically include easily interpretable functions of the data such as correlations between subsets of variables. Under mild conditions, the joint distribution of the model parameters and the summary statistics should be jointly normal in large samples. If the summary statistics are in fact asymptotically sufficient for the model parameters, the joint distribution of the parameter estimates and the summary statistics will be degenerate. More typically, the joint normal distribution will have a covariance matrix with full rank. For example, when estimating the average causal effect of a binary treatment under unconfoundedness, one would expect the parameter of interest to be closely related to the correlation between the outcome and the treatment, and, in addition, to the correlations between some of the additional covariates and the outcome, or to the correlations between some of those covariates and the treatment. Gentzkow and Shapiro discuss how to interpret the covariance matrix in terms of sensitivity of model parameters to model specification. More broadly, their approach is related to proposals in different settings by Conley, Hansen, and Rossi (2012) and Chetty (2009).

Supplementary Analyses in Regression Discontinuity Designs

One of the most interesting supplementary analyses is the McCrary (2008) test in regression discontinuity designs (see also Otsu, Xu, and Matsushita 2013). What makes this analysis particularly interesting is the conceptual distance between the primary analysis and the supplementary analysis. The McCrary test assesses whether there is a discontinuity in the density of the forcing variable at the threshold. In a conventional analysis, it is unusual that the marginal distribution of a variable that is assumed to be exogenous is of any interest to the researcher: often, the entire analysis is conducted conditional on such regressors. However, the identification strategy underlying regression discontinuity designs relies on the assumption that units just to the left and just to the right of the threshold are comparable. That argument is difficult to reconcile if, say, there are substantially more units just to the left than just to the right of the threshold. Again, even though such an imbalance could easily be taken into account in the estimation, in many cases where one would find such an imbalance, it would suggest that the forcing variable is not a characteristic exogenously assigned to individuals, but rather that it is being manipulated in some way.

The classic example is that of an educational regression discontinuity design where the forcing variable is a test score. If the individual grading the test is aware of the importance of exceeding the threshold, and in particular if graders know the student personally, they may assign scores differently than if they were not aware of this. If there was such manipulation of the score, there would likely be a discontinuity in the density of the forcing variable at the threshold; there would be no reason to change the grade for an individual scoring just above the threshold.

Machine Learning and Econometrics

Supervised machine learning focuses primarily on prediction problems: given a dataset with data on an outcome Y_i, which can be discrete or continuous, and some predictors X_i, the goal is to estimate a model on a subset of the data, given the values of the predictors X_i. This subset is called the *training sample,* and it is used for predicting outcomes in the remaining data, which is called the *test sample.* Note that this approach is fundamentally different from the goal of causal inference in observational studies, where we observe data on outcomes and a treatment variable, and we wish to draw inferences about potential outcomes. Kleinberg, Ludwig, Mullainathan, and Obermeyer (2015) argue that many important policy problems are fundamentally prediction problems; see also the article by Mullainathan and Spiess in this issue. A second class of problems, *unsupervised machine learning*, focuses on methods for finding patterns in data, such as groups of similar items, like clustering images into groups, or putting text documents into groups of similar documents. The method can potentially be quite useful in applications involving text, images, or other very high-dimensional data, even though these approaches have not had too much use in the economics literature so far. For an exception, see Athey, Mobius, and Pal (2016) for an example in which unsupervised learning is used to categorize newspaper articles into topics.

An important difference between many (but not all) econometric approaches and supervised machine learning is that supervised machine learning methods typically rely on data-driven model selection, most commonly through cross-validation, and often the main focus is on prediction performance without regard to the implications for inference. For supervised learning methods, the sample is split into a training sample and a test sample, where, for example, the test sample might have 10 percent of observations.

The training sample is itself partitioned into a number of subsamples, or cross-validation samples, often 10 of them. For each subsample, the cross-validation sample m is set aside. The remainder of the training sample is used for estimation. The estimation results are then used to predict outcomes for the left-out subsample m. The final choice of the tuning parameter is the one that minimizes the sum of the squared residuals in the cross-validation samples. Ultimate model performance is assessed by calculating the mean-squared error of model predictions (that is, the sum of squared residuals) on the held-out test sample, which was not used at all for model estimation or tuning. Predictions from these machine learning methods are not typically unbiased, and estimators may not be asymptotically normal and centered around the estimand. Indeed, the machine learning literature places little emphasis on asymptotic normality, and when theoretical properties are analyzed, they often take the forms of worst-case bounds on risk criteria. However, the fact that model performance (in the sense of predictive accuracy on a test set) can be directly measured makes it possible to compare predictive models even when their asymptotic properties are not understood. Enormous progress has been made in the machine learning literature in terms of developing models that do well (according to the stated criteria) in real-world datasets. Here, we focus primarily on problems of causal inference, showing how supervised machine learning methods improve the performance of causal analysis, particularly in cases with many covariates.

Machine Learning Methods for Average Causal Effects

In recent years, researchers have used machine learning methods to help them control in a flexible manner for a large number of covariates. Some of these methods involved adaptions of methods used for the few-covariate case: for example, use of the weighting approach in Hirano, Imbens, Ridder, and Rubin (2001) in combination with machine learning methods such as LASSO and random forests for estimating the propensity score as in McCaffrey, Ridgeway, and Morral (2004) and Wyss et al. (2014). Such methods have relatively poor properties in many cases because they do not necessarily emphasize the covariates that are important for the bias, that is, those that are correlated both with the outcomes and the treatment indicator. More promising methods would combine estimation of the association between the potential outcomes and the covariates, and of the association between the treatment indicator and the covariates. Here we discuss three approaches along these lines (see also Athey, Imbens, Pham, and Wager 2017).

First, Belloni, Chernozhukov, Fernández, and Hansen (2013) propose a double selection procedure, where they first use a LASSO regression to select covariates that are correlated with the outcome, and then again to select covariates that are correlated with the treatment. In a final ordinary least squares regression, they include the union of the two sets of covariates, improving the properties of the estimators for the average treatment effect compared to simple regularized regression of the outcome on the covariates and the treatment.

A second line of research has focused on finding weights that directly balance covariates or functions of the covariates between treatment and control groups, so that once the data has been reweighted, it mimics a randomized experiment more closely. In the literature with few covariates, this approach has been developed in Hainmueller (2012) and Graham, Pinto, and Egel (2012, 2016); for discussion of the case with many covariates, some examples include Zubizarreta (2015) and Imai and Ratkovic (2014). In Athey, Imbens, and Wager (2016), we develop an estimator that combines the balancing with regression adjustment. The idea is that, in order to predict the counterfactual outcomes that the treatment group would have had in the absence of the treatment, it is necessary to extrapolate from control observations. By rebalancing the data, the amount of extrapolation required to account for differences between the two groups is reduced. To capture remaining differences, the regularized regression just mentioned can be used to model outcomes in the absence of the treatment. In effect, the Athey et al. estimator balances the bias coming from imbalance between the covariates in the treated subsample and the weighted control subsample, with the variance from having excessively variable weights.

A third approach builds on the semiparametric literature on influence functions. In general, van der Vaart (2000) suggests estimating the finite dimensional component as the average of the influence function, with the infinite dimensional components estimated nonparametrically. In the context of estimation of average treatment effects this leads to "doubly robust estimators" in the spirit of Robins and Rotnitzky (1995), Robins, Rotnitzky, and Zhao (1995), and van der Laan and Rubin (2006). Chernozhukov et al. (2016) propose using machine learning methods for the infinite dimensional components and incorporate sample-splitting to further improve the properties.

In all three cases, procedures for trimming the data to eliminate extreme values of the estimated propensity score (as in Crump, Hotz, Imbens, and Mitnik 2009) remain important in practice.

Machine Learning for Heterogenous Causal Effects

In many cases, a policy or treatment might have different costs and benefits if applied in different settings. Gaining insight into the nature of such heterogenous treatment effects can be useful. Moreover, in evaluating a policy or treatment, it is useful to know the applications where the benefit/cost ratios are most favorable. However, when machine learning methods are applied to estimating heterogenous treatment effects, they in effect search over many covariates and subsets of the covariate space for the best fit. As a result, such methods may lead to spurious findings of treatment effect differences. Indeed, in clinical medical trials, pre-analysis plans must be registered in advance to avoid the problem that researchers will be tempted to search among groups of the studied population to find one that seems to be affected by the treatment, and may instead end up with spurious findings. In the social sciences, the problem of searching across groups becomes more severe when there are many covariates.

One approach to this problem is to search exhaustively for treatment effect heterogeneity and then correct for issues of multiple hypothesis testing, by which we mean the problems that arise when a researcher considers a large number of statistical hypotheses, but analyzes them as if only one had been considered. This can lead to *false discovery*, because across many hypothesis tests, we expect some to be rejected even if the null hypothesis is true. To address this problem, List, Shaikh, and Xu (2016) propose to give each covariate a "low" or "high" discrete value, and then loop through the covariates, testing whether the treatment effect is different when the covariate is low versus high. Because the number of covariates may be large, standard approaches to correcting for multiple testing may severely limit the power of a (corrected) test to find heterogeneity. List et al. propose an approach based on bootstrapping that accounts for correlation among test statistics; this approach can provide substantial improvements over standard multiple testing approaches when the covariates are highly correlated, because dividing the sample according to each of two highly correlated covariates results in substantially the same division of the data. However, this approach has the drawback that the researcher must specify in advance all of the hypotheses to be tested, along with alternative ways to discretize covariates and flexible interactions among covariates. It may not be possible to explore these combinations fully.

A different approach is to adapt machine learning methods to discover particular forms of heterogeneity by seeking to identify subgroups that have different treatment effects. One example is to examine within subgroups in cases where eligibility for a government program is determined according to criteria that can be represented in a decision tree, similar to the situation when a doctor uses a decision tree to determine whether to prescribe a drug to a patient. Another example is to examine within subgroups in cases where an algorithm uses a table to determine which type of user interface, offer, email solicitation, or ranking of search results to provide to a user. Subgroup analysis has long been used in medical studies (Foster, Taylor, and Ruberg 2011), but it is often subject to criticism due to concerns of multiple hypothesis testing (Assmann, Pocock, Enos, and Kasten 2000).

Among the more common machine learning methods, regression trees are a natural choice for partitioning into subgroups (the classic reference is Breiman, Friedman, Stone, and Olshen 1984). Consider a regression with two covariates. The value of each covariate can be split so that it is above or below a certain level. The regression tree approach would consider which covariate should be split, and at which level, so that the sum of squared residuals is minimized. With many covariates, these steps of choosing which covariate to split, and where to split it, are carried out sequentially, thus resulting in a tree format. The tree eventually results in a partition of the data into groups, defined according to values of the covariates, where each group is referred to as a leaf. In the simplest version of a regression tree, we would stop this splitting process once the reduction in the sum of squared residuals is below a certain level.

In Athey and Imbens (2016), we develop a method that we call *causal trees*, which builds on earlier work by Su et al. (2009) and Zeileis, Hothorn, and Hornik (2008). The method is based on the machine learning method of regression trees, but it uses a different criterion for building the tree: rather than focusing on improvements in mean-squared error of the prediction of outcomes, it focuses on mean-squared error of treatment effects. The method relies on sample splitting, in which half the sample is used to determine the optimal partition of the covariates space (the tree structure), while the other half is used to estimate treatment effects within the leaves. The output of the method is a treatment effect and a confidence interval for each subgroup. In Athey and Imbens (2016), we highlight the fact that the criteria used for tree construction should differ when the goal is to estimate treatment effect heterogeneity rather than heterogeneity in outcomes. After all, the factors that affect the level of outcomes might be quite different from those that affect treatment effects. Although the sample-splitting approach may seem extreme—ultimately only half the data is used for estimating treatment effects—it has several advantages. The confidence intervals are valid no matter how many covariates are used in estimation. In addition, the researcher is free to estimate a more complex model in the second part of the data, for example, if the researcher wishes to include fixed effects in the model, or model different types of correlation in the error structure.

A disadvantage of the causal tree approach is that the estimates are not personalized for each individual; instead, all individuals assigned to a given group have the same estimate. For example, a leaf might contain all male individuals aged 60 to 70, with income above $50,000. An individual whose covariates are near the boundary, for example a 70 year-old man with income of $51,000, might have a treatment effect that is different than the average for the whole group. For the problem of more personalized prediction, Wager and Athey (2015) propose a method for estimating heterogeneous treatment effects based on random forest analysis, where the method generates many different trees and averages the result, except that the component trees are now causal trees (and in particular, each individual tree is estimated using sample splitting, where one randomly selected subsample is used to build the tree while a distinct subsample is used to estimate treatment effects in each leaf). Relative to a causal tree, which identifies a partition and estimates treatment effects within each element of the partition, the causal forest leads to estimates of causal effects that change more smoothly with covariates, and in principle every individual has a distinct estimate. Random forests are known to perform very well in practice for prediction problems, but their statistical properties were less well understood until recently. Wager and Athey show that the predictions from causal forests are asymptotically normal and centered on the true conditional average treatment effect for each individual. They also propose an estimator for the variance, so that confidence intervals can be obtained. Athey, Tibshirani, and Wager (2016) extend the approach to other models for causal effects, such as instrumental variables, or other models that can be estimated using the generalized method of moments (GMM). In each case, the goal is to estimate how a causal parameter of interest varies with covariates.

An alternative approach, closely related, is based on Bayesian Additive Regression Trees (BART) (Chipman, George, and McCulloch 2010), which is essentially a Bayesian version of random forests. Hill (2011) and Green and Kern (2012) apply these methods to estimate heterogeneous treatment effects. Large-sample properties of this method are unknown, but it appears to have good empirical performance in applications.

Other machine-based approaches, like the LASSO regression approach, have also been used in estimating heterogenous treatment effects. Imai and Ratkovic (2013) estimate a LASSO regression model with the treatment indicator interacted with covariates, and uses LASSO as a variable selection algorithm for determining which covariates are most important. In using this approach, it may be prudent to perform some supplementary analysis to verify that the method is not overfitting; for example, one could use a sample-splitting approach, using half of the data to estimate the LASSO regression and then comparing the results to an ordinary least squares regression with the variables selected by LASSO in the other half of the data. If the results are inconsistent, it could indicate that using half the data is not good enough, or it might indicate that sample splitting is warranted to protect against overfitting or other sources of bias that arise when data-driven model selection is used.

A natural application of personalized treatment effect estimation is to estimate optimal policy functions in observational data. A literature in machine learning considers this problem (Beygelzimer and Langford 2009; Beygelzimer et al. 2011); some open questions include the statistical properties of the estimators, and the ability to obtain confidence intervals on differences between policies obtained from these methods. Recently, Athey and Wager (2017) bring in insights from semiparametric efficiency theory in econometrics to propose a new estimator for optimal policies and to analyze the properties of this estimator. Policies can be compared in terms of their "risk," which is defined as the gap between the expected outcomes using the (unknown) optimal policy and the estimated policy. Athey and Wager derive an upper bound for the risk of the policy estimated using their method and show that it is necessary to use a method that is efficient (in the econometric sense) to achieve that bound.

Conclusion

In the last few decades, economists have learned to take very seriously the old admonition from undergraduate econometrics that "correlation is not causality." We have surveyed a number of recent developments in the econometrics toolkit for addressing causality issues in the context of estimating the impact of policies. Some of these developments involve a greater sophistication in the use of methods like regression discontinuity and differences-in-differences estimation. But we have also tried to emphasize that the project of taking causality seriously often benefits from combining these tools with other approaches. Supplementary analyses can help the analyst assess the credibility of estimation and identification strategies. Machine learning methods provide important new tools to improve estimation of causal effects in high-dimensional settings, because in many cases it is important to flexibly control for a large number of covariates as part of an estimation strategy for drawing causal inferences from observational data. When causal interpretations of estimates are more plausible, and inference about causality can reduce the reliance of these estimates on modeling assumptions (like those about functional form), the credibility of policy analysis is enhanced.

■ *We are grateful for comments by the editor and coeditors.*

References

Abadie, Alberto, Alexis Diamond, and Jens Hainmueller. 2010. "Synthetic Control Methods for Comparative Case Studies: Estimating the Effect of California's Tobacco Control Program." *Journal of the American Statistical Association* 105(490): 493–505.

Abadie, Alberto, Alexis Diamond, and Jens Hainmueller. 2014. "Comparative Politics and the Synthetic Control Method." *American Journal of Political Science* 59(2): 495–510.

Abadie, Alberto, and Javier Gardeazabal. 2003. "The Economic Costs of Conflict: A Case Study of the Basque Country." *American Economic Review* 93(1): 113–32.

Abadie, Alberto, and Guido W. Imbens. 2006. "Large Sample Properties of Matching Estimators for Average Treatment Effects." *Econometrica* 74(1): 235–67.

Allcott, Hunt. 2015. "Site Selection Bias in Program Evaluation." *Quarterly Journal of Economics* 130(3): 1117–65.

Altonji, Joseph G., Todd E. Elder, and Christopher R. Taber. 2008. "Using Selection on Observed Variables to Assess Bias from Unobservables When Evaluating Swan–Ganz Catheterization." *American Economic Review* 98(2): 345–50.

Andrews, Donald, and James H. Stock. 2006. "Inference with Weak Instruments." Unpublished paper.

Angrist, Joshua D. 2004. "Treatment Effect Heterogeneity in Theory and Practice." *Economic Journal* 114(494): C52–83.

Angrist, Joshua, and Ivan Fernandez-Val. 2010. "ExtrapoLATE-ing: External Validity and Overidentification in the LATE Framework." NBER Working Paper 16566.

Angrist, Joshua D., Guido W. Imbens, and Donald B. Rubin. 1996. "Identification of Causal Effects Using Instrumental Variables." *Journal of the American Statistical Association* 91(434): 444–55.

Angrist, Joshua D., and Alan B. Krueger. 1999. "Empirical Strategies in Labor Economics." In *Handbook of Labor Economics*, edited by Orley C. Ashenfelter and David Card, 1277–1366. North Holland.

Angrist, Joshua D., and Miikka Rokkanen. 2015. "Wanna Get Away? Regression Discontinuity Estimation of Exam School Effects Away From the Cutoff." *Journal of the American Statistical Association* 110(512): 1331–44.

Arkhangelskiy, Dmitry, and Evgeni Drynkin. 2016. "Sensitivity to Model Specification." Unpublished paper.

Aronow, Peter M., and Cyrus Samii. 2013. "Estimating Average Causal Effects under Interference between Units." arXiv: 1305.6156v1.

Assmann, Susan F., Stuart J. Pocock, Laura E. Enos, and Linda E. Kasten. 2000. "Subgroup Analysis and Other (Mis)uses of Baseline Data in Clinical Trials." *Lancet* 355 (9209): 1064–69.

Athey, Susan, Raj Chetty, and Guido Imbens. 2016. "Combining Experimental and Observational Data: Internal and External Validity." Unpublished paper.

Athey, Susan, Raj Chetty, Guido Imbens, and Hyunseung Kang. 2016. "Estimating Treatment Effects Using Multiple Surrogates: The Role of the Surrogate Score and the Surrogate Index." arXiv: 1603.09326.

Athey, Susan, Dean Eckles, and Guido Imbens. Forthcoming. "Exact *p*-Values for Network Interference." *Journal of the American Statistical Association.*

Athey, Susan, and Guido W. Imbens. 2006. "Identification and Inference in Nonlinear Difference-in-Differences Models." *Econometrica* 74(2): 431–97.

Athey, Susan, and Guido Imbens. 2015. "A Measure of Robustness to Misspecification." *American Economic Review* 105(5): 476–80.

Athey, Susan, and Guido Imbens. 2016. "Recursive Partitioning for Estimating Heterogeneous Causal Effects." *PNAS* 113(27): 7353–60.

Athey, Susan, Guido Imbens, Thai Pham, and Stefan Wager. 2017. "Estimating Average Treatment Effects: Supplementary Analyses and Remaining Challenges." arXiv: 1702.01250.

Athey, Susan, Guido Imbens, and Stefan Wager. 2016. "Efficient Inference of Average Treatment Effects in High Dimensions via Approximate Residual Balancing." arXiv: 1604.07125.

Athey, Susan, Markus Mobius, and Jeno Pal. 2016. "The Impact of Aggregators on News Consumption." Unpublished paper.

Athey, Susan, Julie Tibshirani, and Stefan Wager. 2016. "Solving Heterogeneous Estimating Equations with Gradient Forests." arXiv: 1610.01271.

Athey, Susan, and Stefan Wager. 2017. "Efficient Policy Learning." arXiv: 1702.02896.

Banerjee, Abhijit, Sylvain Chassang, and Erik Snowberg. 2016. "Decision Theoretic Approaches to Experiment Design and External Validity." NBER Working Paper 22167.

Bang, Heejung, and James M. Robins. 2005. "Doubly Robust Estimation in Missing Data and Causal Inference Models." *Biometrics* 61(4): 962–73.

Begg, Colin B., and Denis H. Y. Leung. 2000. "On the Use of Surrogate End Points in

Randomized Trials." *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 163(1): 15–28.

Bekker, Paul A. 1994. "Alternative Approximations to the Distributions of Instrumental Variable Estimators." *Econometrica* 62(3): 657–81.

Belloni, Alexandre, Victor Chernozhukov, Ivan Fernández-Val, and Chris Hansen. 2013. "Program Evaluation and Causal Inference with High-Dimensional Data." arXiv: 1311.2645.

Bertanha, Marinho, and Guido Imbens. 2014. "External Validity in Fuzzy Regression Discontinuity Designs." NBER Working Paper 20773.

Beygelzimer, Alina, and John Langford. 2009. "The Offset Tree for Learning with Partial Labels." *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 129–38.

Beygelzimer, Alina, John Langford, Lihong Li, Lev Reyzin, and Robert E. Schapire. 2011. "Contextual Bandit Algorithms with Supervised Learning Guarantees." *Proceedings of the 14th International Conference on Artificial Intelligence and Statistics (AISTATS)*, pp. 19–26.

Bramoullé, Yann, Habiba Djebbari, and Bernard Fortin. 2009. "Identification of Peer Effects through Social Networks." *Journal of Econometrics* 150(1): 41–55.

Breiman, Leo, Jerome Friedman, Charles J. Stone, and Richard A. Olshen. 1984. *Classification and Regression Trees.* CRC Press.

Brinch, Christian, Magne Mogstad, and Matthew Wiswall. 2015. "Beyond LATE with a Discrete Instrument: Heterogeneity in the Quantity-Quality Interaction in Children." Unpublished paper.

Calonico, Sebastian, Matias D. Cattaneo, and Rocío Titiunik. 2014a. "Robust Nonparametric Confidence Intervals for Regression-Discontinuity Designs." *Econometrica* 82(6): 2295–2326.

Calonico, Sebastian, Matias D. Cattaneo, and Rocío Titiunik. 2014b. "Robust Data-Driven Inference in the Regression-Discontinuity Design." *Stata Journal* 14(4): 909–46.

Card, David. 1990. "The Impact of the Mariel Boatlift on the Miami Labor Market." *Industrial and Labor Relations Review* 43 (2): 245–57.

Card, David, David S. Lee, Zhuan Pei, and Andrea Weber. 2015. "Inference on Causal Effects in a Generalized Regression Kink Design." *Econometrica* 83 (6): 2453–83.

Carrell, Scott E., Bruce I. Sacerdote, and James E. West. 2013. "From Natural Variation to Optimal Policy? The Importance of Endogenous Peer Group Formation." *Econometrica* 81(3): 855–82.

Cattaneo, Matias D. 2010. "Efficient Semiparametric Estimation of Multi-valued Treatment Effects under Ignorability." *Journal of Econometrics* 155(2): 138–54.

Chamberlain, Gary, and Guido Imbens. 2004. "Random Effects Estimators with Many Instrumental Variables." *Econometrica* 72(1): 295–306.

Chandrasekhar, Arun G. 2016. "The Econometrics of Network Formation." Chap. 13 in *The Oxford Handbook on the Economics of Networks*, edited by Yann Bramoullé, Andrea Galeotti, Brian W. Rogers. Oxford University Press.

Chandrasekhar, Arun, and Matthew Jackson. 2016. "A Network Formation Model Based on Subgraphs." arXiv: 1611.07658.

Chernozhukov, Victor, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, and Whitney Newey. 2016. "Double Machine Learning for Treatment and Causal Parameters." arXiv: 1608.00060.

Chetty, Raj. 2009. "Sufficient Statistics for Welfare Analysis: A Bridge between Structural and Reduced-Form Methods." *Annual Review of Economics* 1(1): 451–87.

Chipman, Hugh A., Edward I. George, and Robert E. McCulloch. 2010. "BART: Bayesian Additive Regression Trees." *Annals of Applied Statistics* 4(1): 266–98.

Christakis, Nicholas A., and James H. Fowler. 2007. "The Spread of Obesity in a Large Social Network over 32 Years." *New England Journal of Medicine* (357): 370–79.

Christakis, Nicholas A., James H. Fowler, Guido W. Imbens, and Karthik Kalyanaraman. 2010. "An Empirical Model for Strategic Network Formation." NBER Working Paper 16039.

Conley, Timothy G., Christian B. Hansen, and Peter E. Rossi. 2012. "Plausibly Exogenous." *Review of Economics and Statistics* 94(1): 260–72.

Crépon, Bruno, Esther Duflo, Marc Gurgand, Roland Rathelot, and Philippe Zamora. 2013. "Do Labor Market Policies Have Displacement Effects? Evidence from a Clustered Randomized Experiment." *Quarterly Journal of Economics* 128(2): 531–80.

Crump, Richard K., V. Joseph Hotz, Guido W. Imbens, and Oscar A. Mitnik. 2009. "Dealing with Limited Overlap in Estimation of Average Treatment Effects." *Biometrika* 96(1): 187–99.

Deaton, Angus. 2010. "Instruments, Randomization, and Learning about Development." *Journal of Economic Literature* 48(2): 424–55.

Dehejia, Rajeev H., and Sadek Wahba. 1999. "Causal Effects in Nonexperimental Studies: Reevaluating the Evaluation of Training Programs." *Journal of the American Statistical Association* 94(448): 1053–62.

Dong, Yingying. 2014. "Jump or Kink? Identification of Binary Treatment Regression Discontinuity Design without the Discontinuity." Unpublished paper.

Dong, Yingying, and Arthur Lewbel. 2015. "Identifying the Effect of Changing the Policy Threshold in Regression Discontinuity Models." *Review of Economics and Statistics* 97(5): 1081–92.

Doudchenko, Nikolay, and Guido W. Imbens. 2016. "Balancing, Regression, Difference-in-Differences and Synthetic Control Methods: A Synthesis." arXiv: 1610.07748.

Foster, Jared C., Jeremy M. G. Taylor, and Stephen J. Ruberg. 2011. "Subgroup Identification from Randomized Clinical Data." *Statistics in Medicine* 30(24): 2867–80.

Frangakis, Constantine E., and Donald B. Rubin. 2002. "Principal Stratification in Causal Inference." *Biometrics* 58(1): 21–29.

Gelman, Andrew, and Guido Imbens. 2014. "Why High-Order Polynomials Should Not Be Used in Regression Discontinuity Designs." NBER Working Paper 20405.

Gentzkow, Matthew, and Jesse Shapiro. 2015. "Measuring the Sensitivity of Parameter Estimates to Sample Statistics." Unpublished paper.

Goldberger, Arthur S. 1972. "Selection Bias in Evaluating Treatment Effects: Some Formal Illustrations." Institute for Research on Poverty Discussion Paper 129-72.

Goldberger, Arthur S. 2008. "Selection Bias in Evaluating Treatment Effects: Some Formal Illustrations." In *Advances in Econometrics, Volume 21*, edited by Tom Fomby, R. Carter Hill, Daniel L. Millimet, Jeffrey A. Smith, and Edward J. Vytlacil, 1–31. Emerald Group Publishing Limited.

Goldsmith-Pinkham, Paul, and Guido W. Imbens. 2013. "Social Networks and the Identification of Peer Effects." *Journal of Business and Economic Statistics* 31(3): 253–64.

Graham, Bryan S. 2008. "Identifying Social Interactions through Conditional Variance Restrictions." *Econometrica* 76(3): 643–60.

Graham, Bryan S., Cristine Campos de Xavier Pinto, and Daniel Egel. 2012. "Inverse Probability Tilting for Moment Condition Models with Missing Data." *Review of Economic Studies* 79(3): 1053–79.

Graham, Bryan, Christine Campos de Xavier Pinto, and Daniel Egel. 2016. "Efficient Estimation of Data Combination Models by the Method of Auxiliary-to-Study Tilting (AST)." *Journal of Business and Economic Statistics* 34(2): 288–301.

Green, Donald P., and Holger L. Kern. 2012. "Modeling Heterogeneous Treatment Effects in Survey Experiments with Bayesian Additive Regression Trees." *Public Opinion Quarterly* 76(3): 491–511.

Hahn, Jinyong, Petra Todd, and Wilbert van der Klaauw. 2001. "Identification and Estimation of Treatment Effects with a Regression-Discontinuity Design." *Econometrica* 69(1): 201–09.

Hainmueller, Jens. 2012. "Entropy Balancing for Causal Effects: A Multivariate Reweighting Method to Produce Balanced Samples in Observational Studies." *Political Analysis* 20(1): 25–46.

Heckman, James J., and V. Joseph Hotz. 1989. "Choosing among Alternative Nonexperimental Methods for Estimating the Impact of Social Programs: The Case of Manpower Training." *Journal of the American Statistical Association* 84(408): 862–74.

Heckman, James J., and Edward Vytlacil. 2007. "Econometric Evaluation of Social Programs, Part I: Causal Models, Structural Models and Econometric Policy Evaluation." In *Handbook of Econometrics 6B*, edited by James Heckman and Edward Leamer, 4779–4874. Elsevier.

Hill, Jennifer L. 2011. "Bayesian Nonparametric Modeling for Causal Inference." *Journal of Computational and Graphical Statistics* 20(1): 217–40.

Hirano, Keisuke. 2001. "Combining Panel Data Sets with Attrition and Refreshment Samples." *Econometrica* 69(6): 1645–59.

Hirano, Keisuke, and Guido Imbens. 2004. "The Propensity Score with Continuous Treatments." In *Applied Bayesian Modeling and Causal Inference from Incomplete-Data Perspectives: An Essential Journey with Donald Rubin's Statistical Family*, edited by Andrew Gelman and Xiao-Li Meng, 73–84. Wiley.

Holland, Paul W. 1986. "Statistics and Causal Inference." *Journal of the American Statistical Association* 81(396): 945–60.

Hotz, V. Joseph, Guido W. Imbens, and Julie H. Mortimer. 2005. "Predicting the Efficacy of Future Training Programs Using Past Experiences at Other Locations." *Journal of Econometrics* 125(1–2): 241–70.

Hudgens, Michael G., and M. Elizabeth Halloran. 2008. "Toward Causal Inference with Interference." *Journal of the American Statistical Association* 103(482): 832–42.

Imai, Kosuke, and Marc Ratkovic. 2013. "Estimating Treatment Effect Heterogeneity in Randomized Program Evaluation." *Annals of Applied Statistics* 7(1): 443–70.

Imai, Kosuke, and Marc Ratkovic. 2014. "Covariate Balancing Propensity Score." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 76(1): 243–63.

Imai, Kosuke, and David A. van Dyk. 2004. "Causal Inference with General Treatment Regimes: Generalizing the Propensity Score." *Journal of the American Statistical Association* 99(467): 854–66.

Imbens, Guido W. 2000. "The Role of the Propensity Score in Estimating Dose–Response Functions." *Biometrika* 87(3): 706–10.

Imbens, Guido W. 2003. "Sensitivity to

Exogeneity Assumptions in Program Evaluation." *American Economic Review* 93(2): 126–32.

Imbens, Guido W. 2004. "Nonparametric Estimation of Average Treatment Effects under Exogeneity: A Review." *Review of Economics and Statistics* 86(1): 4–29.

Imbens, Guido W. 2010. "Better LATE Than Nothing: Some Comments on Deaton (2009) and Heckman and Urzua (2009)." *Journal of Economic Literature* 48(2): 399–423.

Imbens, Guido. 2013. "Book Review Feature: Public Policy in an Uncertain World." *Economic Journal* 123(570): F401–411.

Imbens, Guido W. 2014. "Instrumental Variables: An Econometrician's Perspective." *Statistical Science* 29(3): 323–58.

Imbens, Guido W. 2015. "Matching Methods in Practice: Three Examples." *Journal of Human Resources* 50(2): 373–419.

Imbens, Guido W., and Joshua D. Angrist. 1994. "Identification and Estimation of Local Average Treatment Effects." *Econometrica* 62(2): 467–75.

Imbens, Guido W., and Karthik Kalyanaraman. 2012. "Optimal Bandwidth Choice for the Regression Discontinuity Estimator." *Review of Economic Studies* 79(3): 933–59.

Imbens, Guido W., and Thomas Lemieux. 2008. "Regression Discontinuity Designs: A Guide to Practice." *Journal of Econometrics* 142(2): 615–35.

Imbens, Guido W., and Donald B. Rubin. 2015. *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction.* Cambridge University Press.

Imbens, Guido W., Donald B. Rubin, and Bruce I. Sacerdote. 2001. "Estimating the Effect of Unearned Income on Labor Earnings, Savings, and Consumption: Evidence from a Survey of Lottery Players." *American Economic Review* 91(4): 778–94.

Imbens, Guido W., and Jeffrey M. Wooldridge. 2009. "Recent Developments in the Econometrics of Program Evaluation." *Journal of Economic Literature* 47(1): 5–86.

Jackson, Matthew O. 2010. *Social and Economic Networks.* Princeton University Press.

Jackson, Matthew, and Asher Wolinsky. 1996. "A Strategic Model of Social and Economic Networks." *Journal of Economic Theory* 71(1): 44–74.

Jacob, Brian A., and Lars Lefgren. 2004. "Remedial Education and Student Achievement: A Regression-Discontinuity Analysis." *Review of Economics and Statistics* 86(1): 226–44.

Kleinberg, Jon, Jens Ludwig, Sendhil Mullainathan, and Ziad Obermeyer. 2015. "Prediction Policy Problems." *American Economic Review* 105(5): 491–95.

Kowalski, Amanda. 2016. "Doing More When You're Running LATE: Applying Marginal Treatment Effect Methods to Examine Treatment Effect Heterogeneity in Experiments." NBER Paper 22363.

LaLonde, Robert J. 1986. "Evaluating the Econometric Evaluations of Training Programs with Experimental Data." *American Economic Review* 76(4): 604–20.

Leamer, Edward. 1978. *Specification Searches: Ad Hoc Inference with Nonexperimental Data.* Wiley.

Leamer, Edward E. 1983. "Let's Take the Con Out of Econometrics." *American Economic Review* 73(1): 31–43.

Lechner, Michael. 2001. "Identification and Estimation of Causal Effects of Multiple Treatments under the Conditional Independence Assumption." In *Econometric Evaluation of Labour Market Policies,* vol. 13, edited by Michael Lechner and Friedhelm Pfeiffer, 43–58. Physica-Verlag Heidelberg.

Lee, David S. 2008. "Randomized Experiments from Non-random Selection in U.S. House Elections." *Journal of Econometrics* 142(2): 675–97.

Lee, David S., and Thomas Lemieux. 2010. "Regression Discontinuity Designs in Economics." *Journal of Economic Literature* 48(2): 281–355.

List, John A., Azeem M. Shaikh, and Yang Xu. 2016. "Multiple Hypothesis Testing in Experimental Economics." NBER Paper 21875.

Manski, Charles F. 1990. "Nonparametric Bounds on Treatment Effects." *American Economic Review* 80(2): 319–23.

Manski, Charles F. 1993. "Identification of Endogenous Social Effects: The Reflection Problem." *Review of Economic Studies* 60(3): 531–42.

Manski, Charles F. 2013. *Public Policy in an Uncertain World: Analysis and Decisions.* Harvard University Press.

McCaffrey, Daniel F., Greg Ridgeway, and Andrew R. Morral. 2004. "Propensity Score Estimation with Boosted Regression for Evaluating Causal Effects in Observational Studies." *Psychological Methods* 9(4): 403–25.

McCrary, Justin. 2008. "Manipulation of the Running Variable in the Regression Discontinuity Design: A Density Test." *Journal of Econometrics* 142(2): 698–714.

Mele, Angelo. 2013. "A Structural Model of Segregation in Social Networks." Available at SSRN: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2294957.

Nielsen, Helena Skyt, Torben Sorensen, and Christopher Taber. 2010. "Estimating the Effect of Student Aid on College Enrollment: Evidence from a Government Grant Policy Reform." *American Economic Journal: Economic Policy* 2(2): 185–215.

Oster, Emily. 2015. "Diabetes and Diet: Behavioral Response and the Value of Health." NBER Working Paper 21600.

Otsu, Taisuke, Ke-Li Xu, and Yukitoshi Matsushita. 2013. "Estimation and Inference of Discontinuity in Density." *Journal of Business and Economic Statistics* 31(4): 507–24.

Pearl, Judea. 2000. *Causality: Models, Reasoning, and Inference.* Cambridge University Press.

Peri, Giovanni, and Vasil Yasenov. 2015. "The Labor Market Effects of a Refugee Wave: Applying the Synthetic Control Method to the Mariel Boatlift." NBER Working Paper 21801.

Porter, Jack. 2003. "Estimation in the Regression Discontinuity Model." Unpublished paper, University of Wisconsin at Madison.

Prentice, Ross L. 1989. "Surrogate Endpoints in Clinical Trials: Definition and Operational Criteria." *Statistics in Medicine* 8 (4): 431–40.

Robins, James M., and Andrea Rotnitzky. 1995. "Semiparametric Efficiency in Multivariate Regression Models with Missing Data." *Journal of the American Statistical Association* 90(429): 122–29.

Robins, James M., Andrea Rotnitzky, and Lue Ping Zhao. 1995. "Analysis of Semiparametric Regression Models for Repeated Outcomes in the Presence of Missing Data." *Journal of the American Statistical Association* 90(429): 106–21.

Rosenbaum, Paul. 1987. "The Role of a Second Control Group in an Observational Study." *Statistical Science* 2(3): 292–306.

Rosenbaum, Paul R. 2002. "Observational Studies." In *Observational Studies*, 1–17. Springer.

Rosenbaum, Paul R., and Donald B. Rubin. 1983a. "The Central Role of the Propensity Score in Observational Studies for Causal Effects." *Biometrika* 70(1): 41–55.

Rosenbaum, Paul R., and Donald B. Rubin. 1983b. "Assessing Sensitivity to an Unobserved Binary Covariate in an Observational Study with Binary Outcome." *Journal of the Royal Statistical Society. Series B (Methodological)* 45(2): 212–18.

Sacerdote, Bruce. 2001. "Peer Effects with Random Assignment: Results for Dartmouth Roommates." *Quarterly Journal of Economics* 116(2): 681–704.

Shadish, William R., Thomas D. Cook, and Donald T. Campbell. 2002. *Experimental and Quasi-experimental Designs for Generalized Causal Inference.* Houghton Mifflin.

Skovron, Chistopher, and Rocío Titiunik. 2015. "A Practical Guide to Regression Discontinuity Designs in Political Science." Unpublished paper.

Staiger, Douglas, and James H. Stock. 1997. "Instrumental Variables Regression with Weak Instruments." *Econometrica* 65(3): 557–86.

Su, Xiaogang, Chih-Ling Tsai, Hansheng Wang, David M. Nickerson, and Bogong Li. 2009. "Subgroup Analysis via Recursive Partitioning." *Journal of Machine Learning Research* 10: 141–58.

Tamer, Elie. 2010. "Partial Identification in Econometrics." *Annual Review of Economics* 2(1): 167–95.

Thistlewaite, D., and Donald Campbell. 1960. "Regression-Discontinuity Analysis: An Alternative to the Ex-post Facto Experiment." *Journal of Educational Psychology* 51(6): 309–17.

Todd, Petra, and Kenneth I. Wolpin. 2006. "Assessing the Impact of a School Subsidy Program in Mexico: Using a Social Experiment to Validate a Dynamic Behavioral Model of Child Schooling and Fertility." *American Economic Review* 96 (5): 1384–1417.

van der Klaauw, Wilbert. 2008. "Regression-Discontinuity Analysis: A Survey of Recent Developments in Economics." *Labour* 22(2): 219–45.

van der Laan, Mark J., and Daniel Rubin. 2006. "Targeted Maximum Likelihood Learning." *International Journal of Biostatistics* 2(1).

van der Vaart, Aad W. 2000. *Asymptotic Statistics.* Cambridge University Press.

Wager, Stefan, and Susan Athey. 2015. "Estimation and Inference of Heterogeneous Treatment Effects Using Random Forests." arXiv:1510.04342

Wyss, Richard, Allan Ellis, Alan Brookhart, Cynthia Girman, Michele Jonsson Funk, Robert LoCasale, and Til Strümer. 2014 "The Role of Prediction Modeling in Propensity Score Estimation: An Evaluation of Logistic Regression, bCART, and the Covariate-Balancing Propensity Score." *American Journal of Epidemiology* 180(6): 645–55.

Yang, Shu, Guido W. Imbens, Zhanglin Cui, Douglas E. Faries, and Zbigniew Kadziola. 2016. "Propensity Score Matching and Subclassification in Observational Studies with Multi-level Treatments." *Biometrics* 72(4): 1055–65.

Zeileis, Achim, Torsten Hothorn, and Kurt Hornik. 2008. "Model-Based Recursive Partitioning." *Journal of Computational and Graphical Statistics* 17(2): 492–514.

Zubizarreta, Jose R. 2015. "Stable Weights that Balance Covariates for Estimation with Incomplete Outcome Data." *Journal of the American Statistical Association* 110(511): 910–22.

Journal of Economic Perspectives—Volume 31, Number 2—Spring 2017—Pages 33–58

The Use of Structural Models in Econometrics

Hamish Low and Costas Meghir

The aim of this paper is to discuss the role of structural economic models in empirical analysis and policy design. This approach offers some valuable payoffs, but also imposes some costs.

Structural economic models focus on distinguishing clearly between the objective function of the economic agents and their opportunity sets as defined by the economic environment. The key features of such an approach at its best are a tight connection with a theoretical framework alongside a clear link with the data that will allow one to understand how the model is identified. The set of assumptions under which the model inferences are valid should be clear: indeed, the clarity of the assumptions is what gives value to structural models.

The central payoff of a structural econometric model is that it allows an empirical researcher to go beyond the conclusions of a more conventional empirical study that provides reduced-form causal relationships. Structural models define how outcomes relate to preferences and to relevant factors in the economic environment, identifying mechanisms that determine outcomes. Beyond this, they

■ *Hamish Low is Professor of Economics, Cambridge University, Cambridge, United Kingdom, and Cambridge INET and Research Fellow, Institute for Fiscal Studies, London, United Kingdom. Costas Meghir is the Douglas A. Warner III Professor of Economics, Yale University, New Haven, Connecticut; International Research Associate, Institute for Fiscal Studies, London, United Kingdom; Research Fellow, Institute of Labor Economics (IZA), Bonn, Germany; and Research Associate, National Bureau of Economic Research, Cambridge, Massachusetts. Their email addresses are hamish.low@econ.cam.ac.uk and c.meghir@yale.edu.*

† For supplementary materials such as appendices, datasets, and author disclosure statements, see the article page at
https://doi.org/10.1257/jep.31.2.33 doi=10.1257/jep.31.2.33

are designed to analyze counterfactual policies, quantifying impacts on specific outcomes as well as effects in the short and longer run.

The short-run implications can often be compared to what actually happened in the data, allowing for validation of the model. For example, Blundell, Costa Dias, Meghir, and Shaw (2016) model how life-cycle female labor supply and human capital accumulation are affected by tax credit reform. They validate the model by comparing its short-run predictions to those estimated by simple reduced form methods. However, their model also has implications for labor supply and wages beyond the childbearing age, as well as for the educational choice of subsequent cohorts, none of which can be estimated from actual data without an economic model. Such effects are of central importance for understanding the impacts of welfare programs. Similarly, Low and Pistaferri (2015) model the long-run effects of reform to disability insurance, but validate their model using reduced form predictions. This symbiotic interaction of structural models and reduced form approaches, including randomized experiments, provides the strongest tool in the empirical economics toolkit and is emphasized in this paper.

Additional insights come with tradeoffs. Structural economic models cannot possibly capture every aspect of reality, and any effort to do so would make them unwieldy for either theoretical insight or applied analysis. There will always be some economic choices left out of any particular model—the key question is how to judge what aspects to leave out without rendering the quantitative conclusions of the model irrelevant. The principle we advocate to focus on the question of interest, to achieve parsimony, and to understand how much the model distorts reality is the concept of separability (related to Fisher's separation theorem and Gorman's notion of separability, as discussed in Gorman 1995). This leads to the concept of sufficient statistics, which summarize decisions made outside the model. A specific example is that of consumer two-stage budgeting: the first stage defines the total amount to be spent in a particular period, while the second stage allocates that expenditure to a variety of goods within the period. In modeling the within-period allocation, we may not concern ourselves with what determines the intertemporal allocation problem: under suitable separability assumptions, a sufficient statistic for the intertemporal allocation decision is total consumption (MaCurdy 1983; Altonji 1986; Blundell and Walker 1986; Arellano and Meghir 1992).

The validity of the abstraction of a structural model depends on how appropriate the particular separability assumptions being made are. This sort of abstraction is present even if we are modeling a market equilibrium that considers both the supply and demand sides. We focus on a limited system, because anything more would be too complicated to offer insights. For example, we may model the equilibrium in the labor market and pay determination but say nothing explicitly about the product market or capital investment (for example, Burdett and Mortensen 1998).

Structural economic models should be taught and used as part of the standard toolkit for empirical economists. Of course, other parts of that toolkit include treatment effect models based on quasi-experimental methods and randomized experiments, but these present trade-offs of their own: in particular, the interpretation of data can become limited and fragmented without the organizing discipline of economic models. Further, without the ability to simulate counterfactuals and more generally to make claims of external validity, the role of empirical analysis is limited to analyzing historical past events without being able to use this accumulated knowledge in a constructive and organized way.

Solving structural models, especially dynamic stochastic models, involves numerical methods. These numerical methods are used to simulate outcomes and counterfactuals as well as to generate moments for use in estimation. The greatest "entry cost" for a researcher wishing to estimate dynamic structural models is learning to solve them, and as we discuss, there are many steps involved in their solution and estimation. Understanding their solution also helps in understanding how they are identified by the data.

In what follows, we start by defining structural models, distinguishing between those that are fully specified and those that are partially specified. We contrast the treatment effects approach with structural models, using Low, Meghir, and Pistaferri (2010) as an example of how a structural model is specified and the particular choices that were made. We follow with a discussion of combining structural estimation with randomized experiments. We then turn to numerical techniques for solving dynamic stochastic models that are often used in structural estimation, again using Low, Meghir, and Pistaferri as an example. The penultimate section focuses on issues of estimation using the method of moments. The last section concludes.

Defining a Structural Model

We begin by differentiating between fully and partially specified structural models, and then consider their relationship to treatment effect models.

Fully Specified Structural Models

Fully specified structural models make explicit assumptions about the economic actors' objectives and their economic environment and information set, as well as specifying which choices are being made within the model. We call these models fully specified because they allow a complete solution to the individual's optimization problem as a function of the current information set. In the context of labor economics, Keane and Wolpin (1997) and numerous papers by these authors are prime examples of fully specified structural models. Structural models are the foundation for empirical work in industrial organization with key references being Berry, Levinsohn, and Pakes (1995) and Koujianou-Goldberg (1995); however, most of our discussion draws from examples in labor economics and public finance.

A fully specified dynamic model of consumption and labor force participation will account for how employment and savings decisions are made, taking into account future expectations as well as future implications of these decisions. Working today can imply changes in future wages because of skill accumulation, thus altering the future returns to work and/or through changes in the preferences for work (habit formation). The choices that the individual makes depend on beliefs about future opportunities (such as wage rates) and future preferences. Thus, in a fully specified model we need to define the distribution of random events (such as shocks to wages and human capital) often specifying the explicit functional form of the distributions and their persistence. We specify the dynamics of other observable or unobservable variables that affect decisions, distinguishing endogenous changes (such as to wealth due to saving decisions, or to human capital as a result of experience) from exogenous changes (such as to prices or to health). These features are all assumed to be in the individual's information set.

Of course no model is literally complete—all models necessarily abstract from possibly relevant choices. These simplifications take two forms: a choice variable may be completely absent from a model, as for example, in the simplest life-cycle model of consumption under uncertainty, which ignores labor supply and takes income to be some exogenous stochastic process. Low (2005) shows that this assumption can lead to underestimates of precautionary saving behaviour. Alternatively we may condition on a choice, but take it as economically exogenous, as discussed in Browning and Meghir (1991). For example, life-cycle behavior may depend on education, but the level of education is taken as given in modeling consumption: the solution of the consumption function will be conditional on education choice.

To illustrate the issues, consider the structural model in Low, Meghir, and Pistaferri (2010). This is a life-cycle model of consumption and labor supply with a specific focus on quantifying employment and wage risk and measuring the welfare cost of risk, with implications for the design of welfare programs. Individuals choose whether to work, whether to change jobs if the opportunity arises, and how much to save.

The first step is to specify the components of the model. A first component is the intertemporal utility function describing preferences and defining what is chosen. A second component is the intertemporal budget constraint, which depends on the available welfare benefits and taxes. Finally, we need to specify how the individual forms expectations about the future, including shocks to human capital and job loss probabilities and opportunities for new jobs. More broadly, we need to specify how preferences are defined over time and over the states of the world and whether the individual is an expected utility maximizer. Together this characterizes the problem facing the individual. These components also define parameters that need to be estimated from the data after we have argued for how they are identified.

We also need to decide what *not* to model. Of course, this list of omissions is a long one, but for models of life-cycle behavior, the most glaring omissions are marriage and fertility: in our example, male preferences are assumed to be separable from these, as is often done in the literature on male labor supply. Education is taken as given (although it affects choices and opportunity sets). Overall savings are explained but not portfolio allocations. Finally, the model is partial equilibrium, in the sense that counterfactual simulations abstract from changes in wages that may result from aggregate changes in the supply of labor. Perhaps more importantly, the model abstracts from aggregate shocks. This means that the results have little to say about how the welfare effects of idiosyncratic risk vary with the state of the aggregate economy. The judgment is that these other aspects obscure and complicate the model rather than offer important insights given the stated aims. The complications of these extensions are also partly numerical, as we discuss later in this paper.

Some assumptions are made for simplicity and focus, but others are identifying assumptions. For example, the specific distribution of the shocks may be an identifying assumption. A crucial question that arises is the minimal set of assumptions needed for the model to have empirical content and thus be empirically identified. These issues have been much discussed from different perspectives: useful starting points include Rust (1992) and Magnac and Thesmar (2002). Overall, their conclusion is that dynamic discrete choice models need some strong identification assumptions to work. These assumptions can be relaxed somewhat if a continuous outcome variable is involved such as wages (Heckman and Navarro 2007).

The payoff of such assumptions is that we are able to construct a model that is complex in the important dimensions and relatively transparent in the implied mechanisms. In Low, Meghir, and Pistaferri (2010), there are two separate sources of risk—employment and productivity—and a particularly complex budget constraint specifying the details of the available welfare programs. The relative simplicity of the specification hides important numerical complexities because the consumption function may be discontinuous in assets due to the discrete labor supply. The stochastic process of wages is serially correlated, increasing the numerical complexity of the problem. However, within this structure, it is still relatively easy to understand the role of the various sources of risk and how they affect welfare and the way we evaluate various welfare programs. Whether the channel of changed fertility decisions resulting from welfare reform is important for this problem is of course an open question.

Fully specified structural models are particularly useful when we want to understand long-term effects of policy. In a recent paper, Blundell, Costa Dias, Meghir, and Shaw (2016) consider the impact on female careers of tax credits targeted to low-income families with children. A key question is whether tax credits improve longer-term labor market attachment of single mothers by incentivizing them to remain in work and thus avoiding human capital depreciation during the child-rearing period of life. The model quite decisively concludes this is not the case, partly because tax credits in the UK promote part-time work, which is not conducive to building up human capital, and partly because of tax-credit-induced disincentives to work for women within relationships (relative to the situation for single/divorced women). On the other hand, the model also shows that tax credits are by far superior to other commonly used methods of social insurance because of reduced moral hazard. Again, the specification of this model has made a number of simplifying assumptions, the most pertinent of which is to condition on the fertility process and not allow it to change as a result of welfare reform. Despite these sorts of limitations, a structural model that fully specifies behavior can go much further than simply estimating a parameter of interest or testing a particular theoretical hypothesis. To achieve this, a number of simplifying assumptions have to be made, to maintain feasibility and some level of transparency. The key is that the assumptions are made explicit, allowing future research to question results and make progress on that basis.

The discussion would be incomplete without touching upon empirical equilibrium models. Indeed, there are no better examples of completely specified models than those that also address equilibrium issues, since counterfactual analysis takes into account how the interaction between agents on both sides of the market leads to a new outcome. This requires specifying the behavior of all relevant agents and defining equilibrium in the specific context. At the same time, this provides an excellent example of how studies focus on some key features of equilibrium but not on others; this is both because of the need for focus on a particular question and for keeping modeling and computational complexity in check. Heckman, Lochner, and Taber (1998) and Lee and Wolpin (2008) focus on changes in equilibrium in the labor market; Abbott, Gallipoli, Meghir, and Violante (2013) also focus on the labor market equilibrium but in addition endogenize intergenerational links. Chiappori, Costa Dias, and Meghir (forthcoming), on the other hand, focus on equilibrium in the marriage market and on intrahousehold allocations, but do not consider changes in the labor market equilibrium, keeping wages constant. The search literature focuses on how equilibrium in frictional labor markets affects wage determination, as in the seminal paper of Burdett and Mortensen (1998) and a list of further important contributions too long to discuss here. All these studies estimate equilibrium models in some dimension but abstract from adjustments that are not the prime focus of the study. In so doing, they offer empirical insights on some of the important mechanisms at work in the longer run.

Partially Specified Structural Models

Sometimes our focus is on one component of a fully specified model. Consider an individual who maximizes lifetime utility by choosing consumption, savings, and how much to work in each period. We can derive a within-period labor supply function that is consistent with intertemporal choices but does not fully characterize them. Essentially, this is a reorganization of the marginal rate of substitution between consumption and labor supply. Such models rely on a *sufficient statistic* that summarizes choices not being modeled explicitly. In this case, the sufficient statistic is the amount of consumption allocated to the period. The econometric model defines a relationship between labor supply and wages, conditional on consumption and "looks" like a traditional labor supply model. The model is partially specified, in the sense that there is not enough information to solve for the optimal choice as a function of the information set: for example, the labor supply model resulting from the marginal rate of substitution characterization is silent about expectations for the future, the distribution of shocks, and the functioning of credit markets. However, conditioning on consumption makes the relationship between labor supply and wages valid and dependent upon structural parameters that characterize some aspects of utility. By studying this relationship, we can learn something about preferences and about the validity of this marginal rate of substitution representation, but we cannot simulate counterfactuals.

This idea builds on the concept of separability and two-stage budgeting introduced by Gorman (for example, Gorman 1995). In the context of empirical labor supply, this approach has been developed by MaCurdy (1983), Altonji (1986), and Blundell and Walker (1986), where separability is a restriction on preferences. More generally, separability is a way of specifying conditions on preferences and technologies that allow us to focus on some aspect of economic behavior without having to deal explicitly with the broader complications of understanding all aspects of behavior at once. In other words, it formalizes what we mean by a partially specified model and offers a way of understanding where misspecification may occur, which would be a failure of the explicit or implicit separability assumptions.

Partially specified structural models are an important empirical tool. They define testable implications for theory and allow us to estimate important parameters (such as the intertemporal elasticity of substitution or the Marshallian wage elasticity) in a way that is robust to different specifications in the parts of the model that remain unspecified, as discussed in the early simultaneous equations literature as well as Browning and Meghir (1991) and recently in Attanasio, Levell, Low, and Sanchez-Marcos (2017), amongst many others. They are explicit about what is kept constant when considering changes in variables and as such can provide consistent estimates for the parameters, given appropriate econometric methods. However, unlike fully specified models, the counterfactual analysis based on these is incomplete: for example, simulating the effect of taxes using a labor supply model that conditions on consumption will be limited by the inability of the model to capture the resulting intertemporal reallocation of consumption.

One of the most analyzed partially specified models is the Euler equation for consumption. It results from an assumption of intertemporally optimizing individuals and rational expectations. It does not require explicit information on the budget constraint because the level of consumption is used as a sufficient statistic for the marginal utility of wealth. This formulation has been the workhorse for examining the presence of liquidity constraints and for estimating the parameter of intertemporal substitution (for example, Attanasio and Weber 1995; Blundell, Browning, and Meghir 1994; Zeldes 1989). The often-used value for the elasticity of intertemporal substitution of about one originates from this body of work. Similarly, much has been learned by the analysis of the Euler equation for investment with adjustment costs (Bond and Meghir 1994). However, for counterfactual analysis, such as the impact of taxation on savings, the model needs to be completed by specifying the full economic environment as discussed above.

Treatment Effect Models

A treatment effect model focuses on identifying a specific causal effect of a policy or intervention while attempting to say the least possible about the theoretical context. The question is: following a policy change (like the introduction of an education subsidy, or a change in a welfare program), can we estimate the impact on a specific outcome such as education, labor supply, or perhaps transfers between individuals, without specifying a complete model or tying the result to a specific theory? Treatment effect models and their role in program evaluation are developed in Heckman and Robb (1985), Heckman, LaLonde, and Smith (1999), and a subsequent large literature.

The cleanest way of estimating program or treatment effects is experiments where interventions are randomly allocated. Given that in social contexts, compliance with the treatment protocol cannot usually be enforced—that is, subjects allocated to treatment (such as job training) cannot be forced to accept treatment—the randomized experiment will identify the effect of being *offered* treatment, or the intention to treat. Since impacts are possibly heterogeneous, the effect will be an average impact over the population for which randomization took place.

In a treatment model, identification does depend on the assumption that the experiment has not been compromised (either by nonadherence to the protocol or by attrition) and on there being no spillovers from the treatment units (individuals or communities or other groups such as schools) to the control ones, whether directly or through equilibrium mechanisms like price changes and peer effects. Given these important qualifications, we need not assume much about the underlying model of behavior. To get anything more than that out of the experiment, broadening its external validity, will typically require an explicit model, incorporating behavioral and often functional form assumptions.

Sometimes the result of the intention to treat is exactly what we want. However, consider estimating the effect of a welfare program by randomizing its availability (such as randomizing a conditional cash transfer that incentivizes child education and maternal health care, as in Mexico's PROGRESA). The welfare program may change current incentives to work or obtain education, future opportunities, the amount of risk households face, as well as the possibilities of risk-sharing in the communities (for example, Angelucci and De Giorgi 2009). It may even change wages in the affected communities (which it did). The treatment-effects model will isolate the impact of the program on any outcome we look at, but on its own will not be informative about the mechanisms. This limits the lessons from a particular experiment that are generalizable. To obtain more, we need to combine the information from the experiment with a model of household behavior and study how equilibrium in the communities changes. And of course generalizing the results to a scaled-up version of the policy is impossible without a model.

The literature on the effects of taxing higher incomes, discussed in Goolsbee, Hall, and Katz (1999) and Gruber and Saez (2002), provides another example of the issues that arise. Feldstein (1995) measured the impact of decreasing the top tax rate on earnings and incomes by using the 1986 Tax Reform Act. Separate from the issue of the particular merits of this empirical approach to identifying the causal impact, the external validity of the exercise is limited by the fact that the overall effect of reducing the top tax rate depends on how the entire tax schedule was changed and how people are distributed across it, which reduces the generality of the result to the specific context. Even when there is apparent randomization, such as in the comparison between lottery winners and losers in Imbens, Rubin, and Sacerdote (2001), there is still a threat to external validity: those choosing to participate in the lottery are likely to be those whose behavior will be most affected by winning, as shown by Crossley, Low, and Smith (2016).

Not all treatment effect models are created equal: it is important to distinguish those estimated through randomized experiments from those estimated through quasi-experimental methods, such as difference-in-differences, regression discontinuity, matching, and others. The point of randomized experiments is that results do not depend on strong assumptions about individual behavior, if we are able to exclude the important issues discussed above. However, this clarity is lost with quasi-experimental approaches such as difference-in-differences, where the validity of the results typically depends on assumptions relating to the underlying economic behavior that are left unspecified. For example, Athey and Imbens (2006) show that the assumption underlying difference-in-differences is that the outcome variable in the absence of treatment is generated by a model that is (strictly) monotonic in the unobservables, and the distribution of such unobservables must be invariant over time. These assumptions restrict the class of behavioral models that are consistent with the causal interpretation of difference in differences.

For example, suppose we want to estimate the effects of an intervention to increase the years of education. The difference-in-differences approach assumes that the level of education (in the absence of intervention) will be a strictly monotonic function of just one unobservable. Education is typically driven by the comparison of the benefits of education and the costs of education. The benefit can be expressed as the life-cycle value of wages and other outcomes resulting from an education choice. This benefit will in general be a nonlinear function of heterogeneity in wage returns, particularly if individuals are risk-averse. The costs are also likely to be heterogeneous. So the education choice will generally depend on at least two unobserved components, which are unlikely to collapse into one element of heterogeneity. In this case, the model of education will not satisfy the Athey and Imbens (2006) assumptions and a difference-in-differences analysis of an intervention will not have a causal interpretation.

To make things worse, if the outcome variable is discrete (such as "working" or "not working") then a point estimate in difference-in-differences can only be achieved by assuming a functional form: the literature is replete with linear probability models estimating impacts using difference-in-differences. These models look simple and straightforward, but the interpretation of their results as causal impacts rely on strong behavioral and functional-form assumptions. In contrast, results from randomized evaluations "only" rely on the integrity of the experiment itself, including of course, the absence of spillovers.

A further issue is the local nature of the results when the impacts of a policy are heterogeneous. This is best illustrated by the regression discontinuity approach, which identifies impacts for individuals who happen to be located close to the discontinuity. Thus while regression discontinuity has some qualities of a randomized experiment (in the sense that being on either side of the discontinuity is assumed effectively random), in contrast to the experimental approach, the impact is local to a very specific group of people defined by proximity to the discontinuity. These concerns are more broadly relevant for quasi-experimental approaches as discussed in Imbens and Angrist (1994) and in Heckman and Vytlacil (2005).

In short, randomized experiments provide causal effects without having to refer to a specific economic model or structure. Quasi-experimental approaches on the other hand, while not focusing on structural parameters, rely on underlying assumptions about behavior that potentially limit the interpretability of the results as causal. The attraction of these approaches is their simplicity. However their usefulness is limited by the lack of a framework that can justify external validity, which in general requires a more complete specification of the economic model that will allow the mechanisms to be analyzed and the conclusions to be transferred to a different set of circumstances. This is one of the key advantages of structural models: they describe the mechanisms through which effects operate and thus provide the framework for understanding how a particular policy may translate in different environments.

Combining Randomized Experiments and Structural Modeling

A combination of a fully specified model and randomized experiments can enhance analysis in ways that either of the two approaches alone would miss. Indeed, one of the most important recent advances in empirical economics uses dynamic structural models with exogenous sources of variation. The idea is of course not new and goes back at least to Orcutt and Orcutt (1968) as well as to the evaluation of the Gary negative income tax experiment in Burtless and Hausman (1978). Also, Rosenzweig and Wolpin (1980) combine information from quasi-experimental variation to infer structural relations in a twins study to analyze the quality–quantity model of fertility.

The renewed interest in this approach brings together the advantages of credible evaluation that relies on randomized experiments or (arguably) exogenous variation induced by policy changes, with the systematic economic analysis of structural models. A couple of prominent examples include Blundell, Duncan, and Meghir (1998) who use changes in the structure of wages and tax policy reforms to identify a partial model of labor supply, and Kaboski and Townsend (2011) who estimate a model of household investment and borrowing and validate its predictions using Thai data drawn from an expansion of microfinance availability in a large set of villages.

Experimental evidence can be used either to validate a structural model or to aid in the estimation process. These two alternative ways of using the same experimental evidence can be illustrated by comparing Todd and Wolpin (2006) and Attanasio, Meghir, and Santiago (2012). In 1998, the Mexican government experimented with a conditional cash transfer program whose intention was to increase the school participation of children in poor rural areas and improve preventive health care participation by mothers. PROGRESA, as the program became known, was to be evaluated by a cluster randomized control trial. Out of a population of 506 poor rural communities, 320 were assigned to receive the program immediately, while the remaining ones were kept back as a control, only receiving the program two years later. PROGRESA consisted of offering nutritional supplements to young children and a subsidy to families (disbursed to the mother) conditional on children's attendance at school. Mothers had to attend health clinics regularly to be eligible.

The intervention was highly successful. Schultz (2004) carried out the main evaluation of the program and shows that schooling participation increased. But can we learn more from the experiment and the associated data than the magnitude of the treatment effect? Specifically, can we say something about the design of the program and more generally, something about how costs of schooling affect educational participation?

In a standard economic model, the conditional school grants change school participation by counteracting the opportunity cost of schooling. Todd and Wolpin (2006) use this insight to validate a model of education attendance and fertility. They estimate the model based on data from the control group only. They then predict the impact of the experiment by reducing the wage in the model by an amount equivalent to the grant when the child went to school. They thus use the experiment to validate a dynamic model, which as specified, is identifiable from the control data only.

Using data from the experiment, Attanasio, Meghir, and Santiago (2012) identify a richer model (in some dimensions) that implies a more general cost-of-school function. Like Todd and Wolpin (2006), they set up a forward-looking model of educational choice through high school, where the individuals and their families decide each period whether to attend school. The benefits of schooling accrue in the future through better labor market opportunities, identified by the observed schooling attendance in the control group. A more general specification would use observations on the subsequent career to improve identification. The cost of schooling is affected by four elements: 1) forgone child labor income; 2) the amount of the PROGRESA grant for which the individual is eligible if attending school, which varies by the age of the child; 3) past school attendance, which may reduce cost because of habits or because past learning makes schooling now easier; and 4) an unobserved cost of attending school associated with the child's scholastic ability.

This structural model is explicitly dynamic: each year of schooling adds to human capital and future standards of living; there is uncertainty over whether the child will pass the grade; the grant is only available up until age 18; and current attendance affects future costs. The key point is that the model in Attanasio, Meghir, and Santiago (2012) allows the PROGRESA grant to have a different effect from the wage: the authors use the experimental variation to identify this extra effect. The finding is that the impact of the grant is larger than that predicted by the changes in school attendance as a function of forgone wages. This finding poses important questions of interpretation, but it highlights that the experiment allows the model to be extended and to address directly whether the grant has a different effect from the standard opportunity cost. The use of the experiment allowed the relaxation of some of the restrictions from economic theory, thus broadening the scope of the model and the interpretation of the experimental results. A further development of the model would require explaining why the grant has a different impact than forgone wages. Attanasio, Meghir, and Santiago speculate that this has to do with intrahousehold allocations and the fact that the grant goes to women. From the point of view of the discussion here, progress in understanding would require adding such an intrahousehold component and thinking about ways to identify it.

Of course, it is important not to overstate the synergies between structural models and experiments: in most cases, randomized experiments only offer discrete sources of variation—policy is on or off—which is far from the requirements for identification in dynamic models, which would typically require continuous variation (Heckman and Navarro 2007).

The above example illustrates how the experiment can add to the identification potential of the structural model. But what does the structural model add to the experiment? We know how the experiment affected school attendance at various ages. What does the model offer in addition to this finding, and what are the assumptions on which any additional insights are based?

The basic gain from using the structural models is that they allow a better understanding of the mechanisms and analysis of counterfactuals. Attanasio, Meghir, and Santiago (2012) focus on counterfactuals: for example, we can ask whether the grant, which varies by age of the child, would be more effective if structured differently over age, holding total financial cost constant. In terms of mechanisms, Attanasio, Meghir, and Santiago discuss the potential role of intrahousehold allocations; but the age limitation of the grant is an important factor in its effect. They also estimate whether the impact on wages resulting from the change in child labor supply dampened significantly the effect of the program—it did not. A richer model could look at how the program affected risk and risk sharing in the community, thus changing decisions including that of school attendance. Todd and Wolpin (2006) also investigate impacts on fertility. These rich behavioral models offer a deeper insight of just how an intervention can affect the final outcome. Understanding the mechanisms is central to designing policies and avoiding unintended effects as well as for building a better understanding of whether a policy can reasonably be expected to work at all.

The extra richness offered by the model does not come for free. We need to make additional assumptions that were not required for the simple experimental evaluation. Consider the counterfactual that restructures the PROGRESA grant by adjusting the amounts offered at each age. The ability to assess this proposal depends on knowing how education participation varies with the grant at different ages. The amount of the grant did vary with age; however, each age is associated with just one amount mandated by the program—there is no age-by-age experimental variation (although conceptually there could be). To recover how the effect of the grant varies by age, we need to assume that this effect varies smoothly and does not follow the exact pattern of variation of the grant by age. One can be understandably skeptical of results that rely on untestable assumptions about preferences. However, the assumptions in these models tend to be explicit, so promoting transparency and allowing for explicit criticism and improvements. For the purpose of this paper, it provides a good example of the types of assumptions that often need to be made to extend the narrow conclusions of an experiment to a broader context. In a more complex experiment, one can imagine that the amounts themselves within the experiment would be randomized at each age—thus offering stronger identification of this effect. In practice, it is very hard to implement experiments that are complex enough to offer variation in all the directions required for identification of all desired insights and still have sample sizes to allow sufficient statistical power.

Structural models include further restrictions. For example, they often require assumptions about the distribution of random preferences. In Attanasio, Meghir, and Santiago (2012), it is assumed that psychic costs of education can be described by a mixed logistic distribution. A central question in this literature is whether such assumptions are needed, or if they could be relaxed with richer data. In an enlightening paper, Magnac and Thesmar (2002) argue convincingly that a dynamic discrete choice model, such as the one in Attanasio, Meghir, and Santiago (2012), does depend on identifying assumptions relating to the distribution of preferences. The reason is quite intuitive because all outcome variables are discrete.

More can be achieved in models with continuous variables: Heckman and Navarro (2007) develop identification in a dynamic Roy model of education and wages. As they emphasize, the key to identification of the dynamic model is that they use information on measured consequences of treatment—for example, on wages. They show that identification restrictions can be relaxed if one observes explicitly a continuous outcome variable, such as the wage rate, and if the dynamic discrete choice depends on some continuous variable with large support, such as school fees (see also Meghir and Rivkin 2011). In practice, such conditions are usually not met and the functional form restrictions will play a role in analyzing the actual dataset. Heckman and Navarro (2007) also emphasize the use of cross-equation restrictions implied by the theoretical structure of the model. Here there is an important distinction to be made between restrictions on the shape of distributions of unobservables, for which there is rarely any theory, and restrictions that follow a clear reasoning and foundation in theory. While we should minimize ad hoc restrictions, it is also important to realize that empirical analysis can never do away with theoretical foundations and still remain useful as a learning tool. Data from a randomized experiment can however be very helpful here, either in showing that despite the various functional form assumptions, the model matches the unbiased results of the experiment, or in using the experiment to ensure that the resulting estimates reproduce the impacts. In this sense, the experiment can aid in identification.

Combining randomized experiments and credible quasi-experimental variation with structural models seems to bring together the best of both approaches of empirical economics: it identifies causal effects non- or semi-parametrically for specific policies, provides useful identifying information for the structural model, and offers a coherent way for understanding mechanisms and counterfactuals through the organizing lens of economic theory. This approach is growing in influence: beyond the papers already cited, Duflo, Hanna, and Ryan (2012) use a structural model to analyze the results of a school monitoring experiment in India; Kaboski and Townsend (2011) combine information from quasi-experimental evidence from the Thailand Million Baht Village Program with a structural model of small family businesses to understand the mechanisms underlying the workings of microfinance (see also Garlick 2016); and Voena (2015) uses differences-in-differences to evaluate the effect of divorce laws on household behavior and then uses this data to fit a dynamic intrahousehold model with limited commitment in order to analyze policy counterfactuals.

Solving Structural Models

The specification issues discussed above are driven both by the importance of a well-focused and empirically identified economic model as well as by computational feasibility. There have been huge advances in both computational methods and power over the last 30 years allowing much more flexibility in what can be implemented in practice. However, computational constraints remain and to some extent will always be with us. In this section, we discuss computational issues relating to solving these models, which is where most of the difficulty lies. We use Low, Meghir, and Pistaferri (2010) loosely as a case study and discuss in particular the computational implications of relaxing the separability assumptions.

In some situations, structural estimation is simple and relies on linear methods: for example, estimating demand systems in static models or estimating Euler equations with complete markets as in Altug and Miller (1990) or even with incomplete ones as in Meghir and Weber (1996). But more often than not, structural models and particularly dynamic stochastic models involve nonlinear estimation, and require numerical methods to solve the model to generate moments for estimation. The greatest "entry cost" for a researcher wishing to estimate dynamic structural models is learning to solve such models accurately and efficiently. For a broader textbook discussion of these methods, useful starting points include Adda and Cooper (2003) and Miranda and Fackler (2002). Some more recent, faster methods are due to Carroll (2006), Fella (2014), and Barillas and Fernández-Villaverde (2007).

The heart of solving dynamic structural models is the computation of value functions and corresponding decision rules. The value functions associate a numerical value to a decision or set of decisions, conditional on the relevant state variables, and conditional on all future decisions being optimal. The decision rules describe how individuals behave following different realizations of the economic environment. The state variables describe the economic environment. These include variables that are independent of past choices by the agent (and so are treated as exogenous) as well as variables that evolve depending on past decisions (and so are endogenous).

The discussion below relates to finite horizon life-cycle models. In these models, age is part of the state space—which means that the value functions are not stationary. There is a class of structural dynamic models with infinite horizons in which the value functions are stationary, not depending on age. Equilibrium search models of the labor market are usually specified in this way for purposes of convenience. The solution methods for these are related but different, and not touched upon here.

In Low, Meghir, and Pistaferri (2010), the decision rules describe whether an individual at each age would choose to work and how much the individual would save. Decision rules are obtained by comparing the value functions derived from different choices, at a given state of the economic environment. The state variables are wealth, individual productivity, and the matched firm type. The model makes numerous separability assumptions, as discussed above, especially over fertility and marriage: neither children nor marriage are considered choices in the model, and preferences over consumption and employment are assumed to be separable from marriage and fertility. Relaxing these assumptions expands the choice set, increasing the number of value function comparisons. It would also expand the state space to include current marital status, details of any partner, as well as family size, increasing the description of the economy and increasing the number of points at which decision rules have to be solved. The value of the separability assumptions therefore is in reducing the computational burden, as well as in making the model less opaque.

Armed with these decision rules, the researcher can then simulate behavior. There are four core steps in solving a dynamic structural economic model.

In the first step, the points (or nodes) of the state space at which the model needs to be solved are specified. Numerical solution requires defining the bounds of each of the variables, so that one can then think of a multidimensional grid of state variables. The state space fully specifies all aspects of the economic environment which affects the particular choices being analyzed in the model. In Low, Meghir, and Pistaferri (2010), the grid of specific values of the exogenous state variables, such as the permanent wage, are set before solving the model, using approximations to transition probabilities as in Adda and Cooper (2003, chapter 3). For endogenous state variables, such as with wealth, restricting the number of values to a discrete set would restrict the choice set in an arbitrary way. In Low, Meghir, and Pistaferri, this would have meant a discrete set of consumption values, which would introduce jumps in behavior that are not observed in the data. We keep these endogenous state variables continuous. Discretization can nonetheless be used to determine the points where the model is actually solved and then interpolation can be used between these points.

There are many alternative ways to interpolate: linear interpolation is usually robust, particularly when decision rules are not smooth, and in Low, Meghir, and

Pistaferri (2010), we start with this approach. Other methods of interpolation include imposing assumptions about smoothness in the decision rules, which then allows fitting either higher-order local splines or, alternatively, polynomials across the whole state space to provide a global approximation. The tradeoff here is that linear interpolation tends to require more points spanning the state space. Alternative methods need fewer points in the state space but impose further restrictions on the form of the solution.

Carroll (2006) and Fella (2014) discuss using endogenous grid points which are relevant for endogenous state variables. Getting the minimum value of the grid right requires care: in Low, Meghir, and Pistaferri (2010), the lower bound on assets is determined by an exogenously set borrowing limit. Alternatively, the lower bound can be determined by a "no-bankruptcy condition," specifying that borrowing has to be limited to what can be repaid with certainty—a "natural" liquidity constraint.

In the second step, we specify a terminal condition defining the continuation value in the subsequent periods beyond which we do not model decisions. In Low, Meghir, and Pistaferri (2010), the terminal condition is death, but it does not have to be: in Attanasio, Meghir, and Santiago (2012), the terminal condition is defined by the oldest age the child could attend high school, taken to be 18. In general the terminal value is a function of the state variables at that point. In Attanasio, Meghir, and Santiago (2012), the state variable is whatever schooling the child has accumulated by that age. The structure of this terminal value function is either tied directly to the model, with no new parameters, or it needs to be estimated with the rest of the model. Choosing an appropriate terminal point consistent with the model can economize on parameters to be estimated and improve identification. For example, if it is reasonable to assume that no individual lives beyond say 110 and that there are no bequests other than accidental ones, then the terminal value is defined explicitly by the problem and no extra arbitrary modeling assumptions have to be imposed.

In the third step, the "value function" at each node in the state space specified on the grid is solved, starting with the terminal period. The solution involves solving a numerical optimization problem, or a nonlinear equation solution to a first-order condition. The model in Low, Meghir, and Pistaferri (2010) contains a mixture of discrete and continuous choices: over whether to participate in the labor market, and over how much to save. This combination of discrete and continuous choices raises a problem, because changes in asset holdings can lead to changes in participation status and to jumps in the decision rule for consumption. We deal with this nonconcavity by solving for value functions conditional on the discrete choice, and then taking the maximum over these. The number of conditional value functions to solve increases with the number of discrete choices. However, these conditional value functions may themselves not be concave. The solution, numerically, is that if there is "enough" uncertainty about changes in future prices or wages then the expected value function will typically be concave. Nevertheless, this can rarely be proved and depends on the amount of uncertainty in the model. In practice, this means that we need to investigate numerically whether multiple solutions occur.

The fourth and final solution step is to iterate backwards one period at a time, at each period solving for each point in the state space. The solution in earlier periods will be determined taking account of expectations about future outcomes (based on the distribution of potential shocks) and also how the individual will respond in the future to those outcomes (based on the already-solved future decision rules. Because expectations have to be calculated, this involves numerical integration over the unknown random variables: for example, in Low, Meghir, and Pistaferri (2010), these are shocks to wages, job offer arrivals, and firm types. The more underlying random variables are involved, the higher the dimension of integration and consequently the computational costs can rise exponentially. This factor limits in practice the amount and source of uncertainty that one can introduce in a model. Notice also that the distribution of the shocks may depend on past realizations (rather than being independent and identically distributed). For example, if shocks to wages are serially correlated, the realization of a future shock will depend on the value of the current shock. This means that the current shock to wages is in the state space: an extra exogenous continuous state. For this reason, the way we specify the distribution of random events is very important in keeping the problem tractable.

The decision rules solved in this way by backward induction specify what an economic actor will choose given any particular realization of the state of the world. These decision rules will then be combined with particular randomized realizations of the stochastic variables starting at an initial period and simulating forwards. In Low, Meghir, and Pistaferri (2010), the randomized realizations are of permanent shocks to wages, and of wage offers, firm type, and job destruction. These stochastic shocks are responsible for life-time career paths being so different for what otherwise appear to be identical individuals. Inputting one complete set of realizations of these stochastic variables into the decision rules generates the life-cycle path for consumption and labor supply for one individual. This calculation is then repeated a number of times to generate average life-cycle profiles, along with other moments that are needed. We return below to the issue of the number of repetitions when discussing implementation of this approach using the simulated method of moments.

Using Method of Moments for Estimation of Structural Models

The numerical solution is used to generate predictions about behavior for a given set of parameter values. These parameter values need to be estimated.

Estimation of dynamic structural models involves nonlinear optimization with respect to the unknown parameters. However, the key difficulty with this estimation is that we cannot express analytically the functional relationship between the dependent variables and the unknown parameters. In order to see how a change in a parameter changes the dependent variable, the entire model solution needs to be generated afresh. If solving the model once already takes time, the problem is compounded by estimation that requires solving the model repeatedly. Moreover, numerical approximation errors in the solution of the model can compound the estimation complexity. There is an active literature on the way to approach the problem: one is the nested fixed point algorithm (Rust 1987), where the model is solved for each set of parameters that are tried out by the numerical optimization algorithm. A recent alternative, which under certain circumstances can be faster, is the method of Su and Judd (2012).

Beyond the choice of algorithm for optimization, another important choice is the criterion function to be optimized as a function of the parameters. Traditionally, maximum likelihood was used for estimating structural models. This approach is most efficient, exploiting all the information in the specification. However, constructing the likelihood function is impossible or computationally intractable for many models. Estimation now typically uses the method of moments (or indirect inference) (for a formal discussion, see McFadden 1989; Pakes and Pollard 1989; Gourieroux, Monfort, and Renault 1993). For our purposes, we use the term moments in a broad sense to mean any statistic of the data whose counterparts can be computed from model simulations for a given set of model parameters. For example, moments include means, variances, and transition rates between states, as well as regression coefficients from simple "auxiliary" regressions.

With the method of moments, it is easier to tell which features of the data identify which structural parameters. Further, use of multiple datasets is straightforward and the researcher can put emphasis on fitting moments central to the analysis. Finally, this method eliminates the computational burden of using enormous administrative datasets with millions of observations: the data moments need only be computed once; the computational burden will then be due exclusively to the time it takes to solve the model. The downside of this approach is that it does not use all information in the data, and we do not have an easily implementable way of defining which moments need to be used to ensure identification. One must carefully define what are the key features of the data that will identify the parameters. Moreover, in finite samples, the results may be sensitive to the choice of moments.

In Low, Meghir, and Pistaferri (2010), we need to estimate the parameters governing the opportunity set, which include the wage process, job destruction, job arrival rates on and off the job, fixed costs of working, and the parameters, which include the discount rate, elasticity of intertemporal substitution, and disutility of work. Estimation of these parameters can be described in a five-step algorithm:

1) Start with an initial guess at a set of parameter values θ.

2) Numerically solve the model given the parameter vector θ (as described in the previous section).

3) If individuals are *ex ante* identical, simulate the careers of say S individuals using a random number generator for realisations of the stochastic variables, and construct moments from the simulated moments analogous to those constructed from the data. If individuals differ by exogenous observed factors, simulate S careers for each value of the exogenous initial conditions. Similarly if individuals differ by some unobserved characteristic (whose distribution is estimated together with the rest of the model) again simulate S careers for each point of support of the unobservable and then take suitable weighted averages when constructing the moments.

4) Calculate the "criterion function" being minimized. This may be a simple or weighted quadratic distance between the data and the simulated moments.

5) Update the set of parameters θ to minimize the criterion function and return to step 2 and numerically solve the model with the updated parameters.

There are many decisions in implementing this algorithm. Here we discuss the main ones: what parameters to estimate, what moments to use, how to weight the moments, how to optimize to minimize the criterion, and post-estimation, what checks to carry out.

Choosing the Parameters to Estimate

A fully specified economic model requires that all parameters governing the opportunity set and preferences be determined, which can often make the problem unmanageable. The set of parameters can be divided into three: First, some parameters of the economic environment can be obtained directly from the institutional setting or data, requiring the assumption that the particular aspect of the environment is not affected by economic choices made within the model. For example, in Low and Pistaferri (2015), the specification of how health shocks evolve was estimated directly from the data, requiring the assumption that labor supply and other choices did not affect health.

Second, some parameters can be obtained using a partially specified model. Parameters estimated in this way are robust to details of the fully specified structural model. For example, Attanasio, Levell, Low, and Sanchez-Marcos (2017) use an Euler equation to estimate the elasticity of intertemporal substitution to use in a fully specified model. Low, Meghir, and Pistaferri (2010) and Low and Pistaferri (2015) estimate the wage process using a reduced-form procedure with the residuals identifying the wage uncertainty. The disadvantage of the procedure is that estimation is not completely in tune with the fully specified model. However, what may seem to be a shortcoming can also be an advantage: using the partially specified model means many auxiliary assumptions are not imposed on all components of the model.

Finally, parameters that are the key drivers of the economic choices in the model form part of the full structural estimation. In Low, Meghir, and Pistaferri (2010), these parameters were the disutility of work, the fixed cost of work, and job market frictions. In Low and Pistaferri (2015), these parameters also included the acceptance probabilities onto disability insurance.

Selecting Moments

More moments are not necessarily helpful in practice: moments need to be economically important to the model and informative about parameters. In Low, Meghir, and Pistaferri (2010), key moments were employment rates and unemployment duration at different ages. Employment rates were related to fixed cost of work, and durations were related to job arrival rates, although both sets of parameters affect both moments through the structure of the model. Moments used may include reduced form regressions, population means, or elasticities from the literature. Low and Pistaferri (2015) use coefficients from a regression of consumption on health status as moments to inform how health shocks affect the marginal utility of consumption. Other important moments may be transition rates, dispersion, and the time series properties of wages.

Simulating these moments in step 3 of the algorithm above requires randomly generated variables to represent the exogenous stochastic processes in the structural model. For each individual simulated, there is a random realization for each stochastic process. The complete set of random numbers for all individuals should be generated only once at the start of the estimation, and the same set of random numbers should be used in each iteration of the criterion function. As the number of simulations increases to infinity, the simulation error goes to zero, implying the moments become equal to the theoretically implied ones. At this point we are only left with the usual sampling error from the data. In general, due to the number of simulations being finite, simulation error should be taken into account in computing the standard errors of the estimated parameters.

The distributions of the stochastic processes may depend on parameters that need to be estimated. In order to make sure the underlying random draws are the same across iterations we need to draw uniform (0, 1) random variables that can then be transformed to follow whatever distribution the model implies (for example, $N(0, \sigma^2)$ where σ^2 is estimated).

Weighting the Moments

Moments may not be of equal economic importance, or measured with equal precision, or measured in comparable units. These considerations determine the choice of weighting matrix on the moments. Alternatives are the inverse of the full variance–covariance matrix, the inverse of the diagonals of the variance–covariance matrix, the identity matrix, or conversion of deviations into percentage deviations.

The "optimal weighting matrix" is the inverse of the variance–covariance matrix of the moments. This puts greater weight on more precisely estimated moments, and corrects the weighting on moments that are correlated. Ruge-Marcia (2012) shows the advantages of this weighting in a Monte Carlo exercise. However, with small samples, Altonji and Segal (1996) emphasize that the identity matrix (that is, equal weighting) may be the best choice because using hard-to-estimate higher-order moments of the data that constitute the weight matrix may actually introduce substantial bias. Equal weighting does not differentiate the precision with which each moment is estimated, and the units of measurement affect the weighting. The moments can be normalized to convert the difference between moments into the percentage deviation, which is equivalent to using a matrix of the inverse of the moments in the data squared. An alternative is the inverse of the diagonals of the variance–covariance matrix, but the issue remains that the more precisely measured moments get more weight, regardless of how important the moments are for the question at hand.

In Low, Meghir, and Pistaferri (2010), labor participation rates are precisely measured, whereas duration-of-unemployment numbers are imprecisely measured.

Weighting based on precision with which moments are estimated would have meant durations would fit poorly, reducing the relevance of the model. In that study, we reduce the scope of this problem by using only economically relevant moments of the data and converting the moments to be percentage deviations.

Optimization with Simulated Moments

Simulated moments are often not smooth with respect to the parameters and as a result, derivative-based methods of optimization are often inappropriate. A straightforward method is Dantzig's classic simplex method. The simplex method is derivative-free and while it can be computationally slow, it is robust. Recently, Markov Chain Monte Carlo methods for optimization have become more common. This approach requires no more than simulating the model and computing moments given a set of parameters. Chernozhukov and Hong (2003) have shown how Markov Chain Monte Carlo can provide estimators that are asymptotically equivalent to minimizing the method-of-moments criterion. While the Markov Chain Monte Carlo can be slow to converge on some occasions, in practice other alternatives may be much worse. Many researchers make use of parallel computing with multiple chains running at the same time.

Standard Errors and Post-Estimation Checks

After the parameters are estimated in a structural economic model, the list of tasks is not yet complete: additional checks are needed.

First, calculate parameter standard errors. Various papers on simulated method of moments and indirect inference like McFadden (1989), Pakes and Pollard (1989), and Gourieroux, Monfort, and Renault (1993) provide the appropriate results. A practical difficulty is that these approaches require derivatives of the moments with respect to the parameters of the model. Another difficulty arises from the fact that the estimation error in pre-estimated parameters also needs to be taken into account. This correction can become computationally hard.

Second, use the finite difference approach in the first check to show how moments change with estimated structural parameters. This information helps make estimation more transparent, by showing which parameters are pinned down by which moments.

Third, show the 95 percent confidence interval for the difference between each simulated moment and its data counterpart. This provides a metric for judging how well moments fit. In Low and Pistaferri (2015), for example, the model could not match the participation rates of the healthy who were over 45.

Finally, consistency should be checked between any estimates from the partially specified or pre-estimation stage and the implications of the fully specified model. In Low and Pistaferri (2015), the wage process was pre-estimated with a reduced form selection correction. Data on the simulated individuals was used after the estimation to check consistency of the full model with the selection model. A further test compares simulated predictions with additional moments or reduced form evidence, preferably not targeted in estimation. An important validation of Low and Pistaferri was to compare the simulated elasticities of the receipt of disability insurance with respect to generosity to the reduced form estimates in the literature.

The ultimate purpose is to produce an estimated model that is internally consistent, so the estimates can be used for counterfactual analysis. Being explicit about each of these steps can help to provide transparency about the mechanisms and the sources of identification.

Conclusion

Structural economic models are at the heart of empirical economic analysis, offering an organizing principle for understanding data, for testing theory, for analyzing mechanisms through which interventions operate, and for simulating counterfactuals. It has been long understood that econometric identification of such models will necessarily depend on prior assumptions and on theory; but without the organizing device of theory, it is impossible to make progress in our understanding. We argue that the resurgence and increased popularity of the idea of combining randomized experiments or plausible quasi-experimental variation together with structural economic models can strengthen the value of empirical work substantially. Indeed, researchers should think more ambitiously and use theory to define experiments that need to be run to test and estimate important models.

Structural economic models are difficult to use because of computational complexity. Moreover, it is easy to end up with overcomplicated and unwieldy models that offer little insight into mechanisms and whose identifiability is, to say the least, obscure. The trade-off between providing the necessary complexity to be economically meaningful and maintaining transparency is at the heart of good structural modeling. Our approach is to be explicit about what separability assumptions can be invoked: a fully specified structural model will not capture all choices, but will be explicit about which choices are part of the model and which choices are not, and will solve explicitly for all choices in the model. Choices can be left out of a model if they do not affect the choices we are modeling, due to separability in preferences.

With the increasing use of structural models and the progress of both computational power and numerical methods, the economics profession is becoming much more familiar and skilled in the specification and use of structural models. In our view, this all for the good, and it is hard to see how progress can be achieved without both sides of empirical work: experiments generating exogenous variation, and theory-based models.

■ *We thank Tom Crossley, Felix Grey, Gordon Hanson, Enrico Moretti, Ann Norman, and Timothy Taylor for excellent comments and suggestions. Costas Meghir has been supported by the Cowles Foundation and the Institution for Social and Policy Studies (ISPS) at Yale University. All errors and interpretations are our own.*

References

Abbott, Brant, Giovanni Gallipoli, Costas Meghir, and Gianluca L. Violante. 2013. "Education Policy and Intergenerational Transfers in Equilibrium." NBER Working Paper 18782.

Adda, Jérôme, and Russell W. Cooper. 2003. *Dynamic Economics: Quantitative Methods and Applications*. MIT Press.

Altonji, Joseph G. 1986. "Intertemporal Substitution in Labor Supply: Evidence from Micro Data." *Journal of Political Economy* 94(3, Part 2): S176–S215.

Altonji, Joseph G., and Lewis M. Segal. 1996. "Small-Sample Bias in GMM Estimation of Covariance Structures." *Journal of Business and Economic Statistics* 14(3): 353–66.

Altug, Sumru, and Robert A. Miller. 1990. "Household Choices in Equilibrium." *Econometrica* 58(3): 543–70.

Angelucci, Manuela, and Giacomo De Giorgi. 2009. "Indirect Effects of an Aid Program: How Do Cash Transfers Affect Ineligibles' Consumption?" *American Economic Review* 99(1): 486–508.

Arellano, Manuel, and Costas Meghir. 1992. "Female Labour Supply and On-the-Job Search: An Empirical Model Estimated Using Complementary Data Sets." *Review of Economic Studies* 59(3): 537–59.

Athey, Susan, and Guido W. Imbens. 2006. "Identification and Inference in Nonlinear Difference-In-Differences Models." *Econometrica* 74(2): 431–97.

Attanasio, Orazio P., Peter Levell, Hamish Low, and Virginia Sanchez-Marcos. 2017. "Aggregating Elasticities: Intensive and Extensive Margins of Female Labour Supply." Cambridge Working Papers in Economics 1711.

Attanasio, Orazio P., Costas Meghir, and Ana Santiago. 2012. "Education Choices in Mexico: Using a Structural Model and a Randomized Experiment to Evaluate PROGRESA." *Review of Economic Studies* 79(1): 37–66.

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

Barillas, Francisco, and Jesus Fernández-Villaverde. 2007. "A Generalization of the Endogenous Grid Method." *Journal of Economic Dynamics and Control* 31(8): 2698–2712.

Berry, Steven, James Levinsohn, and Ariel Pakes. 1995. "Automobile Prices In Market Equilibrium." *Econometrica* 63(4): 841–90.

Blundell, Richard, Martin Browning, and Costas Meghir. 1994. "Consumer Demand and the Life-Cycle Allocation of Household Expenditures." *Review of Economic Studies* 61(1): 57–80.

Blundell, Richard, Monica Costa Dias, Costas Meghir, and Jonathan Shaw. 2016. "Female Labor Supply, Human Capital, and Welfare Reform." *Econometrica* 84(5): 1705–53.

Blundell, Richard, Alan Duncan, and Costas Meghir. 1998. "Estimating Labor Supply Responses Using Tax Reforms." *Econometrica* 66(4): 827–61.

Blundell, Richard, and Ian Walker. 1986. "A Life-Cycle Consistent Empirical Model of Family Labour Supply Using Cross-Section Data." *Review of Economic Studies* 53(4): 539–58.

Bond, Stephen, and Costas Meghir. 1994. "Dynamic Investment Models and the Firm's Financial Policy." *Review of Economic Studies* 61(2): 197–222.

Browning, Martin, and Costas Meghir. 1991. "The Effects of Male and Female Labor Supply on Commodity Demands." *Econometrica* 59(4): 925–51.

Burdett, Kenneth, and Dale T. Mortensen. 1998. "Wage Differentials, Employer Size, and Unemployment." *International Economic Review* 39(2): 257–73.

Burtless, Gary, and Jerry A. Hausman. 1978. "The Effect of Taxation on Labor Supply: Evaluating the Gary Negative Income Tax Experiment." *Journal of Political Economy* 86(6): 1103–30.

Carroll, Christopher D. 2006. "The Method of Endogenous Gridpoints for Solving Dynamic Stochastic Optimization Problems." *Economics Letters* 91(3): 312–20.

Chernozhukov, Victor, and Han Hong. 2003. "An MCMC Approach to Classical Estimation." *Journal of Econometrics* 115(2): 293–346.

Chiappori, Pierre-Andre, Monica Costa Dias, and Costas Meghir. Forthcoming. "The Marriage Market, Labor Supply and Education Choice." *Journal of Political Economy.*

Crossley, Thomas F., Hamish Low, and Sarah Smith. 2016. "Do Consumers Gamble to Convexify?" *Journal of Economic Behavior and Organization* 131: 276–91.

Duflo, Esther, Rema Hanna, and Stephen P. Ryan. 2012. "Incentives Work: Getting Teachers to Come to School." *American Economic Review* 102(4): 1241–78.

Feldstein, Martin. 1995. "The Effect of Marginal Tax Rates on Taxable Income: A Panel Study of the 1986 Tax Reform Act." *Journal of Political Economy* 103(3): 551–72.

Fella, Giulio. 2014. "A Generalized Endogenous Grid Method for Non-smooth and Non-concave Problems." *Review of Economic Dynamics* 17(2):

329–44.

Garlick, Julia. 2016. “Essays in Development Economics.” Yale PhD thesis.

Gorman, W. M. 1995. *Collected Works of W.M. Gorman*. Edited by C. Blackorby and A. F. Shorrocks. Oxford University Press.

Goolsbee, Austan, Robert E. Hall, and Lawrence F. Katz. 1999. “Evidence on the High-Income Laffer Curve from Six Decades of Tax Reform.” *Brookings Papers on Economic Activity* no. 2, pp. 1–64.

Gourieroux, Christian, Alain Monfort, and Eric Renault. 1993. “Indirect Inference.” *Journal of Applied Econometrics* 8(S1): S85–S118.

Gruber, Jon, and Emmanuel Saez. 2002. “The Elasticity of Taxable Income: Evidence and Implications.” *Journal of Public Economics* 84(1): 1–32.

Hansen, Lars P. 1982. “Large Sample Properties of Generalised Method of Moments Estimators.” *Econometrica* 50(4): 1029–54.

Heckman, James J., Robert J. LaLonde, and Jeffrey A. Smith. 1999. “The Economics and Econometrics of Active Labor Market Programs.” Chap. 31 in *Handbook of Labor Economics,* vol. 3, part A, edited by David Card and Orley Ashenfelter, pp. 1865–2097.

Heckman, James J., Lance Lochner, and Christopher Taber. 1998. “Explaining Rising Wage Inequality: Explorations with a Dynamic General Equilibrium Model of Labor Earnings with Heterogeneous Agents.” *Review of Economic Dynamics* 1(1): 1–58.

Heckman, James J., and Salvador Navarro. 2007. “Dynamic Discrete Choice and Dynamic Treatment Effects.” *Journal of Econometrics* 136(2): 341–96.

Heckman, James J., and Richard Robb Jr. 1985. “Alternative Methods for Evaluating the Impact of Interventions: An Overview.” *Journal of Econometrics* 30(1–2): 239–67.

Heckman, James J., and Edward Vytlacil. 2005. “Structural Equations, Treatment Effects, and Econometric Policy Evaluation.” *Econometrica* 73(3): 669–738.

Imbens, Guido W., and Joshua D. Angrist. 1994. “Identification and Estimation of Local Average Treatment Effects.” *Econometrica* 62(2): 467–475.

Imbens, Guido, Donald Rubin, and Bruce Sacerdote. 2001. “Estimating the Effect of Unearned Income on Labour Earnings, Savings and Consumption: Evidence from a Survey of Lottery Players” *American Economic Review* 91(4): 778–94.

Kaboski, Joseph J., and Robert M. Townsend. 2011. “A Structural Evaluation of a Large-Scale Quasi-Experimental Microfinance Initiative.” *Econometrica* 79(5): 1357–1406.

Keane, Michael P., and Kenneth I. Wolpin. 1997. “The Career Decisions of Young Men.” *Journal of Political Economy* 105(3): 473–522.

Koujianou-Goldberg, Pinelopi. 1995. “Product Differentiation and Oligopoly in International Markets: The Case of the U.S. Automobile Industry.” *Econometrica* 63(4): 891–951.

Lee, Donghoon, and Kenneth I. Wolpin. 2006. “Intersectoral Labor Mobility and the Growth of the Service Sector.” *Econometrica* 74(1): 1–46.

Low, Hamish. 2005. “Self-Insurance in a Life-Cycle Model of Labour Supply and Savings.” *Review of Economic Dynamics* 8(4): 945–75.

Low, Hamish, Costas Meghir, and Luigi Pistaferri. 2010. “Wage Risk and Employment Risk over the Life Cycle.” *American Economic Review* 100(4): 1432–67.

Low, Hamish, and Luigi Pistaferri. 2015. “Disability Insurance and the Dynamics of the Incentive–Insurance Trade-off.” *American Economic Review* 105(10): 2986–3029.

MaCurdy, Thomas E. 1983. “A Simple Scheme for Estimating an Intertemporal Model of Labor Supply and Consumption in the Presence of Taxes and Uncertainty.” *International Economic Review* 24(2): 265–89.

Magnac, Thierry, and David Thesmar. 2002. “Identifying Dynamic Discrete Decision Processes.” *Econometrica* 70(2): 801–816.

McFadden Daniel. 1989. “A Method of Simulated Moments for Estimation of Discrete Response Models without Numerical Integration.” *Econometrica* 57(5): 995–1026.

Meghir, Costas, and Steven Rivkin. 2011. “Econometric Methods for Research in Education.” Chap. 1 in *Handbook of the Economics of Education*, edition 1, vol. 3, edited by Erik Hanushek, Stephen Machin, and Ludger Woessmann. Elsevier.

Meghir, Costas, and Guglielmo Weber. 1996. “Intertemporal Nonseparability or Borrowing Restrictions? A Disaggregate Analysis using a U.S. Consumption Panel.” *Econometrica 64*(5): 1151–81.

Miranda, Mario J., and Paul L. Fackler. 2002. *Applied Computational Economics and Finance.* MIT Press

Orcutt, Guy H., and Alice G. Orcutt. 1968. “Incentive and Disincentive Experimentation for Income Maintenance Policy Purposes.” *American Economic Review* 58(4): 754–72.

Pakes, Ariel, and David Pollard. 1989. “Simulation and the Asymptotics of Optimization Estimators.” *Econometrica* 57(5): 1027–57.

Rosenzweig, Mark R., and Kenneth I. Wolpin. 1980. “Testing the Quantity–Quality Fertility Model: The Use of Twins as a Natural Experiment.” *Econometrica* 48(1): 227–40.

Ruge-Murcia, F. 2012. “Estimating Nonlinear

DSGE Models by the Simulated Method of Moments: With an Application to Business Cycles." *Journal of Economic Dynamics and Control* 36(6): 914–38.

Rust, John. 1987. "Optimal Replacement of GMC Bus Engines: An Empirical Model of Harold Zurcher." *Econometrica* 55(5): 999–1033.

Rust, John. 1992. "Do People Behave According to Bellman's Principle of Optimality." Working Papers in Economics E-92-10, Hoover Institution Stanford.

Schultz, T. Paul. 2004. "School Subsidies for the Poor: Evaluating the Mexican Progresa Poverty Program." *Journal of Development Economics* 74(1): 199–250.

Su, Che-Lin, and Kenneth L. Judd. 2012. "Constrained Optimization Approaches To Estimation of Structural Models." *Econometrica* 80(5): 2213–30.

Todd, Petra E., and Kenneth I. Wolpin. 2006. "Assessing the Impact of a School Subsidy Program in Mexico: Using a Social Experiment to Validate a Dynamic Behavioral Model of Child Schooling and Fertility." *American Economic Review* 96(5): 1384–1417.

Voena, Alessandra. 2015. "Yours, Mine, and Ours: Do Divorce Laws Affect the Intertemporal Behavior of Married Couples?" *American Economic Review* 105(8): 2295–2332.

Zeldes, Stephen P. 1989. "Consumption and Liquidity Constraints: An Empirical Investigation." *Journal of Political Economy* 97(2): 305–46.

Journal of Economic Perspectives—Volume 31, Number 2—Spring 2017—Pages 59–86

Twenty Years of Time Series Econometrics in Ten Pictures

James H. Stock and Mark W. Watson

Twenty years ago, empirical macroeconomists shared some common understandings. One was that a dynamic causal effect—for example, the effect on output growth of the Federal Reserve increasing the federal funds rate—is properly conceived as the effect of a shock, that is, of an unanticipated autonomous change linked to a specific source. Following Sims (1980), the use of vector autoregressions to estimate the dynamic causal effect of shocks on economic variables was widespread. There was also an understanding that vector autoregressions, because they impose as little structure on the data as possible, cannot answer questions about changes in policy regimes, such as the macroeconomic consequences of the Fed adopting a new policy rule. For such questions, more structured models grounded in economic theory are needed. At the same time, there was an increasing recognition that the available methods needed significant work. The schemes used to identify structural shocks in vector autoregressions were often seen as unconvincing by researchers outside the field, and the small structural models of the time were not econometrically estimated, miring that enterprise in an unhelpful debate over how to calibrate such models. In addition, there were chinks emerging in the theoretical econometric underpinnings of inference in time series data, as well as opportunities for using the much larger datasets becoming available, if only the tools to do so could be developed. The time was ripe for progress.

■ *James H. Stock is the Harold Hitchings Burbank Professor of Political Economy, Harvard University, Cambridge, Massachusetts. Mark W. Watson is Howard Harrison and Gabrielle Snyder Beck Professor of Economics and Public Affairs, Princeton University, Princeton, New Jersey. Their email addresses are james_stock@harvard.edu and mwatson@princeton.edu.*

† For supplementary materials such as appendices, datasets, and author disclosure statements, see the article page at
https://doi.org/10.1257/jep.31.2.59 doi=10.1257/jep.31.2.59

This review tells the story of the past 20 years of time series econometrics through ten pictures. These pictures illustrate six broad areas of progress in time series econometrics: estimation of dynamic causal effects; estimation of dynamic structural models with optimizing agents (specifically, dynamic stochastic equilibrium models); methods for exploiting information in "big data" that are specialized to economic time series; improved methods for forecasting and for monitoring the economy; tools for modeling time variation in economic relationships; and improved methods for statistical inference.

These pictures remind us that time series methods remain essential for shouldering real-world responsibilities. The world of business, finance, and government needs reliable information on where the economy is and where it is headed. Policymakers need analysis of possible policies, and macroeconomists need to improve their understanding of the workings of modern, evolving economies. Taken together, the pictures show how 20 years of research have improved our ability to undertake these professional responsibilities. These pictures also remind us of the close connection between econometric theory and the empirical problems that motivate the theory, and of how the best econometric theory tends to arise from practical empirical problems.

A review of 20 years of research must make some arbitrary decisions. One of our decisions is to focus on empirical macroeconomics, not finance. Fortunately, there are good surveys of the many developments in financial econometrics: for example, see the papers in Aït-Sahalia and Hansen (2010). Another concerns the choice of figures. Our ten figures are not meant to single out superstar papers (although some are) but rather to represent important lines of research: each figure illustrates a broader research program. In choosing these figures, we first looked for influential early papers from the late 1990s and early 2000s that framed subsequent research. This yielded five figures from papers with an average of 1,486 Google Scholar citations each. We then looked for figures more recently published that illustrate key findings or methods in a relatively mature line of research, yielding four more figures. Our final figure, which is not from published research, illustrates an open empirical challenge for research ahead.

Causal Inference and Structural Vector Autoregressions

An ongoing question in empirical macroeconomics is how to determine the causal effect of a policy change. For example, what is the effect of an autonomous, unexpected, policy-induced change in the monetary policy target rate—that is, a monetary policy shock—on output, prices, and other macro variables? The underlying problem is simultaneous causality: for example, the federal funds interest rate depends on changes in real GDP through a monetary policy rule (formal or informal), and GDP depends on the federal funds interest rate through induced changes in investment, consumption, and other variables. Thus, one cannot determine the effect of a change in the federal funds interest rate simply by using the rate (perhaps along with lagged values of the rate) as a right-hand-side variable in a regression to explain

GDP. Somehow, a researcher needs to isolate the exogenous variation in the federal funds interest rate, and for that you need external information.

Since the seminal work of Sims (1980), vector autoregressions have been a standard tool for estimating the causal effects over time of a shock on a given macro variable. This tool evolved into "structural" vector autoregressions, which are based on the idea that the unanticipated movements in the variables—that is, their forecast errors—are induced by structural shocks. The goal of structural vector autoregressions is to impose sufficient restrictions so that one or more structural shocks can be identified: specifically, that one or more shocks can be represented as an estimable linear combination of the forecast errors. The result of this analysis is the estimation of a dynamic path of causal effects, which in macroeconometrics is called a "structural impulse response function."

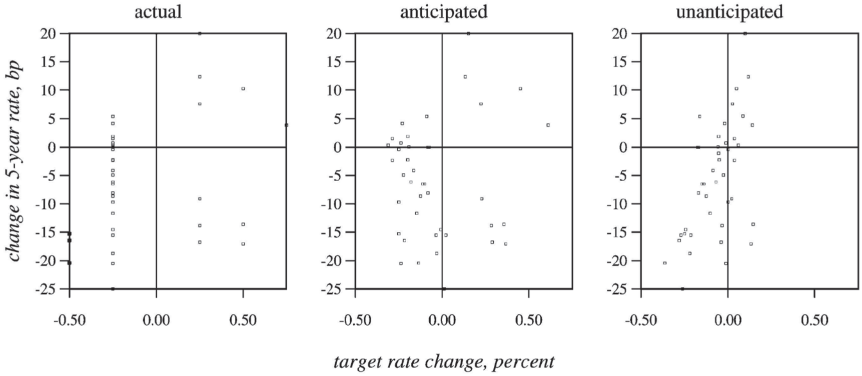
However, many applications of the original methods for identification of structural autoregressions that were dominant in the 1980s and 1990s have not withstood close scrutiny (as articulated, for example, by Rudebusch 1998). For example, a popular method for identifying monetary policy shocks in the 1980s and 1990s was to assume that economic activity and prices respond to a monetary policy shock with a lag, but that monetary policy responds systematically to contemporaneous nonmonetary shocks to the other variables. Under this assumption, the predicted value in a regression of the federal funds rate on its lags and on current and lagged values of the other variables is the endogenous policy response, and the residual is the unanticipated exogenous component—that is, the monetary policy shock. But this identifying assumption is not credible if the other variables include other asset prices, such as long-term interest rates.

Thus, this area needed new approaches. Broadly speaking, these new approaches bring to bear external information: information outside the linear system of equations that constitutes the vector autoregression. The development of new methods for estimating causal effects has been one of the main advances in microeconometrics over the past two decades (as discussed in several other articles in this symposium), and the focus on credible identification has parallels in the structural vector autoregression literature.

Using External Information to Estimate the Shock Directly

This brings us to our first picture, which is from Kuttner (2001). Kuttner's interest was in estimating the dynamic causal effect of a monetary policy shock on long-term interest rates, which is part of the broader program of estimating their dynamic causal effect on macroeconomic variables. Because the Fed controls the federal funds interest rate, one might initially think that the fed funds rate is exogenous; but not so, because some of the changes are responses to changes in economic activity which have their own effect on long-term interest rates. Rather, the exogenous part of the fed funds rate—the monetary policy shock—is the part that is not a response to economic activity. Kuttner's innovation was to draw on external information to identify the shock. Specifically, he knew that the Federal Reserve Open Market Committee announced its decisions at a specific time after its meetings, and he also had evidence (along with the theory of efficient financial markets) that the

Figure 1
Changes in the 5-year Treasury Rate and in the Target Federal Funds Rate on Federal Reserve Open Market Committee (FOMC) Announcement Dates
(fed funds changes are decomposed into anticipated and unanticipated components using changes in the fed funds futures market on announcement dates)



Source: Kuttner (2001), Figure 2.
Note: The figure shows changes in the 5-year Treasury rate (on the y-axes) and in the fed funds target rate (on the x-axes) on Federal Reserve Open Market Committee (FOMC) announcement dates. Unanticipated changes in the fed funds rate—which are the monetary policy shocks—are identified as changes in the fed funds futures rate from before to after the announcement of a change in the FOMC's target for the fed funds rate. The anticipated change is the actual change in the fed funds futures rate, minus the unanticipated change.

fed funds future rate was an efficient forecast of future fed funds rates. Thus, he was able to measure the unexpected part of the change in the federal funds futures rate as the change in the fed funds rate before and after the announcement. Assuming that no other relevant news was released during the announcement window, this change in the fed funds futures rate measures the change in market expectations of the fed funds rate resulting from the announcement—that is, it measures the monetary policy shock associated with the announcement. By using this external information, he could directly estimate the monetary policy shock.

Kuttner's figure (our Figure 1) shows that this unanticipated component of the change in the target rate is associated with changes in the five-year Treasury rate (right panel), but anticipated changes are not (center). As a result, there is no particular relationship between the actual announced target and the five-year rate (left). We interpret this figure as a compelling plot of the "first stage" in instrumental variables regression: it shows that an instrument (the unanticipated component of the target change on the announcement day) is correlated with an endogenous variable (the five-year interest rate).

The idea of using external information to identify shocks for structural vector autoregression analysis traces back to Romer and Romer (1989), who used textual and historical information to identify some exogenous monetary policy shocks. In addition to Kuttner (2001), Cochrane and Piazzesi (2002), and Faust, Rogers, Swanson, and Wright (2003), and Bernanke and Kuttner (2005) are early papers that use interest rate changes around Federal Reserve announcement dates to identify monetary policy shocks. In a similar spirit, Hamilton (2003) and Kilian (2008) use external information on international oil supply disruptions to estimate the effect of oil supply shocks on the economy.

This line of attack aims to measure the exogenous shock directly from external information, such as knowledge of the interest rate markets around announcement dates. If the shock can actually be measured, then estimation of structural impulse response functions is straightforward: because the shock is uncorrelated with other shocks, one can simply regress a variable of interest on current and lagged values of the shock, and the resulting coefficients trace out the dynamic causal effect (for example, Stock and Watson 2011, chap. 15). But doing so requires a particular strong form of external information: that the shock can be accurately measured.

Identification by External Instruments

If the external information succeeds in measuring only part of the shock or produces a noisy measurement of the shock, then the measured shock has the interpretation as an instrumental variable and regressions on the measure have the interpretation as the first stage in two-stage least squares. Arguably, many of the shock measures proposed to date yield imperfect measures. For example, changes in federal funds futures around an announcement reveal only a part of the monetary policy shock. In this case, the external shock measure is an instrumental variable: it is exogenous (that is, it is uncorrelated with other structural shocks) if properly constructed, and it is relevant because it is correlated with the true shock. Hamilton (2003) uses his measured international oil shock measure as instrument in a single-equation setting. In a vector autoregression, the technicalities differ from standard instrumental variables regression because the observed endogenous variables are forecast errors, not the original variables themselves. Still, the two criteria for a valid instrument, relevance and exogeneity, are the same in the structural vector autoregression application as in standard instrumental variable regression.

The explicit use of external instruments in structural vector autoregressions is fairly recent. This method is described in Stock (2008), Ramey (2016), and Stock and Watson (2016). Empirical applications of identification of structural impulse response functions using external instruments include Stock and Watson (2012a), Mertens and Ravn (2013), and Gertler and Karadi (2015).

Identification by Heteroskedasticity

Another method for identifying impulse response functions developed during the past 20 years exploits the observation that changes in the *variance* of the shocks can serve to identify the impulse response functions if those responses remain constant despite the heteroskedasticity of the shocks. Suppose that there are two known regimes, a high- and a low-volatility regime. Identification by heteroskedasticity works by generating two sets of moment equations, one for each regime. Although neither set can be solved on its own (the identification problem), assuming that the impulse response functions are the same across both regimes imposes enough parametric restrictions that together the two sets of equations can be solved, and thus the impulse response functions can be identified. This clever insight was developed for regime-shift heteroskedasticity by Rigobon (2003) and Rigobon and Sack (2003, 2004), and for conditional heteroskedasticity by Sentana and Fiorentini (2001) and Normandin and Phaneuf (2004). Lütkepohl (2013) offers a survey and discussion.

Identification by Sign Restrictions

An altogether different approach to identification in structural vector autoregressions is to use restrictions on the sign of impulse responses to identify the economic shocks. For many shocks, disparate macro theories often agree on the signs of their effects, at least over short horizons. Although several early papers build off this insight, the method developed by Uhlig (2005) is the most widely used. In his application, Uhlig restricted the impulse response with respect to a monetary policy shock identified by requiring that, on impact and over the next five months, the response of overall prices, commodity prices, and nonborrowed reserves to a contractionary monetary policy shock are not positive, and that the response of the federal funds interest rate is not negative. Identification using sign restrictions can be compelling and has been widely adopted.

At a mathematical level, using sign restrictions is fundamentally different than the other methods that identify shocks: with enough restrictions, those methods lead, in large samples, to a unique impulse response function, whereas the sign restrictions approach only determines a set that includes the impulse response. That is, sign-identified impulse response functions are not point-identified, but instead are set-identified.

Set identification of impulse response functions raises subtle issues of inference, which have only recently been appreciated. Following Uhlig (2005), the standard approach is Bayesian, but just as the identification scheme in classical structural vector autoregression methods can strongly influence results, the prior distribution over the unidentified region of the impulse response parameter space strongly influences Bayesian inference, even in large samples. These methods therefore require great care to produce transparent, valid, and robust inference. Recent papers tackling inference in sign-identified structural vector autoregressions are Fry and Pagan (2011), Moon, Schorfheide, and Granziera (2013), Giacomini and Kitagawa (2014), Baumeister and Hamilton (2015), and Plagborg-Møller (2016). For additional discussion and references to the recent methodological literature see Stock and Watson (2016, Section 4).

Estimation of Dynamic Stochastic General Equilibrium Models

Dynamic stochastic general equilibrium models are models of forward-looking, optimizing economic agents who live in an economy subject to unexpected shocks. The development of methods for solving and estimating these models, combined with their grounding in optimizing economic theory, has made them a central tool of monetary policy analysis at central banks.

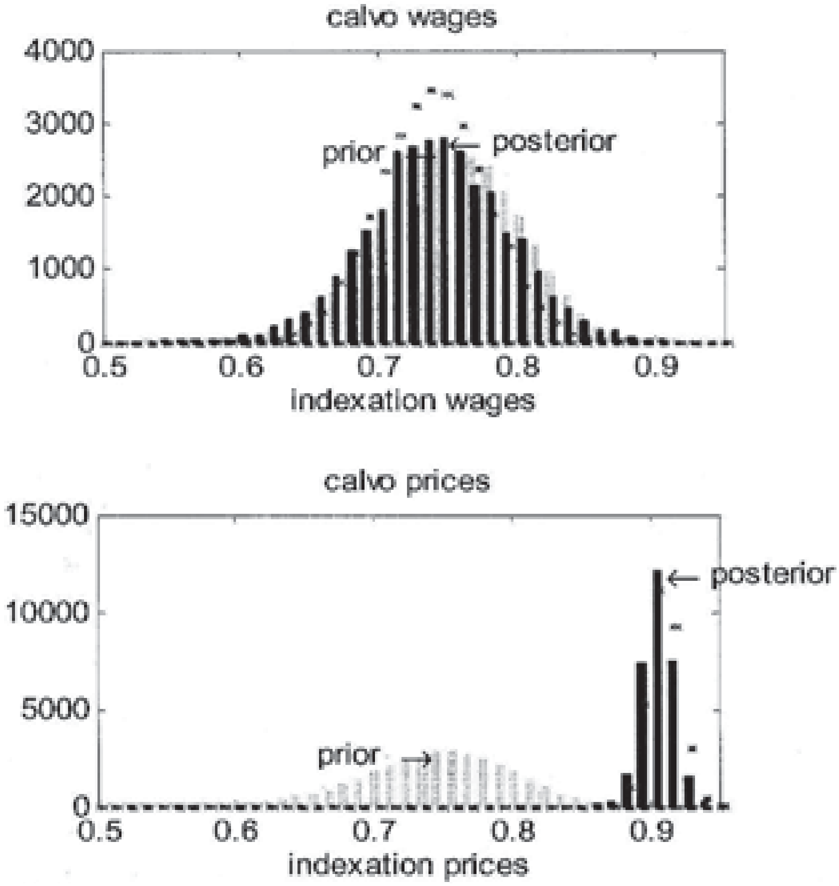
One of the first full-system estimations of a dynamic stochastic general equilibrium model was by Ireland (1997), who estimated a three-equation (GDP, prices, and money) system by maximum likelihood. However, maximizing the likelihood proves far more difficult numerically than averaging over the likelihood using a Bayesian prior, and today the dominant methods for estimating dynamic stochastic general equilibrium models are Bayesian. These methods were first used by DeJong, Ingram, and Whiteman (2000), Schorfheide (2000), and Otrok (2001) for small dynamic stochastic general equilibrium systems. Smets and Wouters (2003) showed that these methods can be applied to larger dynamic stochastic general equilibrium models that are rich enough to be a starting point for monetary policy analysis.

Figure 2, taken from Smets and Wouters (2003), represents the breakthroughs made over the past 20 years in the estimation of dynamic stochastic general equilibrium models. In their model, the "Calvo wage" parameter in the first panel is the probability that a worker's wage does not change, and the "Calvo price" parameter in the second panel is the probability that the firm's price does not change. As Figure 2 illustrates, the method works: The computational problems encountered when fitting dynamic stochastic general equilibrium models using frequentist methods such as maximum likelihood are sidestepped by computing posteriors, facilitated by a suite of tools developed in the modern Bayesian computational literature. For some parameters, such as the "Calvo price" parameter, the data are highly informative: incorporating the data results in much stickier prices than the authors' prior, so that the posterior and prior distributions are quite different. But for other parameters, such as the "Calvo wage" parameter, the data are much less informative, so that the prior and posterior essentially coincide. Thus, the Calvo wage parameter is in effect calibrated by the researcher, so the resulting complete model combines estimation where the data are informative with calibration where they are not.

This property of estimation *cum* calibration means that care needs to be taken in interpreting measures of uncertainty arising from the model. From a frequentist perspective, a classic justification of Bayesian methods is that coverage intervals ("Bayes credible sets") computed using the Bayesian posterior are essentially the same as frequentist confidence intervals in large samples, as long as a continuous prior does not rule out parameter values. (This is the celebrated Bernstein–von Mises theorem.) But for dynamic stochastic general equilibrium models, because the data are uninformative for some parameters—that is, some parameters are poorly identified—this equivalence does not hold and the uncertainty measures are heavily influenced by the shape of the prior. We return to this issue below, when we discuss weak identification.

The literature on estimation of dynamic stochastic general equilibrium models is vast and, because it quickly gets into specialized computational devices, it can be difficult to penetrate. For example, models of the Smets–Wouters sort rely on log-linearized approximations to decision rules, which both makes the models fairly easy to solve and means that the Kalman filter can be used to compute the Gaussian likelihood. Much of the recent methodological research on estimation of these models has focused on avoiding the log-linearization step. Among other things,

Figure 2
Prior and Posterior Distributions for Two Structural Parameters in a Dynamic Stochastic General Equilibrium Model



Source: Smets-Wouters (2003), Figure 1c (upper panel).
Note: This figures represents the breakthroughs made over the past 20 years in the estimation of dynamic stochastic general equilibrium models. In the model of Smets-Wouters (2003), the "Calvo wage" parameter in the first panel is the probability that a worker's wage does not change, and the "Calvo price" parameter in the second panel is the probability that the firm's price does not change.

avoiding log-linearization can improve the ability to analyze the effects of risk and uncertainty. However, there are substantial computational challenges in estimating nonlinear models, so that log-linearization remains common in practice. Canova (2007) provides an accessible textbook treatment of the linearize/Kalman filter/Bayes approach. Herbst and Schorfheide (2015) provide an up-to-date textbook treatment that focuses on computationally efficient methods for evaluating the posterior of linearized models. Fernández-Villaverde, Rubio-Ramírez, and Schorfheide (2016) provide a detailed overview of methods that avoid linearization.

Dynamic Factor Models and "Big Data"

The idea of using a large number of series to understand macroeconomic fluctuations is an old one, dating back at least as far as the economic indexes and forecasts of the Harvard Economic Service in the 1920s (Friedman 2009) and to Burns and Mitchell's (1946) use of 1,277 time series to study business cycles. The challenge of using large numbers of series is the proliferation of parameters in standard time series models. While there were large macroeconomic models developed in the 1960s, and versions of them remain in use today, the restrictions that reduced the number of parameters in those models were heavily criticized as being arbitrary, having neither statistical nor economic foundations. Although low-dimensional vector autoregressions had become a standard macroeconometric tool by the mid-1990s, an outstanding challenge was increasing the number of variables, both to improve forecasting and to span a wider range of forecast errors, and thus structural shocks. The technical challenge was that in an unrestricted vector autoregression, the number of parameters increases with the square of the number of variables. Methods were needed to manage this proliferation of parameters if time series methods were to be used with large numbers of variables.

Dynamic factor models impose parametric restrictions in a way that is consistent with empirical evidence and a broad set of modern theoretical models. In a dynamic factor model, a given observable variable—say, the growth rate of consumption of services—is written as the sum of a common component and an idiosyncratic component. The common component depends on unobserved (or latent) common variables, called factors, which evolve over time; the idiosyncratic component is uncorrelated with the common component and has limited correlation with the other idiosyncratic components. The idiosyncratic component captures measurement error and series-specific disturbances that have no broader macroeconomic consequences. Thus, in a dynamic factor model, a small number of unobserved factors explain the comovements of a large number of macroeconomic variables.

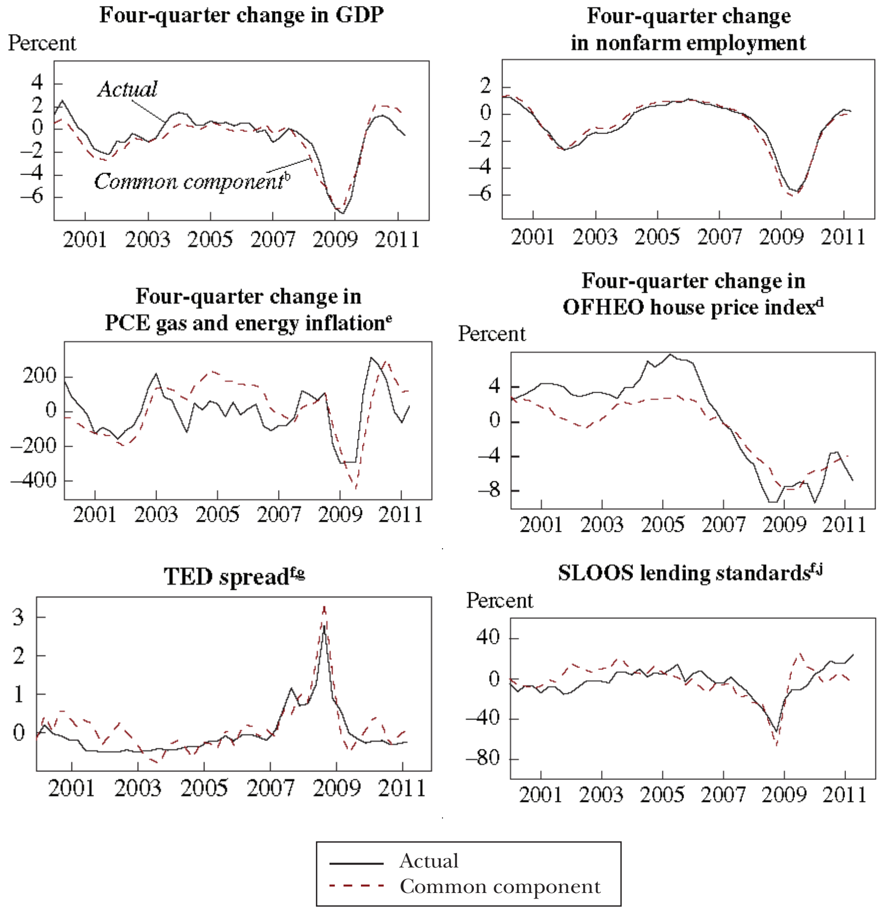
This brings us to our next figure, which is taken from Stock and Watson (2012a). Figure 3 shows the predicted value of six US quarterly macro variables from a 200-variable, six-factor dynamic factor model; this predicted value is called the "common component" of the series. The in-sample R^2 of the common component for four-quarter growth in GDP (that is, the R^2 of the regression of the four-quarter growth in GDP on the four-quarter growth of the six factors) is 73 percent; the average R^2 of the common component over 21 major expenditures variables from the national income and product accounts is 56 percent; and the average R^2 for all 200 variables is 46 percent. The parameters in this dynamic factor model were fitted using data from 1959–2007, so the post-2007 values of the common component represent the pseudo out-of-sample fit. At the visual level, for these and many other series, the fit is essentially the same in-sample and out-of-sample, suggesting that the parameters of the dynamic factor model remained largely stable during and after the financial crisis.

As Figure 3 illustrates, dynamic factor models fit the data. Techniques for dynamic factor analysis now can handle arbitrarily many series. One convenient way to estimate

Figure 3

Selected US macroeconomic Time Series: Actual Values and Common Components

(where the common components are the fitted values using the factors from a 200-variable, 6-factor dynamic factor model fit using data from 1959–2007)



Four-quarter change in GDP

Four-quarter change in nonfarm employment

Four-quarter change in PCE gas and energy inflation[e]

Four-quarter change in OFHEO house price index[d]

TED spread[f,g]

SLOOS lending standards[f,j]

Source: Stock-Watson (2012a), Figure 2.

Note: Figure 3 shows the predicted value of six US quarterly macro variables from a 200-variable, 6-factor dynamic factor model; this predicted value is called the "common component" of the series. The parameters in this dynamic factor model were fitted using data from 1959–2007, so the post-2007 values of the common component represent the pseudo out-of-sample fit.

the factors is principal components analysis, in which the factors are estimated by least squares. When estimated using many series, the principal component factor estimates can be treated as data for subsequent regressions (Stock and Watson 2002; Bai 2003; Bai and Ng 2006). To implement this approach, one needs to decide how many factors to use, and Bai and Ng (2002) show how to use information criteria to estimate the number of factors. This approach can be expanded to arbitrarily many series without substantially increasing the computational burden, indeed these models provide a twist on the usual "curse of dimensionality:" in dynamic factor models, the precision of the estimation of the factors improves as the number of data series increases, so that the curse becomes a blessing.

Because of theoretical and empirical work over the past 20 years, dynamic factor models have become a leading method for the joint modeling of large numbers—hundreds—of economic time series. Dynamic factor models have natural applications to macroeconomic monitoring and forecasting, a topic we take up below. They also can be used to estimate the effect of a structural shock, such as a monetary policy shock, on multiple economic variables. These economy-wide shocks drive the common factors, and because the factors can be estimated, the economic shocks can be estimated up to a nonsingular linear transformation. As a result, the techniques for shock analysis developed for structural vector autoregressions, including the new methods discussed above, carry over directly to dynamic factor models. By using many variables, dynamic factor models can more plausibly capture macro-structural shocks than can low-dimensional vector autoregressions. Moreover, the estimated structural impulse response functions are internally consistent across all the variables. In Stock and Watson (2016), we survey dynamic factor models, with a focus on structural shock analysis.[1]

Dynamic factor models are not the only method available for high-dimensional modeling. A different approach is to use a Bayesian prior distribution over the vector autoregression parameters to reduce the influence of the data on any one parameter estimate and thus to reduce the amount of noise across parameter estimates. In some applications, large numbers of restrictions arise naturally: for example, global vector autoregression reduces the dimensionality of the vector autoreregression parameter space by restricting domestic variables to depend on foreign variables only through a small number of weighted averages of global variables (Chudik and Pesaran 2016).

While this discussion has focused on the development of econometric methods for analyzing high-dimensional time series models, the other major development that has facilitated this work is the ready availability of data. The Federal Reserve Bank of St. Louis's FRED database, which migrated to an online platform in 1995, has been a boon to researchers and to the general public alike. A recent useful addition to FRED is FRED-MD, a monthly dataset currently comprised of 128 major economic time series for use in high-dimensional macroeconomic modeling (McCracken and Ng 2016); a beta-version with quarterly data (FRED-QD) is now available too. These datasets provide a common testbed for high-dimensional time series modeling and relieve researchers from the arduous task of updating a large dataset in response to new and revised data. A more specialized database, maintained by the Federal Reserve Bank of Philadelphia, archives and organizes

[1] A variant of a dynamic factor model is the factor-augmented vector autoregression (Bernanke, Boivin, and Eliasz 2005), in which one or more of the factors are modeled as observed. For example, because the Federal Reserve controls the federal funds interest rate, Bernanke, Boivin, and Eliasz (2005) argue that the target interest rate is itself a macroeconomic factor. Alternatively, factor-augmented vector autoregression can be interpreted as augmenting a low-dimensional vector autoregression with information from a first-step dynamic factor model. See Stock and Watson (2016) for a discussion of the relation between dynamic factor models and factor-augmented vector autoregressions.

real-time economic data; these data are especially valuable to those who want to test tools for real-time monitoring and forecasting.

Macroeconomic Monitoring and Forecasting

Two important related functions of macroeconomists in business and government are tracking the state of the economy and predicting where the economy is headed. During the 1960s and 1970s, these two functions—macroeconomic monitoring and macroeconomic forecasting—relied heavily on expert judgment. The 1980s and 1990s saw new efforts by time series econometricians to place macroeconomic monitoring and forecasting on a more scientific footing: that is, to be replicable, to use methods that are transparent and have well-understood properties, to quantify uncertainty, and to evaluate performance using out-of-sample experience. While these advances provided macroeconomic monitoring and forecasting with a solid foundation, much work remained to be done. This work included improving methods for quantifying and conveying forecast uncertainty; dramatically expanding the number of data series that could be used, both to enable real-time monitoring to use the most recently released information and to improve forecasts; and developing reliable forecasting tools that take into account the evolution of the economy. Here, we discuss the first two of these: forecast uncertainty and macroeconomic monitoring. Issues of model instability go far beyond macroeconomic monitoring and forecasting, so we defer that discussion to the next section.

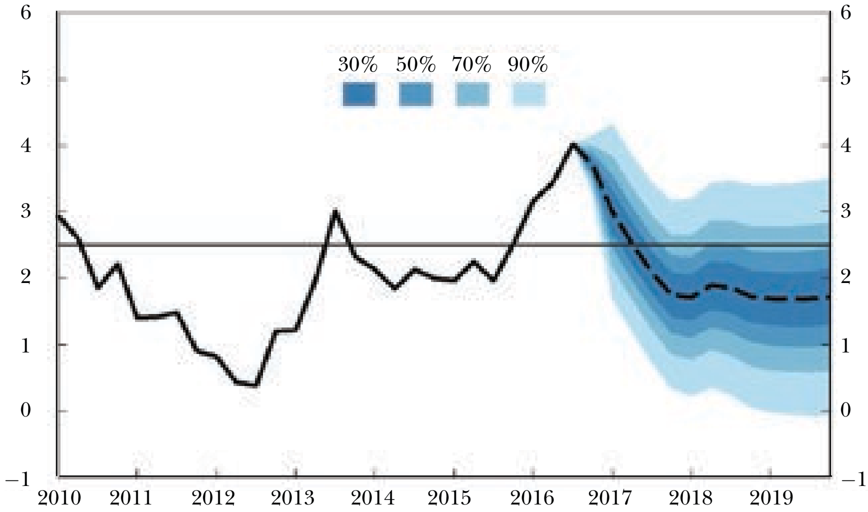
Estimating and Conveying Forecast Uncertainty

A fundamental problem of economic forecasting is that many economic variables are inherently very difficult to forecast, and despite advances in data availability, theory, and computational power, we have not seen dramatic improvements in forecast accuracy over the past decades. One implication of this observation is that economic forecasters should focus on communicating not just point estimates, but likely future ranges or distributions of the variable.

Our next figure highlights the development and adoption of density forecasts over the past 20 years. Figure 4 is a real-time release of a so-called fan chart from the Bank of Norway's Monetary Policy Report for December 2016. A fan chart communicates uncertainty by providing a density that describes the distribution of possible future values of the series being forecast, in this case Norwegian consumer price inflation.[2] The Bank of England was an early leader in the use of density forecasts and fan charts to communicate uncertainty to the public, and these methods are now widely adopted. Methods for constructing density forecasts are reviewed in

[2] The forecast uncertainty is better communicated in color! See the real thing at the websites of the Bank of England Inflation Report (http://www.bankofengland.co.uk/publications/Pages/inflationreport) and the Norges Bank Monetary Policy Report (http://www.norges-bank.no/en/Published/Publications/Monetary-Policy-Report-with-financial-stability-assessment/).

Figure 4
Fan Chart (Density Forecast) for Consumer Price Index (CPI) inflation in Norway
(percent; four-quarter change)



Source: Reproduced from Chart 2.2c of Norges Bank Monetary Policy Report for December 2016, which used data from Statistics Norway and Norges Bank.
Notes: This chart shows the distribution of possible future values of Norwegian consumer price inflation, projections for 2016 Q4 through 2019 Q4.

Elliott and Timmermann (2016, ch. 13), and Corradi and Swanson (2006) survey methods for evaluating the accuracy of density forecasts.

Beyond the clear communication of uncertainty, the past 20 years of academic work on forecasting has focused on extending the scientific foundations for forecasting. These include methods for evaluating forecasts (including density forecasts), selecting variables for forecasting, and detecting forecast breakdown. While judgment will inevitably play a role in interpreting model-based forecasts, a central goal of this research program is to reduce the amount of judgment involved in constructing a forecast by developing reliable models and tools for evaluating those models. For a graduate textbook treatment, see Elliott and Timmerman (2016), and for additional detail see Elliott and Timmermann (2013).

Macroeconomic Monitoring

Twenty years ago, economists who monitored the economy in real time used indexes of economic indicators and regression models for updating expectations of individual releases (such as the monthly employment report), combined with a large dose of judgment based on a narrative of where the economy was headed. While this approach uses data, it is not scientific in the sense of being replicable, using well-understood methods, quantifying uncertainty, or being amenable to later evaluation. Moreover, this method runs the risk of putting too much weight on the most recent but noisy data releases, putting too little weight on other data, and being internally inconsistent because each series is handled separately. Because knowing the current state of the economy in real-time is an ongoing, arguably increasingly important responsibility of policymakers, time series econometricians at central banks and in academia have put considerable effort into improving the foundations and reliability of real-time macroeconomic monitoring.

Our next figure illustrates a central line of research in macroeconomic monitoring: the use of large models, in particular dynamic factor models, to incorporate real-time data releases to provide an internally consistent framework for estimating current economic conditions. Figure 5 is taken from the February 10, 2017, weekly update published by the New York Federal Reserve Bank. The dynamic factor model used by the New York Fed incorporates the most recently available data on 36 major economic indicators to provide a weekly estimate of the growth of GDP in the current quarter. Figure 5 shows the evolution of this real-time forecast of current-quarter GDP growth—for obvious reasons, called a "nowcast" of GDP—for the fourth quarter of 2016.

In August, the prognosis was for growth slightly above 2 percent at an annual rate, but by the first Friday in the fourth quarter (October 7), the nowcast had fallen to 1.3 percent. The November 18 nowcast rose to 2.4 percent on the strength of retail sales and housing starts data released that week. Then weak industrial production data, along with weak housing data released less than two hours before the December 16 update, pushed that nowcast down to 1.8 percent. As it happened, the advance estimate of fourth-quarter GDP growth released January 27 was 1.9 percent, slightly less than the estimate of 2.1 percent made on January 20.

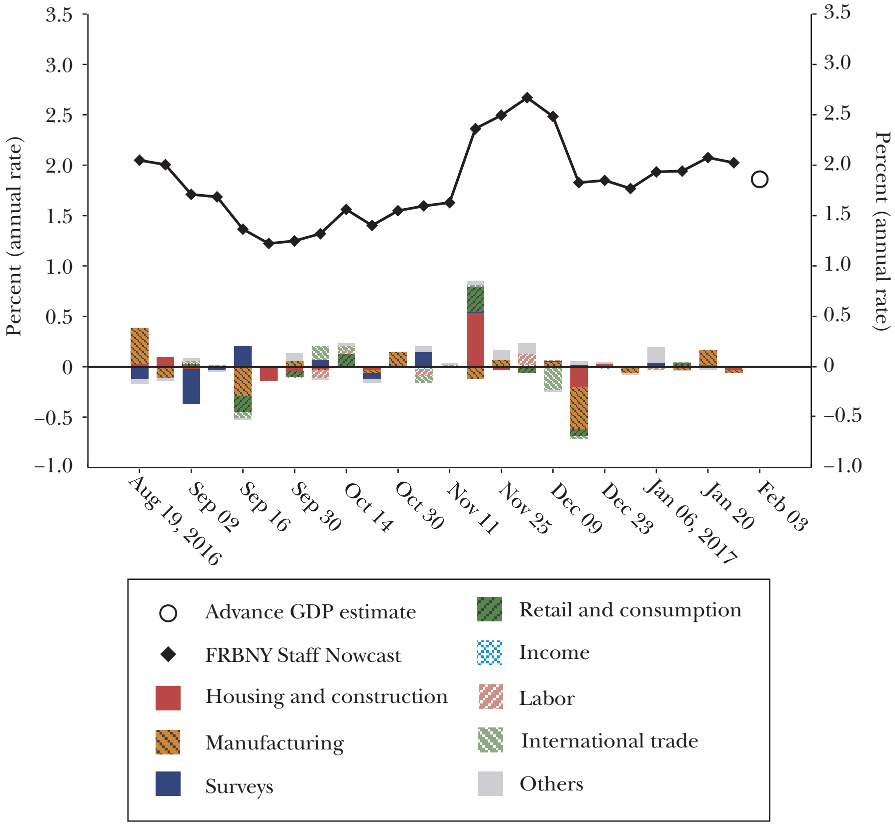
Under the hood of this real-time tracking product is a powerful set of tools for updating estimated factors in dynamic factor models using real-time data flows. The use of dynamic factor models for real-time macroeconomic monitoring incorporating staggered data releases dates to the NBER experimental coincident index (Stock and Watson 1989). By today's standards, that index was primitive: a monthly release that encompassed only four variables. The current suite of tools for handling large series and complicated data flows are exposited in detail in Bańbura, Giannone, Modugno, and Reichlin (2013). The New York Fed's model is updated (using the Kalman filter) as new data arrives, yielding an updated estimate of the single latent factor which in turn provides an updated estimate of the current-quarter value of GDP growth. By using a single flexible model, the news content of each series is exploited in a disciplined and internally consistent way. Some announcements contain substantial news, but many do not, and using a single model to evaluate these releases—rather than a suite of small models or judgment—provides a scientific way to use the real-time data flow.

The New York Fed report is one of several that use dynamic factor models to provide real-time, publicly available reports on the state of the economy. The EUROCOIN index, maintained by the Centre for Economic Policy Research and the Bank of Italy, is a real-time monthly index computed using a dynamic factor model with approximately 145 variables, calibrated to estimate monthly eurozone GDP growth (Altissimo, Cristadoro, Forni, Lippi, and Veronese 2010). The Chicago Fed National Activity Index is a monthly index of real economic activity

Figure 5

Contributions of Daily Data Releases to the Federal Reserve Bank of New York Real-Time Nowcast of 2016Q4 GDP Growth

(bars represent weekly contributions of data revisions to changes in the nowcast)



Source: Federal Reserve Bank of New York Nowcasting report, February 10, 2017.
Note: Figure 5 shows the evolution of a real-time forecast of 2016 fourth-quarter GDP—for obvious reasons, sometimes called a "nowcast." Technically, the points through September 31, 2016, are forecasts of fourth quarter GDP growth; the points October 1 through December 31, 2016, are nowcasts; and the points January 1, 2017, to the end of the series are backcasts of fourth quarter GDP growth.

constructed as the single factor in an 85-variable dynamic factor model. The Federal Reserve Bank of Philadelphia maintains the Aruoba-Diebold-Scotti (2009) index, which is updated weekly using a six-variable dynamic factor model with one quarterly series (GDP), four monthly series, and one weekly series. The Federal Reserve Bank of Atlanta's real-time nowcasting tool, GDPNow, uses a dynamic factor model combined with a GDP accounting approach to estimate current-quarter GDP.

There are other methods for nowcasting and mixed-frequency data. One popular tool for single-equation prediction using mixed-frequency data is the MIDAS model (Ghysels, Sinko, and Valkanov 2007), in which high-frequency data are temporally aggregated using data-dependent weights. For a survey of methods of mixed-frequency nowcasting and forecasting, see Foroni and Marcellino (2013).

Model Instability and Latent Variables

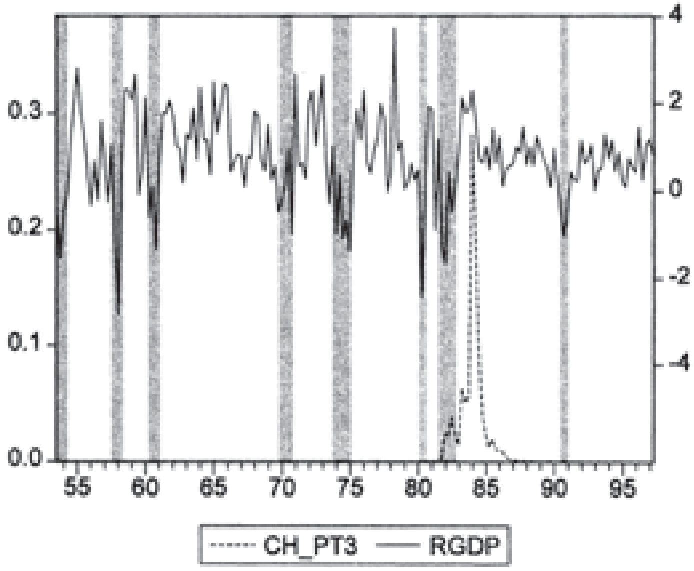
A large empirical literature has documented instability in both large- and small-dimensional time series models. A particularly well-known example of this instability is the Great Moderation, the period from 1984 to 2007 in which the volatility of many macroeconomic time series was greatly reduced. Examples of some of the many papers that document instability in the parameters of time series models include Stock and Watson (1996) for univariate time series forecasts, Stock and Watson (2003) for inflation forecasts using asset prices, and Welch and Goyal (2008) for equity premium forecasts. The methods in this literature draw in part on tests for breaks, time variation, and out-of-sample stability that date to the early 1990s.

This widespread nature of instability in time series relations raises the question of how to modify time series models so that they can be useful even in the presence of instability. An early approach was to model instability as deterministic regime shifts, but while useful, that approach is often unsatisfying because, outside of applications to a policy regime shift, the single-break model is an approximation and in any event there is rarely a reason to think that another shift will not occur. After all, the Great Moderation was followed by the financial crisis. A more appealing modeling strategy is to allow model parameters to evolve over time according to a stochastic process. If those time-varying parameters multiply observed variables, then the model has a linear state space (hidden Markov) structure and the Gaussian likelihood can be computed (using the Kalman filter). If, however, the time-varying parameters multiply latent variables, then it has an inherently nonlinear structure. Estimating such models is challenging, and it was clear 20 years ago that the rudimentary methods available needed to be improved.

The next two figures illustrate developments in the estimation of nonlinear latent variable models over the past 20 years. The first, Figure 6, is from Kim and Nelson (1999); the figure shows real GDP growth (the solid line), and the posterior probability of a break in the variance in GDP (dashed line). Based on this figure, Kim and Nelson (1999) concluded that US GDP growth had entered a period of low volatility, and that the most likely date for this transition was 1984Q1. This conclusion was reached independently using break test methods by McConnell and Perez-Quiros (2000). This low-volatility period, which lasted through 2007 (and to which the economy seems to have returned) subsequently became known as the Great Moderation.

Aside from its seminal empirical finding, Figure 6 illustrates a major methodological development in handling nonlinear and/or non-Gaussian time series models with latent variables. Kim and Nelson's (1999) model falls in this category: it allows for a one-time shift in the mean and variance of GDP growth, layered on top of Hamilton's (1989) stochastic regime shift model with recurrent shifts in the mean (which, Hamilton found, aligned with business cycle turning points). A challenge in these models is estimating the time path of the latent variable given all the data,

Figure 6
US GDP Growth and the Posterior Probability of a Regime Change in its Innovation Variance



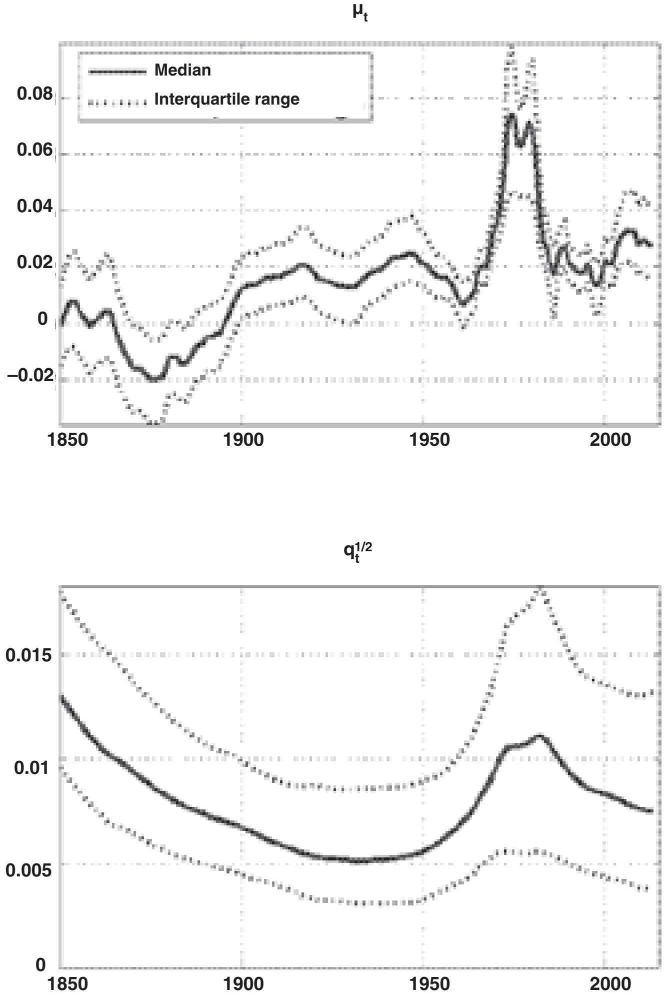
Source: Kim and Nelson (1999), Fig. 3.A.
Note: The figure shows real GDP growth (the solid line), and the posterior probability of a break in the variance in GDP (dashed line). Based on this figure, Kim and Nelson (1999) concluded that US GDP growth had entered a period of low volatility, and that the most likely date for this transition was 1984Q1.

the so-called smoothing problem, along with the model parameters. To estimate their parameters and to solve this smoothing problem—to produce Figure 6—Kim and Nelson used Markov Chain Monte Carlo methods, which break down their complicated nonlinear non-Gaussian model into a sequence of Monte Carlo simulations using simpler models. Over the past 20 years, Markov Chain Monte Carlo has become a widely used tool for estimating seemingly intractable nonlinear/non-Gaussian models. With this tool, Kim and Nelson were able to obtain the posterior distribution of a one-time structural break in the variance which, as Figure 6 shows, strongly points to a reduction in the variance of GDP growth early in 1984.

The next figure, Figure 7, shows two panels from Cogley and Sargent (2015) that illustrate the incorporation of stochastic volatility into latent state variables. Cogley and Sargent use a univariate model that decomposes the rate of inflation into unobserved permanent and transitory (measurement error) components, both of which have innovations with time-varying variances. These variances are modeled as latent stochastic volatility processes. From a technical perspective, the situation is similar to that faced by Kim and Nelson (1999) in that the resulting model expresses the observed data as a nonlinear function of unobserved random variables (the permanent and transitory components of inflation and their volatilities). While the details differ, the Cogley–Sargent model is also readily estimated by Markov Chain Monte Carlo methods.

Figure 7

Trend Inflation (Upper Panel) and the Standard Deviation of the Trend Innovation (Lower Panel) in an Unobserved Components–Stochastic Volatility Model of US Inflation, 1850–2012



Source: Cogley-Sargent (2015), Fig. 7(A, C).

Note: Figure 7 illustrates the incorporation of stochastic volatility into latent state variables. Cogley and Sargent (2015) use an unobserved-components/stochastic-volatility model to study the evolution of the US inflation process from 1850 to 2012. Their posterior estimate of trend inflation is shown in the first panel, and their estimate of the time-varying standard deviation of changes in the trend is shown in the second panel. They find the periods of greatest variance in the trend to be during the Civil War and during the period of inflation and disinflation in the 1970s and early 1980s.

Cogley and Sargent (2015) use this unobserved-components/stochastic-volatility model to study the evolution of the US inflation process from 1850 to 2012. Their posterior estimate of trend inflation is shown in the first panel, and their estimate of the time-varying standard deviation of changes in the trend is shown in the second panel. They find the periods of greatest variance in the trend to be during the Civil War and during the period of inflation and disinflation in the 1970s and early 1980s.

The literature on nonlinear/non-Gaussian filtering is complex, nuanced, and massive. See Durbin and Koopman (2012) for a textbook treatment of linear and nonlinear filtering methods.

More Reliable Inference

Finally, the past 20 years has seen important work that aims to improve the quality of statistical inferences. In the mid-1990s, several influential studies found that widely used methods for computing test statistics with time series data could reject far too often or, said differently, that confidence intervals could fail to include the true parameter value far less frequently than the claimed 95 percent coverage rate. Theoretical econometricians recognized that more work was needed, particularly in the areas of instrumental variables where the instrument might be weak, standard errors for regression with serially correlated errors, and regression with highly persistent regressors.

Weak Instruments and Weak Identification

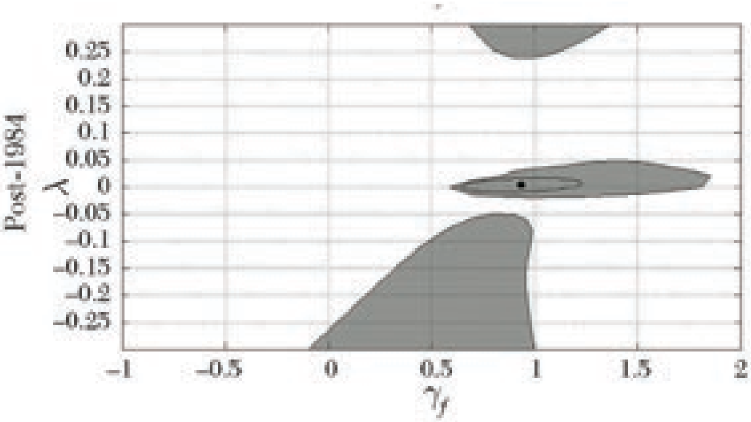
A weak instrument has a small correlation with the variable it is instrumenting, given the other included variables. For decades, conventional wisdom held that a weak instrument would simply produce large standard errors, which would correctly convey that the information in that variable is scant. But a series of papers in the 1990s showed that the consequences of a so-called weak instrument were more serious: the estimator will in general be biased, conventional standard errors are misleading, and these problems can occur in very large samples.[3] This problem, which is more generally referred to as weak identification, also arises in generalized method of moments estimation. Although weak instruments have received the most attention in microeconometrics, the inferential challenges posed by weak identification also have played a role in time series econometrics over the past 20 years.

The next figure, taken from Mavroeidis, Plagborg-Møller, and Stock (2014), illustrates the problems with using conventional asymptotic standard errors and confidence intervals in instrumental variables methods when one has weak instruments. Figure 8 shows confidence sets for two key parameters of the hybrid New Keynesian Phillips Curve; on the vertical axis, λ is the coefficient on marginal cost (or, in other specifications, the unemployment gap or output gap) and, on the

[3] Key papers on this subject from the 1990s include Nelson and Startz (1990a, 1990b) and Hansen, Heaton, and Yaron (1996) (Monte Carlo simulations), Bound, Jaeger, and Baker (1995) (empirical application), and Staiger and Stock (1997) (econometric theory).

Figure 8

Point Estimate and 90% Confidence Sets for Hybrid New Keynesian Phillips Curve Parameters: Standard Generalized Method of Moments (Ellipse) and Weak-Instrument Robust (Gray)



Source: Mavroeidis, Plagborg-Møller, and Stock (2014), Fig. 11a.
Note: Figure 8 shows confidence sets for two key parameters of the hybrid New Keynesian Phillips Curve. The dot is the point estimate using generalized method of moments, and the small ellipse around the point estimate is the corresponding nominal 90 percent confidence set computed using textbook asymptotics. The gray regions in the figure comprise a 90 percent confidence set that is robust to the use of weak instruments. The figures show that the weak-identification robust confidence sets differ dramatically from the standard asymptotic confidence ellipse. See text for details.

horizontal axis, γ_f is the coefficient on forward-looking rational expectations (sometimes interpreted as relating to the fraction of forward-looking agents). The results in this figure were computed using data from 1984–2011, where, following Galí and Gertler (1999), the labor share is the proxy for marginal cost, and the instruments are three lags each of marginal cost and the change in inflation, pruned down from Galí and Gertler's (1999) original set of 24 instruments (which yield similar qualitative results). The dot is the point estimate using generalized method of moments, and the small ellipse around the point estimate is the corresponding nominal 90 percent confidence set computed using textbook asymptotics. The gray regions in the figure comprise a 90 percent confidence set that is robust to the use of weak instruments. The obvious conclusion from Figure 8 is that the weak-identification robust confidence sets differ dramatically from the standard asymptotic confidence ellipse. Mavroeidis, Plagborg-Møller, and Stock (2014) argue that the reason for this divergence is that the instruments used in this generalized method of moments estimation are weak. This problem of weak identification arises broadly in New Keynesian Phillips Curve applications (for example, Henry and Pagan 2004; Mavroeidis 2004; Nason and Smith 2008).

Weak identification also arises in other contexts, like in the estimation of intertemporal consumption-based asset pricing models (Stock and Wright 2000) and estimation of monetary policy reaction functions using generalized method of moments (Consolo and Favero 2009). Weak identification arises in some types of inference in structural autoregressions (for example, Pagan and Robertson 1998; Chevillon, Mavroeidis, and Zhan 2016; for more references, see Stock and Watson 2016, Section 4). It also arises in complicated ways in the estimation of dynamic stochastic equilibrium models (for example, Andrews and Mikusheva 2015; Qu 2014).

In linear instrumental variable regressions, one commonly used diagnostic is to check if the F-statistic testing the hypothesis that the coefficient(s) on the instrument(s) in the first stage of two stage least squares—the so-called first-stage F-statistic—is less than 10; if so, weak identification is potentially a problem. This specific approach is specialized to the homoskedastic setting with uncorrelated errors; approaches to extending this to heteroskedasticity are proposed by Montiel Olea and Pflueger (2013) and Andrews (2016).

In the simplest models—the textbook regression model with a single endogenous regressor and errors that are homoskedastic and serially uncorrelated—there are now methods for dealing with weak instruments with very good size and power, both asymptotically and in finite samples. As one departs from this model, most notably when the number of parameters gets large and/or the model is nonlinear in the parameters, the toolkit is less complete and theoretical work remains under way.

Inference with Serially Correlated and Potentially Heteroskedastic Errors

In time series data with a serially correlated error term, each additional observation does not provide entirely new information about the regression coefficient. Moreover, many time series regressions exhibit clear signs of heteroskedasticity. In this setting, the ordinary least squares standard error formula does not apply and instead standard errors that are robust to heteroskedasticity and autocorrelation must be used. For example, this problem arises when the dependent variable is a multi-period return or a multiple-period-ahead variable. The problems of heteroskedasticity and autocorrelation also arise in generalized method of moments models when the data are serially correlated.

In practice, the most commonly used standard errors that are heteroskedasticity- and autocorrelation-robust are computed using methods from seminal papers by Newey and West (1987) and Andrews (1991). These methods compute standard errors by replacing the estimate of the variance of the product of the regressor and the error in the usual heteroskedasticity-robust formula for the variance of the ordinary least squares estimator with a weighted average of the autocovariances of that product; the number of autocovariances averaged is determined by the so-called "bandwidth" parameter. But even 20 years ago, there were inklings that the performance of hypothesis tests and confidence intervals constructed using these standard errors in typical macroeconometric applications fell short of the asymptotic performance used to justify the tests. In an early Monte Carlo simulation, den Haan and Levin (1997) studied the rejection rates of tests using these standard errors under the null hypothesis—that is, the size of the test. Depending on the persistence in the data, they found that a test that should reject 5 percent of the time under the null will in practice reject 10 or even 20 percent of the time. If the aim of a research project is, say, to test for predictability in multiyear stock returns using monthly data, this over-rejection could easily lead to an incorrect conclusion that returns are predictable when in fact they are not.

Understanding the source of these size distortions and improving upon Newey–West/Andrews standard errors therefore became a major line of research by theoretical econometricians over the past 20 years, which is succinctly surveyed by Müller (2014, Sections 2–3). In brief, this line of work finds that to construct tests with a rejection rate closer to the desired 5 percent, it is necessary to use bandwidths much larger than those suggested by Newey–West and Andrews. But doing so results in a complication: the test statistic no longer has the usual large-sample normal distribution and, in general, nonstandard critical values must be used. These ideas were set out by Kiefer, Vogelsang, and Bunzel (2000), and their insights prompted a large literature aimed at understanding and refining their large-bandwidth approach. This theoretical literature has now produced multiple methods that yield far smaller size distortions than tests based on Newey–West/Andrews standard errors, and which also have better power than the Kiefer–Vogelsang–Bunzel test. Moreover, some of these tests have standard critical values, simplifying their use in practice.

Applied econometricians typically are eager to use the most recent econometric method when they demonstrably improve upon the methods of the past. Curiously, this has not been the case for heteroskedasticity- and autocorrelation-robust inference, where empirical practice continues to be dominated by Newey–West/Andrews standard errors. The new methods are easy to use, straightforward to understand, and have a lineage that traces back 40 years. It is time for empirical researchers in time series econometrics to take the next step and to adopt these improved methods for heteroskedasticity- and autocorrelation-robust inference.

Long-run Relations, Cointegration, and Persistent Regressors

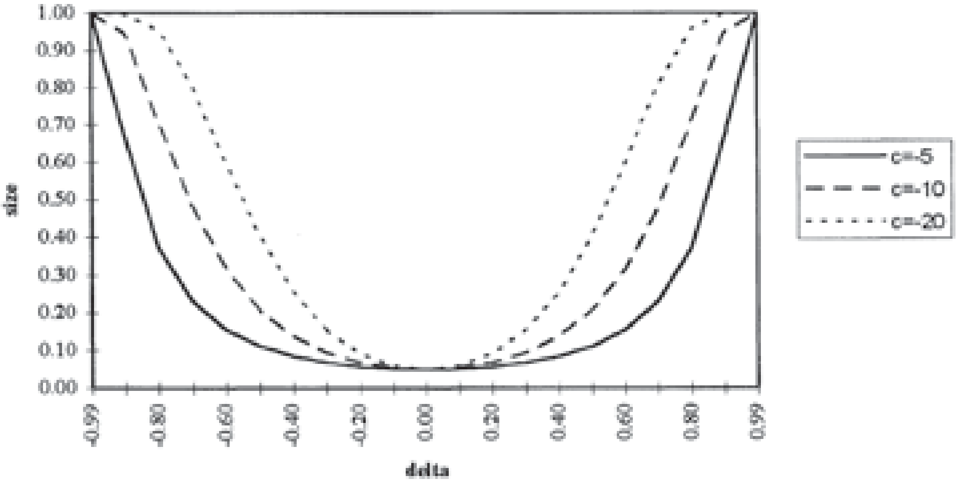
The basic insight of cointegration—the development for which Clive Granger received the Nobel Prize in 2003—is that multiple persistent macroeconomic variables move together at low frequencies, that is, they share common long-term trends. Moreover, these low-frequency comovements connect with basic economic theories such as balanced economic growth. But while there was a surge of work on cointegration in the 1980s and 1990s, such work has received less emphasis since then.

Our final historical figure, from Elliott (1998), illustrates a technical roadblock hit by this research program. Elliott's figure, our Figure 9, portrays the null rejection rate of a test of the value of a cointegrating coefficient in a simple model with two cointegrated variables. The test maintains that each of the variables is integrated of order one, that is, has a unit autoregressive root, an assumption that is part of the cointegration model. Figure 9 shows that small departures from this unit-root assumption (as measured by *c*, which is the difference between the true largest root and one, multiplied by the sample size) can cause major problems for tests and confidence intervals about the value of that cointegrating coefficient: tests that are supposed to reject 5 percent of the time under the null can reject with very high rates (shown on the vertical axis), particularly when the correlation δ (shown on the horizontal axis) between innovations in the error and in

Figure 9

Asymptotic Size of Tests of Values of the Cointegrating Coefficient Using Efficient Cointegrating Estimators and Their Standard Errors when the Time Series Follow Local-to-Unity Processes with Parameter *c*

(delta is the correlation between innovations in the error and in the regressor)



Source: Elliott (1998), Figure 1(a).
Note: The figures portrays the null rejection rate of a test of the value of a cointegrating coefficient in a simple model with two cointegrated variables. The test maintains that each of the variables is integrated of order one, that is, has a unit autoregressive root, an assumption that is part of the cointegration model. The figure shows that small departures from this unit-root assumption (as measured by c, which is the difference between the true largest root and one, multiplied by the sample size) can cause major problems for tests and confidence intervals about the value of that cointegrating coefficient.

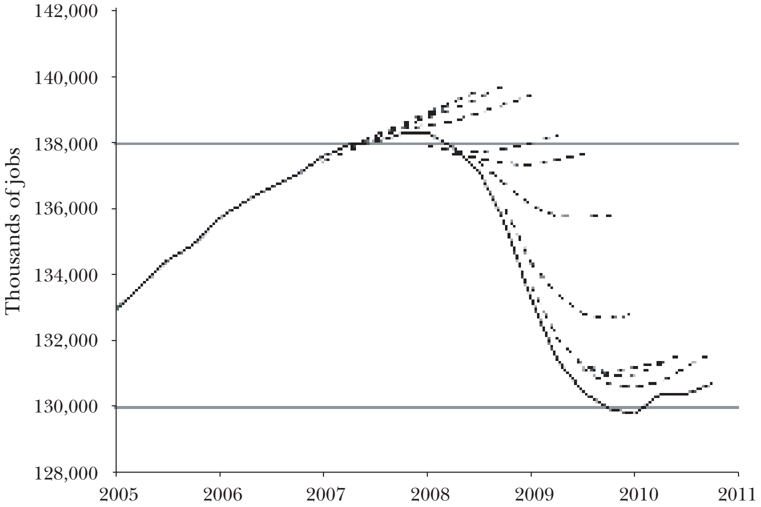
the regressor is large. In fact, this problem arises for deviations from a unit root that are too small to be detected with high probability, even in arbitrarily large samples. As a result, standard methods of inference developed for cointegration models are not robust to effectively undetectable departures from the model, making such inference unreliable.

While subsequent work has produced novel ideas by econometric theorists, the proposed methods have drawbacks and no alternative set of procedures have emerged. In fact, the literature has shown that the problem documented in Figure 9 goes beyond the local-to-unity model used by Elliott (1998) and other researchers in this area. Related problems of inference also arise in regressions in which a regressor is persistent, as can occur in applications with financial data.

It is important to stress that these challenges are technical ones; the basic insight of cointegration that variables move together at low frequencies is a deep one that connects with core economic theories such as balanced growth and the term structure of interest rates. But inference, and perhaps modeling, of those comovements can be more complicated than had originally been thought.

Figure 10

"The Mother of All Forecast Errors": Survey of Professional Forecasters Median Forecast for Nonfarm Business Employment during the 2007–2009 Recession and Early Recovery



Source: Philadelphia Fed Survey of Professional Forecasters.
Note: This figure shows the real-time median forecast of the log of nonfarm employment recorded by the Survey of Professional Forecasters in the quarters leading up to and through the financial crisis. Even well after the crisis began and real-time information about the collapse of the economy was available, these forecasters consistently predicted a mild recession.

Challenges Ahead

We close by mentioning a few of the research challenges for time series econometrics. Our final figure shows that despite the substantial improvements in forecasting methods over the past decades, much work remains. When we teach, we call Figure 10 the "Mother of All Forecast Errors." This figure shows the real-time median forecast of the log of nonfarm employment recorded by the Survey of Professional Forecasters in the quarters leading up to and through the financial crisis. Even well after the crisis began and real-time information about the collapse of the economy was available, these forecasters consistently predicted a mild recession. A small part of these errors is due to revisions between preliminary and final data, but most of these errors, we believe, represent a failure of forecasting models to capture the severity of the shocks and their devastating effect on the economy. Forecasters certainly were not the only economists to misjudge events leading up to and during the financial crisis! But this is an article about time series methods, and in our view, tackling the challenge of Figure 10 is a priority.

Another open challenge lies in the big data sphere. The methods of the past 20 years—dynamic factor models and large Bayesian vector autoregressions—have made it possible to include arbitrarily many series in forecasting systems and to incorporate data releases in real time, and the result has been large improvements in macroeconomic monitoring. However, there is some evidence that the parametric restrictions (or priors) that make these methods work discard potentially important information. In the context of dynamic factor models, the question is whether there is useful information in the higher factors beyond the handful that would normally be included (such as the six factors used to produce Figure 3). Some studies have looked at this question, with mixed results; for example, Carrasco and Rossi (2016) give some positive results, while we give some negative results in Stock and Watson (2012b). A more ambitious question is whether there is exploitable nonlinear structure in these data that could perhaps be revealed by modern machine learning methods. While it is tempting to dive in and use a battery of machine learning methods to attack these data, one must remember that data snooping can lead to unintentional overstatement of results. One advantage of dynamic factor models, after all, is that they are closely linked to dynamic macro models (Sargent 1989; Boivin and Giannoni 2006). We suspect that the next steps towards exploiting additional information in large datasets will need to use new statistical methods guided by economic theory.

Separately, there are important open questions relating to low-frequency time series econometrics. For example, what does historical evidence tell us about whether the recent slowdown in US productivity is permanent or temporary? The answer to this question is crucial for many long-term economic issues, such as the future of Social Security and valuing policies to mitigate climate change. Another, technically related set of questions returns to the basic insight of cointegration and the challenge posed by Elliott's (1989) figure (Figure 9): there are clearly low-frequency comovements in the data, and macroeconometricians need a set of tools for quantifying those comovements that does not hinge on adopting a particular model, such as a unit root model, for the underlying trends. These are technically difficult problems, and Müller and Watson (2016a, 2016b) propose possible avenues for tackling them.

Finally, there are a number of opportunities for expanding identification and estimation of macro models by using information in microeconometric data. Here, opportunities range from estimation of parameters describing individual preferences and firm behavior, to the possibility of using rich micro data to improve macro monitoring and forecasting.

The earliest empirical work in macroeconomics relied on time series data; indeed the first instrumental variables regression was estimated in 1926 using time series data. The past 20 years has seen a continuation of the vigorous development of methods for using time series data. These methods draw on improved computational capacity, better data availability, and new understandings in econometric and statistical theory. The core driver of these developments is the need of policymakers for reliable guidance on the effects of contemplated policies, along with their shared need with the private sector to understand where the economy is and where it is going. Those needs will not go away. If anything, they become more urgent in our volatile and ever-changing economic environment. Although the challenges facing time series econometricians are difficult, so have they been in the past, and exciting and highly relevant research programs beckon.

■ *We thank Gray Calhoun, Mikkel Plagborg-Møller, and the authors of the papers from which we took figures, for helpful comments.*

References

Aït-Sahalia, Yacine, and Lars Peter Hansen. 2010. *Handbook of Financial Economics*, Vol. 1: *Tools and Techniques* and Vol. 2: *Applications*. Elsevier.

Altissimo, Filippo, Riccardo Cristadoro, Mario Forni, Marco Lippi, and Giovanni Veronese. 2010. "New EuroCOIN: Tracking Economic Growth in Real Time." *Review of Economics and Statistics* 92(4): 1024–34.

Andrews, Donald W. K. 1991. "Heteroskedasticity and Autocorrelation Consistent Covariance Matrix Estimation." *Econometrica* 59(3): 817–858.

Andrews, Isaiah. 2016. "Valid Two-Step Identification-Robust Confidence Sets for GMM." http://economics.mit.edu/files/11848.

Andrews, Isaiah, and Anna Mikusheva. 2015. "Maximum Likelihood Inference in Weakly Identified Dynamic Stochastic General Equilibrium Models." *Quantitative Economics* 6(1): 123–52.

Aruoba, S. Borağan, Francis X. Diebold, and Chiara Scotti. 2009. "Real-Time Measurement of Business Conditions." *Journal of Business & Economic Statistics* 27(4): 417–27.

Bai, Jushan. 2003. "Inferential Theory for Factor Models of Large Dimensions." *Econometrica* 71(1): 135–72.

Bai, Jushan, and Serena Ng. 2002. "Determining the Number of Factors in Approximate Factor Models." *Econometrica* 70(1): 191–21.

Bai, Jushan, and Serena Ng. 2006. "Confidence Intervals for Diffusion Index Forecasts and Inference for Factor-Augmented Regressions." *Econometrica* 74(4): 1133–50.

Bank of Norway. 2016. "Monetary Policy Report with Financial Stability Assessment 4/16." December 15, 2016. http://www.norges-bank.no/en/Published/Publications/Monetary-Policy-Report-with-financial-stability-assessment/416-Monetary-Policy-Report/.

Bańbura, Marta, Domenico Giannone, Michele Modugno, and Lucrezia Reichlin. 2013. "Now-Casting and the Real-Time Data Flow." Chap. 4 in *Handbook of Economic Forecasting*, vol. 2, edited by Graham Elliott, and Alan Timmermann. Elsevier, North-Holland.

Baumeister, Christiane, and James D. Hamilton. 2015. "Sign Restrictions, Structural Vector Autoregressions, and Useful Prior Information." *Econometrica* 83(5): 1963–99.

Bernanke, Ben S., Jean Boivin, and Piotr Eliasz. 2005. "Measuring the Effects of Monetary Policy: A Factor-Augmented Vector Autoregressive (FAVAR) Approach." *Quarterly Journal of Economics* 120(1): 387–422.

Bernanke, Ben S., and Kenneth N. Kuttner. 2005. "What Explains the Stock Market's Reaction to Federal Reserve Policy?" *Journal of Finance* 60(3): 1221–57.

Boivin, Jean, and Marc Giannoni. 2006. "DSGE Models in a Data-Rich Environment." NBER Working Paper 12772.

Bound, John, David A. Jaeger, and Regina M. Baker. 1995. "Problems with Instrumental Variables Estimation When the Correlation between the Instruments and the Endogenous Explanatory Variable Is Weak." *Journal of the American Statistical Association* 90: 443–50.

Burns, Arthur F., and Wesley C. Mitchell. 1946. *Measuring Business Cycles*. NBER.

Canova, Fabio. 2007. *Methods for Applied Macroeconomic Research*. Princeton University Press.

Chevillon, Guillaume, Sophocles Mavroeidis, and Zhaoguo Zhan. 2016. "Robust Inference in Structural VARs with Long-Run Restrictions." https://sites.google.com/site/sophoclesmavroeidis/research/lrsvar.pdf?attredirects=0.

Chudick, Alexander, and M. Hashem Pesaran. 2016. "Theory and Practice of GVAR Modelling." *Journal of Economic Surveys* 30(1): 165–97.

Cochrane, John H., and Monica Piazzesi. 2002. "The Fed and Interest Rates: A High-Frequency Identification." *American Economic Review* 92(2): 90–95.

Cogley, Timothy, and Thomas J. Sargent. 2015. "Measuring Price-Level Uncertainty and Instability in the United States, 1850–2012." *Review of Economics and Statistics* 97(4): 827–38.

Consolo, Agostino, and Carlo A. Favero. 2009. "Monetary Policy Inertia: More a Fiction than a Fact?" *Journal of Monetary Economics* 56(6): 900–906.

Corradi, Valentina, and Norman R. Swanson. 2006. "Predictive Density Evaluation." Chap. 5 in *Handbook of Economic Forecasting,* vol. 1, edited by Graham Elliott, Clive W. J. Granger, and Allan Timmermann. Elsevier: North Holland.

Corrasco, Marine, and Barbara Rossi. 2016. "In-Sample Inference and Forecasting in Misspecified Factor Models." *Journal of Business and Economic Statistics* 34(3): 313–38.

DeJong, David N., Beth F. Ingram, and Charles H. Whiteman. 2000. "A Bayesian Approach to Dynamic Macroeconomics." *Journal of Econometrics* 98(2): 203–223.

den Haan, W. J., and A. Levin. 1997. "A Practitioners Guide to Robust Covariance Matrix Estimation." Chap. 12 of *Handbook of Statistics,* Vol. 15: *Robust Inference,* edited by G. S. Maddala and C. R. Rao. Elsevier.

Durbin, J., and S. J. Koopman. 2012. *Time Series Analysis by State Space Methods,* 2nd edition. Oxford University Press.

Elliott, Graham. 1998. "On the Robustness of Cointegration Methods When Regressors Almost Have Unit Roots." *Econometrica* 66(1): 149–58.

Elliott, Graham, and Alan Timmermann, eds. 2013. *Handbook of Economic Forecasting,* vol. 2. Elsevier.

Elliott, Graham, and Alan Timmermann. 2016. *Economic Forecasting*. Princeton University Press.

Faust, Jon, John H. Rogers, Eric Swanson, and Jonathan H. Wright. 2003. "Identifying the Effects of Monetary Policy Shocks on Exchange Rates Using High Frequency Data." *Journal of the European Economic Association* 1(5): 1031–57.

Federal Reserve Bank of New York. 2017. "Nowcasting Report, February 10, 2017." Available at: https://www.newyorkfed.org/research/policy/nowcast.

Fernández-Villaverde, Jesús, Juan F. Rubio-Ramírez, and Frank Schorfheide. 2016. "Solution and Estimation Methods for DSGE Models." Chap. 9 in *Handbook of Macroeconomics*, vol. 2, edited by John B. Taylor and Harald Uhlig. Elsevier.

Foroni, Claudia, and Massimiliano Marcellino. 2013. "A Survey of Econometric Methods for Mixed-Frequency Data." Norges Bank Research Working Paper 2013-6.

Friedman, Walter F. 2009. "The Harvard Economic Service and the Problems of Forecasting." *History of Political Economy* 41(1): 57–88.

Fry, Renée, and Adrian Pagan. 2011. "Sign Restrictions in Structural Vector Autoregressions: A Critical Review." *Journal of Economic Literature* 49(4): 938–60.

Galì, Jordi, and Mark Gertler. 1999. "Inflation Dynamics: A Structural Econometric Analysis." *Journal of Moneteary Economics* 44(2): 195–222.

Gertler, Mark, and Peter Karadi. 2015. "Monetary Policy Surprises, Credit Costs, and Economic Activity." *American Economic Journal: Macroeconomics* 7(1): 44–76.

Ghysels, Eric, Arthur Sinko, and Rossen Valkanov. 2007. "MIDAS Regressions: Further Results and New Directions." *Econometric Reviews* 26(1): 53–90.

Giacomini, Raffaella, and Toru Kitagawa. 2014. "Inference about Non-Identified SVARs." CEPR Discussion Paper 10287.

Hamilton, James D. 1989. "A New Approach to the Economic Analysis of Nonstationary Time Series and the Business Cycle." *Econometrica* 57(2): 357–84.

Hamilton, James D. 2003. "What Is an Oil Shock?" *Journal of Econometrics* 113(2): 363–98.

Hansen, Lars P., John Heaton, and Amir Yaron. 1996. "Finite Sample Properties of Some Alternative GMM Estimators." *Journal of Business and Economic Statistics* 14(3): 262–80.

Henry, S. G. B., and A. R. Pagan. 2004. "The Econometrics of the New Keynesian Policy Model: Introduction." *Oxford Bulletin of Economics and Statistics* 66(Supplement): 581–607.

Herbst, Edward P., and Frank Schorfheide. 2015. *Bayesian Estimation of DSGE Models.* Princeton University Press.

Ireland, Peter N. 1997. "A Small, Structural, Quarterly Model for Monetary Policy Evaluation." *Carnegie-Rochester Conference Series on Public Policy* 47: 83–108.

Kiefer, Nicholas M., Timothy J. Vogelsang, and Helle Bunzel. 2000. "Simple Robust Testing of Regression Hypotheses." *Econometrica* 68(3): 695–714.

Kilian, Lutz. 2008. "Exogenous Oil Supply Shocks: How Big Are They and How Much Do They Matter for the U.S. Economy?" *Review of Economics and Statistics* 90(2): 216–40.

Kim, Chang-Jin, and Charles R. Nelson. 1999. "Has the U.S. Economy Become More Stable? A Bayesian Approach Based on a Markov-Switching Model of the Business Cycle." *Review of Economics and Statistics* 81(4): 608–616.

Kuttner, Kenneth N. 2001. "Monetary Policy Surprises and Interest Rates: Evidence from the Fed Funds Futures Market." *Journal of Monetary Economics* 47(3): 523–44.

Lütkepohl, Helmut. 2013. "Identifying Structural Vector Autoregressions via Changes in Volatility." *Advances in Econometrics,* vol. 32, pp. 169–203.

Mavroeidis, Sophocles. 2004. "Weak Identification of Forward-Looking Models in Monetary Economics." *Oxford Bulletin of Economics and Statistics* 66(Supplement): 609–35.

Mavroeidis, Sophocles, Mikkel Plagborg-Møller, and James H. Stock. 2014. "Empirical Evidence on Inflation Expectations in the New Keynesian Phillips Curve." *Journal of Economic Literature* 52(1): 124–88.

McConnell, Margret M., and Gabriel Perez-Quiros. 2000. "Output Fluctuations in the United States: What Has Changed Since the Early 1980's." *American Economic Review* 90(5): 1464–76.

McCracken, Michael W., and Serena Ng. 2016. "FRED-MD: A Monthly Database for Macroeconomic Research." *Journal of Business & Economic Statistics* 34(4): 574–89.

Mertens, Karel, and Morten O. Ravn. 2013. "The Dynamic Effects of Personal and Corporate Income Tax Changes in the United States." *American Economic Review* 103(4): 1212–47.

Montiel Olea, José Luis, and Carolin Pflueger. 2013. "A Robust Test for Weak Instruments." *Journal of Business and Economic Statistics* 31(3): 358–69.

Moon, Hyungsik Roger, Frank Schorfheide, and Eleonara Granziera. 2013. "Inference for VARs Identified with Sign Restrictions." http://sites.sas.upenn.edu/schorf/files/svar-paper.pdf.

Müller, Ulrich K. 2014. "HAC Corrections for Strongly Autocorrelated Time Series." *Journal of Business and Economic Statistics* 32(3): 311–22.

Müller, Ulrich K., and Mark W. Watson. 2016a. "Measuring Uncertainty about Long-Run Predictions." *Review Economic Studies* 84(4): 1711–40.

Müller, Ulrich K., and Mark W. Watson. 2016b. "Long-Run Covariability." Unpublished paper, Princeton University.

Nason, James M., and Gregor W. Smith. 2008. "Identifying the New Keynesian Phillips Curve." *Journal of Applied Econometrics* 23(5): 525–51.

Nelson, Charles R., and Richard Startz. 1990a. "The Distribution of the Instrumental Variable Estimator and Its *t* Ratio When the Instrument Is a Poor One." *Journal of Business* 63(1, Part 2): S125–S140.

Nelson, Charles R., and Richard Startz. 1990b. "Some Further Results on the Exact Small Sample Properties of the Instrumental Variable Estimator." *Econometrica* 58(4): 967–76.

Newey, Whitney K., and Kenneth D. West. 1987. "A Simple Positive Semi-Definite, Heteroskedasticity and Autocorrelation Consistent Covariance Matrix." *Econometrica* 55(3): 703–708.

Normandin, Michel, and Louis Phaneuf. 2004. "Monetary Policy Shocks: Testing Identification Conditions under Time-Varying Conditional Volatility." *Journal of Monetary Economics* 51(6): 1217–43.

Otrok, Christopher. 2001. "On Measuring the Welfare Costs of Business Cycles." *Journal of Monetary Economics* 47(1): 61–92.

Pagan, A. R., and J. C. Robertson. 1998. "Structural Models of the Liquidity Effect." *Review of Economics and Statistics* 80(2): 202–217.

Plagborg-Møller, Mikkel. 2016. "Bayesian Inference on Structural Impulse Response Functions." http://scholar.harvard.edu/files/plagborg/files/irf_bayes.pdf.

Qu, Zhongjun. 2014. "Inference in DSGE Models with Possible Weak Identification." *Quantitative Economics* 5(2): 457–94.

Ramey, V. A. 2016. "Macroeconomic Shocks and their Propagation." Chap. 2 in *Handbook of Macroeconomics* vol. 2., edited by J. B. Taylor and H. Uhlig, 71–162. Elsevier.

Rigobon, Roberto. 2003. "Identification through Heteroskedasticity." *Review of Economics and Statistics* 85(4): 777–92.

Rigobon, Roberto, and Brian Sack. 2003. "Measuring the Reaction of Monetary Policy to the Stock Market." *Quarterly Journal of Economics* 118(2): 639–69

Rigobon, Roberto, and Brian Sack. 2004. "The Impact of Monetary Policy on Asset Prices." *Journal of Monetary Economics* 51(8): 1553–75.

Romer, Christina D., and David H. Romer. 1989. "Does Monetary Policy Matter? A New Test in the Spirit of Friedman and Schwartz." In *NBER Macroeconomics Annual 1989,* vol. 4, edited by Olivier Blanchard and Stanley Fisher, 121–70. MIT Press.

Rudebusch, Glenn D. 1998. "Do Measures of Monetary Policy in a VAR Make Sense?" *International Economic Review* 39(4): 907–931.

Sargent, Thomas J. 1989. "Two Models of Measurements and the Investment Accelerator." *Journal of Political Economy* 97(2): 251–87.

Schorfheide, Frank. 2000. "Loss Function-based Evaluation of DSGE Models." *Journal of Applied Econometrics* 15(6): 645–70.

Sentana, Enrique, and Gabriele Fiorentini. 2001. "Identification, Estimation, and Testing of Conditionally Heteroskedastic Factor Models." *Journal of Econometrics* 102(2): 143–64.

Sims, Christopher A. 1980. "Macroeconomics and Reality." *Econometrica* 48(1): 1–48.

Smets, Frank, and Raf Wouters. 2003. "An Estimated Dynamic Stochastic General Equilibrium Model of the Euro Area." *Journal of the European Economic Association* 1(5): 1123–75.

Staiger, Douglas, and James H. Stock. 1997. "Instrumental Variables Regression with Weak Instruments." *Econometrica* 65(3): 557–86.

Stock, James H. 2008. *What's New in Econometrics: Time Series, Lecture 7.* Short course lectures, NBER Summer Institute, at http://www.nber.org/minicourse_2008.html.

Stock, James H., and Mark W. Watson. 1989. "New Indexes of Coincident and Leading Economic Indicators." *NBER Macroeconomics Annual 1989*, edited by Olivier J. Blanchard and Stanley Fischer, 351–93. MIT Press.

Stock, James H., and Mark W. Watson. 1996. "Evidence on Structural Instability in Macroeconomic Time Series Relations." *Journal of Business and Economic Statistics* 14(1): 11–30.

Stock, James H., and Mark W. Watson. 2002. "Forecasting Using Principal Components from a Large Number of Predictors." *Journal of the American Statistical Association* 97(460): 1167–79.

Stock, James H., and Mark W. Watson. 2003. "Forecasting Output and Inflation: The Role of Asset Prices." *Journal of Economic Literature* 41(3): 788–829.

Stock, James H., and Mark W. Watson. 2011. *Introduction to Econometrics,* 3rd Edition. Pearson.

Stock, James H., and Mark W. Watson. 2012a. "Disentangling the Channels of the 2007–09 Recession." *Brookings Papers on Economic Activity*, no. 1, 81–135.

Stock, James H., and Mark W. Watson. 2012b. "Generalized Shrinkage Methods for Forecasting Using Many Predictors." *Journal of Business & Economic Statistics* 30(4): 481–93.

Stock, James H., and Mark W. Watson. 2016. "Dynamic Factor Models, Factor-Augmented Autoregressions, and Structural Vector Autoregressions in Macroeconomics." Chap. 8 in *Handbook of Macroeconomics,* vol. 2, edited by John B. Taylor and Harald Uhlig, 415–526. Elsevier.

Stock, James H., and Jonathan H. Wright. 2000. "GMM with Weak Identification." *Econometrica* 68(5): 1055–96.

Uhlig, Harald. 2005. "What are the Effects of Monetary Policy on Output? Results from an Agnostic Identification Procedure." *Journal of Monetary Economics* 52(2): 381–419.

Welch, Ivo, and Amit Goyal. 2008. "A Comprehensive Look at the Empirical Performance of Equity Premium Prediction." *Review of Financial Studies* 21(4): 1455–1508.

Journal of Economic Perspectives—Volume 31, Number 2—Spring 2017—Pages 87–106

Machine Learning: An Applied Econometric Approach

Sendhil Mullainathan and Jann Spiess

Machines are increasingly doing "intelligent" things: Facebook recognizes faces in photos, Siri understands voices, and Google translates websites. The fundamental insight behind these breakthroughs is as much statistical as computational. Machine intelligence became possible once researchers stopped approaching intelligence tasks procedurally and began tackling them empirically. Face recognition algorithms, for example, do not consist of hard-wired rules to scan for certain pixel combinations, based on human understanding of what constitutes a face. Instead, these algorithms use a large dataset of photos labeled as having a face or not to estimate a function $f(x)$ that predicts the presence y of a face from pixels x. This similarity to econometrics raises questions: Are these algorithms merely applying standard techniques to novel and large datasets? If there are fundamentally new empirical tools, how do they fit with what we know? As empirical economists, how can we use them?[1]

We present a way of thinking about machine learning that gives it its own place in the econometric toolbox. Central to our understanding is that machine learning

[1] In this journal, Varian (2014) provides an excellent introduction to many of the more novel tools and "tricks" from machine learning, such as decision trees or cross-validation. Einav and Levin (2014) describe big data and economics more broadly. Belloni, Chernozhukov, and Hanson (2014) present an econometrically thorough introduction on how LASSO (and close cousins) can be used for inference in high-dimensional data. Athey (2015) provides a brief overview of how machine learning relates to causal inference.

■ *Sendhil Mullainathan is the Robert C. Waggoner Professor of Economics and Jann Spiess is a PhD candidate in Economics, both at Harvard University, Cambridge, Massachusetts. Their email addresses are mullain@fas.harvard.edu and jspiess@fas.harvard.edu.*

† For supplementary materials such as appendices, datasets, and author disclosure statements, see the article page at
https://doi.org/10.1257/jep.31.2.87 doi=10.1257/jep.31.2.87

not only provides new tools, it solves a different problem. Machine learning (or rather "supervised" machine learning, the focus of this article) revolves around the problem of *prediction*: produce predictions of y from x. The appeal of machine learning is that it manages to uncover generalizable patterns. In fact, the success of machine learning at intelligence tasks is largely due to its ability to discover complex structure that was not specified in advance. It manages to fit complex and very flexible functional forms to the data without simply overfitting; it finds functions that work well out-of-sample.

Many economic applications, instead, revolve around *parameter estimation*: produce good estimates of parameters β that underlie the relationship between y and x. It is important to recognize that machine learning algorithms are not built for this purpose. For example, even when these algorithms produce regression coefficients, the estimates are rarely consistent. The danger in using these tools is taking an algorithm built for $\hat{y}$, and presuming their $\hat{\beta}$ have the properties we typically associate with estimation output. Of course, prediction has a long history in econometric research—machine learning provides new tools to solve this old problem.[2] Put succinctly, machine learning belongs in the part of the toolbox marked $\hat{y}$ rather than in the more familiar $\hat{\beta}$ compartment.

This perspective suggests that applying machine learning to economics requires finding relevant $\hat{y}$ tasks. One category of such applications appears when using new kinds of data for traditional questions; for example, in measuring economic activity using satellite images or in classifying industries using corporate 10-K filings. Making sense of complex data such as images and text often involves a prediction pre-processing step. In another category of applications, the key object of interest is actually a parameter β, but the inference procedures (often implicitly) contain a prediction task. For example, the first stage of a linear instrumental variables regression is effectively prediction. The same is true when estimating heterogeneous treatment effects, testing for effects on multiple outcomes in experiments, and flexibly controlling for observed confounders. A final category is in direct policy applications. Deciding which teacher to hire implicitly involves a prediction task (what added value will a given teacher have?), one that is intimately tied to the causal question of the value of an additional teacher.

Machine learning algorithms are now *technically* easy to use: you can download convenient packages in R or Python that can fit decision trees, random forests, or LASSO (Least Absolute Shrinkage and Selection Operator) regression coefficients. This also raises the risk that they are applied naively or their output is misinterpreted. We hope to make them *conceptually* easier to use by providing a crisper

[2] While the ideas we describe as central to machine learning may appear unfamiliar to some, they have their roots and parallels in nonparametric statistics, including nonparametric kernel regression, penalized modeling, cross-validation, and sieve estimation. We refer to Györfi, Kohler, Krzyzak, and Walk (2002) for a general overview, and to Hansen (2014) more specifically for counterparts in sieve estimation.

understanding of how these algorithms work, where they excel, and where they can stumble—and thus where they can be most usefully applied.[3]

How Machine Learning Works

Supervised machine learning algorithms seek functions that predict well out of sample. For example, we might look to predict the value y of a house from its observed characteristics x based on a sample of n houses (y_i, x_i). The algorithm would take a loss function $L(\hat{y}, y)$ as an input and search for a function $\hat{f}$ that has low expected prediction loss $E_{(y,x)}[L(\hat{f}(x), y)]$ on a *new* data point from the same distribution. Even complex intelligence tasks like face detection can be posed this way. A photo can be turned into a vector, say a 100-by-100 array so that the resulting x vector has 10,000 entries. The y value is 1 for images with a face and 0 for images without a face. The loss function $L(\hat{y}, y)$ captures payoffs from proper or improper classification of "face" or "no face."

Familiar estimation procedures, such as ordinary least squares, already provide convenient ways to form predictions, so why look to machine learning to solve this problem? We will use a concrete application—predicting house prices—to illustrate these tools. We consider 10,000 randomly selected owner-occupied units from the 2011 metropolitan sample of the American Housing Survey. In addition to the values of each unit, we also include 150 variables that contain information about the unit and its location, such as the number of rooms, the base area, and the census region within the United States. To compare different prediction techniques, we evaluate how well each approach predicts (log) unit value on a separate hold-out set of 41,808 units from the same sample. All details on the sample and our empirical exercise can be found in an online appendix available with this paper at http://e-jep.org.

Table 1 summarizes the findings of applying various procedures to this problem. Two main insights arise from this table. First, the table highlights the need for a hold-out sample to assess performance. In-sample performance may overstate performance; this is especially true for certain machine learning algorithms like random forests that have a strong tendency to overfit. Second, on out-of-sample performance, machine learning algorithms such as random forests can do significantly better than ordinary least squares, even at moderate sample sizes and with a limited number of covariates. Understanding machine learning, though, requires looking deeper than these quantitative gains. To make sense of how these

[3] This treatment is by no means exhaustive: First, we focus specifically on "supervised" machine learning where prediction is central, and do not discuss clustering or other "unsupervised" pattern recognition techniques. Second, we leave to more specialized sources the more hands-on practical advice, the discussion of computational challenges that are central to a computer-science treatment of the subject, and the overview of cutting-edge algorithms.

Table 1
Performance of Different Algorithms in Predicting House Values

	Prediction performance (R^2)		*Relative improvement over ordinary least squares by quintile of house value*				
Method	*Training sample*	*Hold-out sample*	1st	2nd	3rd	4th	5th
Ordinary least squares	47.3%	41.7% [39.7%, 43.7%]	–	–	–	–	–
Regression tree tuned by depth	39.6%	34.5% [32.6%, 36.5%]	–11.5%	10.8%	6.4%	–14.6%	–31.8%
LASSO	46.0%	43.3% [41.5%, 45.2%]	1.3%	11.9%	13.1%	10.1%	–1.9%
Random forest	85.1%	45.5% [43.6%, 47.5%]	3.5%	23.6%	27.0%	17.8%	–0.5%
Ensemble	80.4%	45.9% [44.0%, 47.9%]	4.5%	16.0%	17.9%	14.2%	7.6%

Note: The dependent variable is the log-dollar house value of owner-occupied units in the 2011 American Housing Survey from 150 covariates including unit characteristics and quality measures. All algorithms are fitted on the same, randomly drawn training sample of 10,000 units and evaluated on the 41,808 remaining held-out units. The numbers in brackets in the hold-out sample column are 95 percent bootstrap confidence intervals for hold-out prediction performance, and represent measurement variation for a fixed prediction function. For this illustration, we do not use sampling weights. Details are provided in the online Appendix at http://e-jep.org.

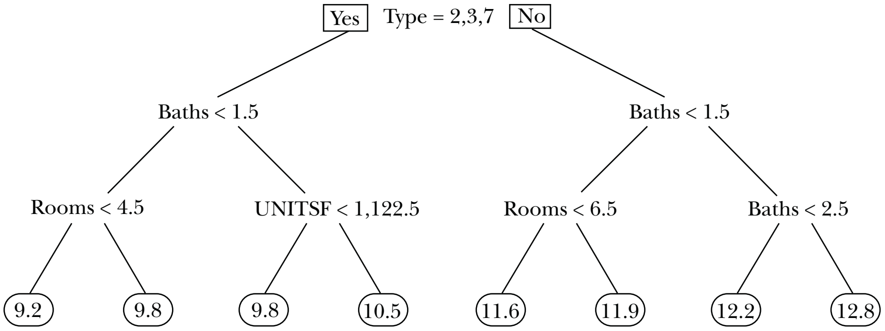
procedures work, we will focus in depth on a comparison of ordinary least squares and regression trees.

From Linear Least-Squares to Regression Trees

Applying ordinary least squares to this problem requires making some choices. For the ordinary least squares regression reported in the first row of Table 1, we included all of the main effects (with categorical variables as dummies). But why not include interactions between variables? The effect of the number of bedrooms may well depend on the base area of the unit, and the added value of a fireplace may be different depending on the number of living rooms. Simply including all pairwise interactions would be infeasible as it produces more regressors than data points (especially considering that some variables are categorical). We would therefore need to hand-curate which interactions to include in the regression. An extreme version of this challenge appears in the face-recognition problem. The functions that effectively combine pixels to predict faces will be highly nonlinear and interactive: for example, "noses" are only defined by complex interactions between numerous pixels.

Machine learning searches for these interactions automatically. Consider, for example, a typical machine learning function class: regression trees. Like a linear function, a regression tree maps each vector of house characteristics to a predicted

Figure 1
A Shallow Regression Tree Predicting House Values



Note: Based on a sample from the 2011 American Housing Survey metropolitan survey. House-value predictions are in log dollars.

value. The prediction function takes the form of a tree that splits in two at every node. At each node of the tree, the value of a single variable (say, number of bathrooms) determines whether the left (less than two bathrooms) or the right (two or more) child node is considered next. When a terminal node—a leaf—is reached, a prediction is returned. An example of a tree is given in Figure 1. We could represent the tree in Figure 1 as a linear function, where each of the leaves corresponds to a product of dummy variables ($x_1 = 1_{TYPE=2,3,7} \times 1_{BATHS<1.5} \times 1_{ROOMS<4.5}$ for the leftmost leaf) with the corresponding coefficient ($\alpha_1 = 9.2$). Trees are thus a highly interactive function class.

The Secret Sauce

How can a tree even be fitted here? A deep enough tree would fit perfectly—each observation would end up in its own leaf. That tree will have perfect *fit,* but of course this is really perfect *overfit:* out of sample, this tree would perform terribly for prediction. The (over)fitting conundrum is not specific to trees. The very appeal of machine learning is high dimensionality: flexible functional forms allow us to fit varied structures of the data. But this flexibility also gives so many possibilities that simply picking the function that fits best in-sample will be a terrible choice. So how does machine learning manage to do out-of-sample prediction?

The first part of the solution is *regularization.* In the tree case, instead of choosing the "best" overall tree, we could choose the best tree among those of a certain depth. The shallower the tree, the worse the in-sample fit: with many observations in each leaf, no one observation will be fit very well. But this also means there will be less overfit: the idiosyncratic noise of each observation is averaged out. Tree depth is an example of a regularizer. It measures the complexity of a function. As we regularize less, we do a better job at approximating the in-sample variation, but for the same reason, the wedge between in-sample and out-of-sample fit will typically increase. Machine learning algorithms typically have a regularizer associated with them. By choosing the level of regularization appropriately, we can have some benefits of flexible functional forms without having those benefits be overwhelmed by overfit.

How do we choose the optimal depth of the tree? In machine learning terms, how do we choose the level of regularization ("tune the algorithm")? This is the second key insight: *empirical tuning*. The essential problem of overfitting is that we would like the prediction function to do well *out of sample,* but we only fit in-sample. In empirical tuning, we create an out-of-sample experiment inside the original sample. We fit on one part of the data and ask which level of regularization leads to the best performance on the other part of the data.[4] We can increase the efficiency of this procedure through cross-validation: we randomly partition the sample into equally sized subsamples ("folds"). The estimation process then involves successively holding out one of the folds for evaluation while fitting the prediction function for a range of regularization parameters on all remaining folds. Finally, we pick the parameter with the best estimated average performance.[5] The second row of Table 1 shows the performance of a regression tree where we have chosen depth in this way.

This procedure works because prediction quality is observable: both predictions $\hat{y}$ and outcomes y are observed. Contrast this with parameter estimation, where typically we must rely on assumptions about the data-generating process to ensure consistency. Observability by itself would not make prediction much easier since the algorithm would still need to sort through a very large function class. But regularization turns this choice into a low-dimensional one—we only need to choose the best tuning parameter. Regularization combines with the observability of prediction quality to allow us to fit flexible functional forms and still find generalizable structure.

Most of Machine Learning in One Expression[6]

This structure—regularization and empirical choice of tuning parameters—helps organize the sometimes bewildering variety of prediction algorithms that one encounters. There is a function class $\mathcal{F}$ (in this case, trees) and a regularizer $R(f)$ (in the specific example, depth of tree) that expresses the complexity of a function

[4] One approach to the tuning problem is deriving the optimal level of regularization analytically for each procedure and under assumptions on the sampling process, such as AIC (Akaike Information Criterion), BIC (Bayesian Information Criterion), and SURE (Stein's Unbiased Risk Estimate). This theoretical guidance is helpful when available and applicable, but assumptions may prove hard to verify as the reason for undertaking nonparametric analysis may be that we are unsure about features of the data-generating processes in the first place. In other cases, theoretical results give only asymptotic guidance that remain an unverifiable promise in finite samples.

[5] In some cases, the researcher will adjust the empirical loss minimizer to account for measurement error and/or sample size differences in mapping observed performance to the level of regularization. An example is the "one standard-error rule" for LASSO tuning discussed in Hastie, Tibshirani, and Friedman (2009).

[6] We adapted the title of this section from a post on Francis X. Diebold's "No Hesitations" blog, http://fxdiebold.blogspot.com/2017/01/all-of-machine-learning-in-one.html.

Table 2
Some Machine Learning Algorithms

Function class $\mathcal{F}$ (and its parametrization)	*Regularizer $R(f)$*
Global/parametric predictors	
Linear $\beta' x$ (and generalizations)	Subset selection$\|\|\beta\|\|_0 = \sum_{j=1}^{k} 1_{\beta_j \neq 0}$ LASSO $\|\|\beta\|\|_1 = \sum_{j=1}^{k} \|\beta_j\|$ Ridge $\|\|\beta\|\|_2{}^2 = \sum_{j=1}^{k} \beta_j^2$ Elastic net $\alpha\|\|\beta\|\|_1 + (1 - \alpha)\ \|\|\beta\|\|_2{}^2$
Local/nonparametric predictors	
Decision/regression trees	Depth, number of nodes/leaves, minimal leaf size, information gain at splits
Random forest (linear combination of trees)	Number of trees, number of variables used in each tree, size of bootstrap sample, complexity of trees (see above)
Nearest neighbors	Number of neighbors
Kernel regression	Kernel bandwidth
Mixed predictors	
Deep learning, neural nets, convolutional neural networks	Number of levels, number of neurons per level, connectivity between neurons
Splines	Number of knots, order
Combined predictors	
Bagging: unweighted average of predictors from bootstrap draws	Number of draws, size of bootstrap samples (and individual regularization parameters)
Boosting: linear combination of predictions of residual	Learning rate, number of iterations (and individual regularization parameters)
Ensemble: weighted combination of different predictors	Ensemble weights (and individual regularization parameters)

(more precisely the complexity of its representation).[7] Picking the prediction function then involves two steps: The first step is, conditional on a level of complexity, to pick the best in-sample loss-minimizing function.[8] The second step is to estimate the optimal level of complexity using empirical tuning (as we saw in cross-validating the depth of the tree). In Table 2, we give an incomplete overview of methods that follow this pattern.

[7] We write the regularizer as a mapping from the function itself. In cases where functions are not uniquely parametrized (and for practical purposes in many applications), we implicitly refer to a set of parameters that define a function for a given parametrization of the function class. Also, the complexity itself may be estimated from the training data.

[8] We summarize this central step in the expression

$$\text{minimize } \underbrace{\textstyle\sum_{i=1}^{n} L(f(x_i), y_i}_{\text{in-sample loss}}, \text{ over } \overbrace{\quad f \in F \quad}^{\text{function class}} \text{ subject to } \underbrace{\quad R(f) \leq c \quad}_{\text{complexity restriction}}.$$

For example, in our framework, the LASSO (probably the machine learning tool most familiar to economists) corresponds to 1) a quadratic loss function, 2) a class of linear functions (over some fixed set of possible variables), and 3) a regularizer which is the sum of absolute values of coefficients.[9] This effectively results in a linear regression in which only a small number of predictors from all possible variables are chosen to have nonzero values: the absolute-value regularizer encourages a coefficient vector where many are exactly zero. The third row of Table 1 shows the performance of LASSO in predicting house prices. Ridge regression is a close cousin: it simply uses a quadratic regularizer instead of the sum of absolute values.

In some of the most successful machine learning methods, multiple predictors from the same function class are combined into a single prediction function and tuned jointly. The fourth row in Table 1 shows the performance of a random forest; it outperforms ordinary least squares on the hold-out by over 9 percent in terms of overall R^2. The random forest is an average over many (in this case, 700) trees. Each tree is fitted on a bootstrap sample of the original training set and constrained to a randomly chosen subset of variables. The predictions of the trees are then averaged. The regularization variables in this algorithm include the complexity of each individual tree (such as its depth), the number of variables used in each tree, the size of each bootstrap sample, and the number of trees.

The last row in Table 1 lists an ensemble method that runs several separate algorithms (in this case tree, LASSO, and forest) and then averages their predictions, with weights chosen by cross-validation. The fact that the ensemble comes out on top in Table 1—with an out-of-sample R^2 of almost 46 percent—is no isolated case. While it may be unsurprising that such ensembles perform well *on average*—after all, they can cover a wider array of functional forms—it may be more surprising that they come on top in virtually *every* prediction competition.

Other models that we have not estimated in our data also fit this framework. For example, neural nets are popular prediction algorithms for image recognition tasks. For one standard implementation in binary prediction, the underlying function class is that of nested logistic regressions: The final prediction is a logistic transformation of a linear combination of variables ("neurons") that are themselves such logistic transformations, creating a layered hierarchy of logit regressions. The complexity of these functions is controlled by the number of layers, the number of neurons per layer, and their connectivity (that is, how many variables from one level enter each logistic regression on the next).

Econometric Guidance

Viewed this way, there are several choices to be made when using a machine learning approach. First, this approach involves choosing the functions we fit and how we regularize them: Should I use a regression tree or linear functions? If I choose a tree, do I express its complexity by its depth, the minimal number of units

[9] For some readers, a more familiar equation for the LASSO is the Lagrangian dual formulation, where the Lagrange multiplier λ plays the role of the tuning parameter.

in each leaf, or the minimal improvement in prediction quality at every split? Available guidance in the machine learning literature is largely based on a combination of simulation studies and expert intuition. They are complemented by recent theoretical results in econometrics that shed light on the comparative performance of different regularizers, such as Abadie and Kasy (2017) for LASSO and close relatives.

Practically, one must decide how to encode and transform the underlying variables. In our example of house prices, do we include base area *per room* as a variable, or just total area? Should we use logarithmic scales? Normalize to unit variances? These choices about how to represent the features will interact with the regularizer and function class: A linear model can reproduce the log base area per room from log base area and log room number easily, while a regression tree would require many splits to do so. In a traditional estimator, replacing one set of variables by a set of transformed variables from which it could be reconstructed would not change the predictions, because the set of functions being chosen from has not changed. But with regularization, including these variables can improve predictions because—at any given level of regularization—the set of functions might change. If the number of bathrooms *per bedroom* is what we suspect will matter in the price-setting process, creating that variable explicitly lowers the complexity cost for including this variable. Economic theory and content expertise play a crucial role in guiding where the algorithm looks for structure first. This is the sense in which "simply throw it all in" is an unreasonable way to understand or run these machine learning algorithms. For example, in visual tasks, some understanding of geometry proves crucial in specifying the way in which neurons are connected within neural nets.

A final set of choices revolves around the tuning procedure: Should out-of-sample performance be estimated using some known correction for overfitting (such as an adjusted R^2 when it is available) or using cross-validation? How many folds should be used in cross-validation, and how exactly should the final tuning parameter be chosen? While asymptotic results show that cross-validation tuning approximates the optimal complexity (Vaart, Dudoit, and Laan 2006), available finite-sample guidance on its implementation—such as heuristics for the number of folds (usually five to ten) or the "one standard-error rule" for tuning the LASSO (Hastie, Tibshirani, and Friedman 2009)—has a more ad-hoc flavor. Design choices must be made about function classes, regularizers, feature representations, and tuning procedures: there are no definitive and universal answers available. This leaves many opportunities for econometric research.

Quantifying Predictive Performance

While these design choices leave plenty of freedom, having a reliable estimate of predictive performance is a nonnegotiable requirement for which strong econometric guarantees are available. In the house-price example, we divide the sample into a training and a test (hold-out) sample. This implements a *firewall principle*: none of the data involved in fitting the prediction function—which includes cross-validation to tune the algorithm—is used to evaluate the prediction function that is produced. As a result, inference on predictive performance of a fixed predictive function is a straightforward task of mean estimation: the distribution of realized loss in the hold-out (taking any clustering into account) directly yield consistent estimates of performance and confidence intervals.

Econometric theory plays a dual role here. First, econometrics can guide design choices, such as the number of folds or the function class. Guidance in these choices can help improve prediction quality and the power of any test based on it. Second, given the fitted prediction function, it must enable us to make inferences about estimated fit. The hold-out sample exactly allows us to form properly sized tests about predictive properties of the fitted function.

What Do We (Not) Learn from Machine Learning Output?

It is tempting to do more with the fitted function. Why not also use it to learn something about the "underlying model": specifically, why not use it to make inferences about the underlying data-generating process? Even if correlations are not causal, might they not reveal useful underlying structure? The LASSO regression of Table 1 ends up not using the number of dining rooms as a right-hand variable. Does that reveal that the number of dining rooms is an unimportant variable in determining house prices (given the other available variables)? It is tempting to draw such conclusions, and such conclusions could be economically meaningful: for example, in predicting wages, the weight placed on race by a machine learning algorithm seems like it could be a proxy for discrimination. Statistical packages contribute to these inferences by outputting measures of variable importance in the fitted functions.

One obvious problem that arises in making such inferences is the lack of standard errors on the coefficients. Even when machine-learning predictors produce familiar output like linear functions, forming these standard errors can be more complicated than seems at first glance as they would have to account for the model selection itself. In fact, Leeb and Pötscher (2006, 2008) develop conditions under which it is impossible to obtain (uniformly) consistent estimates of the distribution of model parameters after data-driven selection.

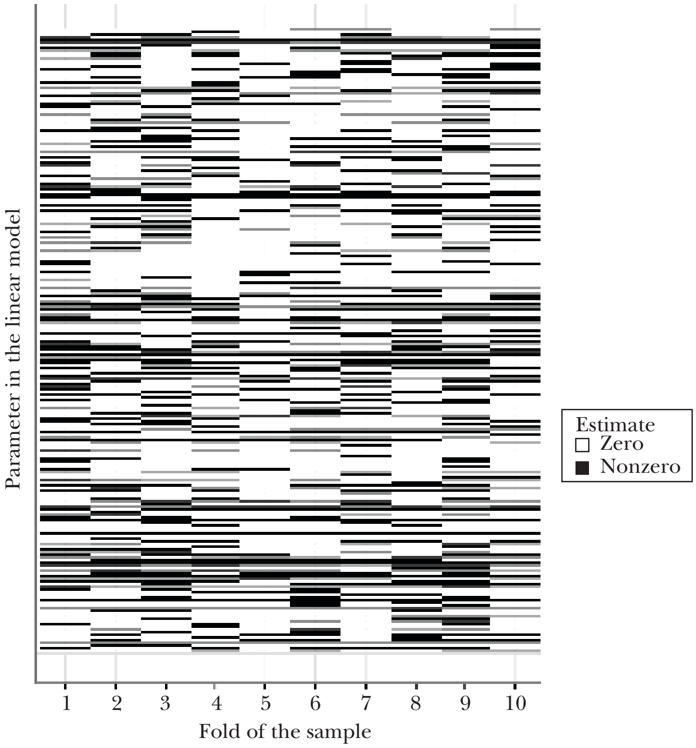
But there is an even bigger challenge. To illustrate the problem, we repeated the house-value prediction exercise on subsets of our sample from the American Housing Survey. First, we randomly cut the sample into ten partitions of approximately 5,000 units each. On each partition, we re-estimate the LASSO predictor. Through its regularizer, LASSO produces a sparse prediction function, so that many coefficients are zero and are "not used"—in this example, we find that more than half the variables are unused in each run.

Figure 2 shows how the variables that are used vary from partition to partition. Each row represents one of *x* variables used. Each column represents a different partition. We color each cell black if that variable is used by the LASSO model in that partition. Figure 2 documents a fundamental problem: a variable used in one partition may be unused in another. In fact, there are few stable patterns overall.

These instabilities do not reflect instability in prediction quality—in fact, the R^2 remains roughly constant from partition to partition. The problem arises because if

Figure 2

Selected Coefficients (Nonzero Estimates) across Ten LASSO Regressions



Note: We repeated the house-value prediction exercise on subsets of our sample from the American Housing Survey. First, we randomly cut the sample into ten partitions of approximately 5,000 units each. On each partition, we re-estimate the LASSO predictor, with LASSO regularization parameter fixed. The figure shows how the variables that are used vary from partition to partition. Each row represents one of *x* variables used. Each column represents a different partition. We color each cell black if that variable is used by the LASSO model (has a nonzero coefficient) in that partition. The figure documents a fundamental problem: a variable used in one partition may be unused in another. In fact, there are few stable patterns overall. For details, see discussion in text and online appendix available with this paper at http://e-jep.org.

the variables are correlated with each other (say the number of rooms of a house and its square-footage), then such variables are substitutes in predicting house prices. Similar predictions can be produced using very different variables. Which variables are actually chosen depends on the specific finite sample. In traditional estimation, correlations between observed variables would be reflected in large standard errors that express our uncertainty in attributing effects to one variable over the other.

This problem is ubiquitous in machine learning. The very appeal of these algorithms is that they can fit many different functions. But this creates an Achilles' heel: more functions mean a greater chance that two functions with very different coefficients can produce similar prediction quality. As a result, how an algorithm chooses between two very different functions can effectively come down to the flip of a coin. In econometric terms, while the lack of standard errors illustrates the limitations to making inference *after* model selection, the challenge here is (uniform) model selection consistency itself.

Regularization also contributes to the problem. First, it encourages the choice of less complex, but wrong models. Even if the best model uses interactions of number of bathrooms with number of rooms, regularization may lead to a choice of a simpler (but worse) model that uses only number of fireplaces. Second, it can bring with it a cousin of omitted variable bias, where we are typically concerned with correlations between observed variables and unobserved ones. Here, when regularization excludes some variables, even a correlation between observed variables and other *observed* (but excluded) ones can create bias in the estimated coefficients.

Recovering Structure: Estimation ($\hat{\beta}$) vs Prediction ($\hat{y}$)

We face a challenge. On the one hand, these machine learning algorithms by their very construction—tuning and evaluation out of sample—seek a generalizable structure and are evaluated on their capacity to find it. These algorithms do detect structure in $\hat{y}$: when predictive quality is high, some structure must have been found. Some econometric results also show the converse: when there is structure, it will be recovered at least asymptotically (for example, for prediction consistency of LASSO-type estimators in an approximately sparse linear framework, see Belloni, Chernozhukov, and Hansen 2011). On the other hand, we have seen the dangers of naively interpreting the estimated $\hat{\beta}$ parameters as indicating the discovered structure.

Of course, assumptions about the data-generating process would allow us to take the estimated $\hat{\beta}$ parameters more literally. The discussion above suggests that we must limit the correlations between the observed variables. This is seen clearly in Zhao and Yu (2006) who establish asymptotic model-selection consistency for the LASSO. Besides assuming that the true model is "sparse"—only a few variables are relevant—they also require the "irrepresentable condition" between observables: loosely put, none of the irrelevant covariates can be even moderately related to the set of relevant ones.

In practice, these assumptions are strong. The instability in Figure 2, for example, suggests that they are not realistic in the house price example. But since we know this model is finding some structure, can we characterize it? A key area of future research in econometrics and machine learning is to make sense of the estimated prediction function without making strong assumptions about the underlying true world.

How Machine Learning Can Be Applied

Our starting point for applications of machine learning algorithms is guided by both the strength of machine learning—it provides a powerful, flexible way of making quality predictions—and its weakness: absent strong and mostly unverifiable assumptions, machine learning does not produce stable estimates of the underlying parameters. Therefore, we look for $\hat{y}$ problems, places where improved prediction has large applied value.

New Data

The phrase "big data" emphasizes a change in the scale of data. But there has been an equally important change in the *nature* of this data. Machine learning can deal with unconventional data that is too high-dimensional for standard estimation methods, including image and language information that we conventionally had not even thought of as data we can work with, let alone include in a regression.

Satellites have been taking images of the earth for decades, which we can now use not just as pixelated vectors, but as economically meaningful input. Donaldson and Storeygard (in this journal, 2016) provide an overview of the growing literature in economics using satellite data, including how luminosity at night correlates with economic output (Henderson, Storeygard, and Weil 2012) or estimating future harvest size (Lobell 2013). Satellite images do not directly contain, for example, measures of crop yield. Instead, they provide us with a large x vector of image-based data; these images are then matched (in what we hope is a representative sample) to yield data which form the y variable. This translation of satellite images to yield measures is a prediction problem. Machine learning is the essential tool by which we extract and scale economically meaningful signals from this data.

These new sources of data are particularly relevant where reliable data on economic outcomes are missing, such as in tracking and targeting poverty in developing countries (Blumenstock 2016). Jean et al. (2016) train a neural net to predict local economic outcomes from satellite data in five African countries. Machine learning also yields economic predictions from large-scale network data; for example, Blumenstock, Cadamuro, and On (2015) use cell-phone data to measure wealth, allowing them to quantify poverty in Rwanda at the individual level. Image recognition can of course be used beyond satellite data, and localized prediction of economic outcomes is relevant beyond the developing world: as one example, Glaeser, Kominers, Luca, and Naik (2016) use images from Google Street View to measure block-level income in New York City and Boston.

Language provides another new powerful source of data. As with satellite images, online posts can be made meaningful by labeling them with machine learning. Kang, Kuznetsova, Luca, and Choi (2013) use restaurant reviews on Yelp.com to predict the outcome of hygiene inspections. Antweiler and Frank (2004) classify text on online financial message boards as bullish, bearish, or neither. Their algorithm trains on a small number of manual classifications, and scales these labels up to 1.5 million messages as a basis for the subsequent analysis, which shows that online messages help explain market volatility, with statistically significant, if economically modest, effects on stock returns.

Financial economists rely heavily on corporate financial information, such as that available in Compustat. But companies release detailed reports on their financial positions above and beyond these numbers. In the United States, publicly traded companies must file annual 10-K forms. Kogan, Levin, Routledge, Sagi, and Smith (2009) predict volatility of roughly 10,000 such firms from market-risk disclosure text within these forms, and show that it adds significant predictive information to past volatility. Hoberg and Phillips (2016) extract similarity of firms from their 10-K business description texts, generating new time-varying industry classifications for these firms.

Machine learning can be useful in preprocessing and imputing even in traditional datasets. In this vein, Feigenbaum (2015a, b) applies machine-learning classifiers to match individuals in historical records: he links fathers and sons across censuses and other data sources, which allows him to quantify social mobility during the Great Depression. Bernheim, Bjorkegren, Naecker, and Rangel (2013) link survey responses to observable behavior: A subset of survey respondents take part in a laboratory experiment; a machine learning algorithm trained on this data predicts actual choices from survey responses, giving economists a tool to infer actual from reported behavior.

Prediction in the Service of Estimation

A second category of application is in tasks that we approach as estimation problems. Even when we are interested in a parameter $\hat{\beta}$, the tool we use to recover that parameter may contain (often implicitly) a prediction component. Take the case of linear instrumental variables understood as a two-stage procedure: first regress $x = \gamma' z + \delta$ on the instrument z, then regress $y = \beta' x + \epsilon$ on the fitted values $\hat{x}$. The first stage is typically handled as an estimation step. But this is effectively a prediction task: only the predictions $\hat{x}$ enter the second stage; the coefficients in the first stage are merely a means to these fitted values.

Understood this way, the finite-sample biases in instrumental variables are a consequence of overfitting. Overfitting means that the in-sample fitted values $\hat{x}$ pick up not only the signal $\gamma' z$, but also the noise δ. As a consequence, $\hat{x}$ is biased towards x, and the second-stage instrumental variable estimate $\hat{\beta}$ is thus biased towards the ordinary least squares estimate of y on x. Since overfit will be larger when sample size is low, the number of instruments is high, or the instruments are weak, we can see why biases arise in these cases (Bound, Jaeger, and Baker 1995; Bekker 1994; Staiger and Stock 1997).

This analogy carries through to some of the classical solutions to finite-sample bias. Above, we used hold-out sets (in evaluating the prediction function) or cross-validation (in choosing the tuning parameter) to separate the data used in the fitting of the function from the data used in the forming of predicted values; this ensured, for example, that our evaluations of a function's prediction quality were unbiased. These same techniques applied here result in split-sample instrumental variables (Angrist and Krueger 1995) and "jackknife" instrumental variables (Angrist, Imbens, and Krueger 1999). Overfitting has wider consequences: the flipside of excessive in-sample overfitting is bad out-of-sample prediction. In fact, predicting well requires managing overfitting, which was the goal of both regularization and empirical tuning. These techniques are applicable to the first stage of instrumental variable analysis as well. In particular, a set of papers has already introduced regularization into the first stage in a high-dimensional setting, including the LASSO (Belloni, Chen, Chernozhukov, and Hansen 2012) and ridge regression (Carrasco 2012; Hansen and Kozbur 2014). More recent extensions include nonlinear functional forms, all the way to neural nets (Hartford, Leyton-Brown, and Taddy 2016).

Practically, even when there appears to be only a few instruments, the problem is effectively high-dimensional because there are many degrees of freedom in how instruments are actually constructed. For example, several papers use college proximity as an instrument in estimating returns to schooling (for example, Card 1999, Table 4). How exactly college proximity is used, however, varies. After all, it can be included linearly, logarithmically, or as dummies (and if so, with different thresholds) and can be interacted with other variables (such as demographic groups most likely to be affected). The latitude in making these choices makes it even more valuable to consider the first stage as a prediction problem. It allows us to let the data explicitly pick effective specifications, and thus allows us to recover more of the variation and construct stronger instruments, provided that predictions are constructed and used in a way that preserves the exclusion restriction.[10]

Many other inference tasks also have a prediction problem implicit inside them. In controlling for observed confounders, we do not care about the parameters associated with the control variables as an end in themselves. For example, Lee, Lessler, and Stuart (2010) use machine-learning algorithms to estimate the propensity score. Chernozhukov, Chetverikov, Demirer, Duflo, Hansen, and Newey (2016) take care of high-dimensional controls in treatment effect estimation by solving two simultaneous prediction problems, one in the outcome and one in the treatment equation.

Similar opportunities arise even in cases where we have experimental data. Consider the problem of verifying balance between treatment and control groups (such as when there is attrition). Or consider the seemingly different problem of testing for effects on many outcomes. Both can be viewed as prediction problems (Ludwig, Mullainathan, and Spiess 2017). If treatment assignment can be predicted better than chance from pretreatment covariates, this is a sign of imbalance. If treatment assignment can be predicted from a set of outcomes, the treatment must have had an effect. Estimating heterogeneous treatment effects can also be viewed as a prediction problem, though the parallel is nonobvious and implementing the transformation is a major contribution of the papers in this literature. Typically, heterogeneous treatment effects might be estimated as coefficients on interaction terms in a linear regression. Consider instead the prediction task of mapping unit-level attributes to individual effect estimates. Of course, individual-level treatment effects are not directly observed. Despite this, machine-learning methods have been successfully applied to map out treatment effect heterogeneity. Athey and Imbens (2016) use sample-splitting to obtain valid (conditional) inference on

[10] In particular, we have to avoid "forbidden regressions" (Angrist and Pischke 2008) in which correlation between first-stage residuals and fitted values exists and creates bias in the second stage.

treatment effects that are estimated using decision trees, as previously suggested by Imai and Strauss (2011). Wager and Athey (2015) extend the construction to random forests, while Grimmer, Messing, and Westwood (2016) employ ensemble methods. These heterogenous treatment effects can be used to assign treatments; Misra and Dubé (2016) illustrate this on the problem of price targeting, applying Bayesian regularized methods to a large-scale experiment where prices were randomly assigned.

Expressing parts of these inference tasks as prediction problems also makes clear the limitation on their output. A carefully constructed heterogeneity tree provides valid estimates of treatment effects in every leaf (Athey and Imbens 2016). At the same time, one must be careful in interpreting the particular representation of the chosen tree, as it may suffer from instabilities similar to those in the American Housing Survey data above. Suppose the algorithm chooses a tree that splits on education but not on age. Conditional on this tree, the estimated coefficients are consistent. But that does not imply that treatment effects do not also vary by age, as education may well covary with age; on other draws of the data, in fact, the same procedure could have chosen a tree that split on age instead.

Prediction in Policy

Consider the following policy problem: Shortly after arrest, a judge must decide where defendants will wait for trial, at home or in jail. This decision, by law, must be based on a prediction by the judge: If released, will the defendant return for their court appearance or will they skip court, and will they potentially commit further crimes? Statistical tools have helped improve policies in several ways (such as randomized control trials helping to answer "does the policy work?"). In this case, one might wonder whether a predictive algorithm could similarly help improve the judge's decision (Kleinberg et al. 2017).

Prediction policy problems, such as the bail problem, appear in many areas (Kleinberg, Ludwig, Mullainathan, and Obermeyer 2015). For instance, a large literature estimates the impact of hiring an additional teacher—this is meant to inform the decision of whether to hire more teachers. The decision of *which* teacher to hire, however, requires a prediction: the use of information available at time of hiring to forecast individual teacher quality (Kane and Staiger 2008; Dobbie 2011; Jacob et al. 2016). Chalfin et al. (2016) provide some preliminary evidence of how machine learning may improve predictive accuracy in these and other personnel decisions. Chandler, Levitt, and List (2011) predict highest-risk youth so that mentoring interventions can be appropriately targeted. Abelson, Varshney, and Sun (2014), McBride and Nichols (2016), and Engstrom, Hersh, and Newhouse (2016) use machine learning to improve poverty targeting relative to existing poverty scorecards. These predictive problems intimately relate to questions we already seek to answer: the impact of an extra teacher depends on how that teacher is chosen; the impact of a transfer program depends on how well targeted it is. Given the active research contributing to these policy discussions, expanding the focus to these adjacent prediction questions seems promising.

Economists can play a crucial role in solving prediction policy problems. First, even though prediction is important, machine learning is not enough: familiar econometric challenges arise. In deciding whether an algorithm could improve on the judge, one must resolve a basic counterfactual issue: we only know the crimes committed by those released. Many predictions problems share the feature that the available data is dictated by existing decision rules, and insights from the causal inference could prove helpful in tackling these problems; for example, Kleinberg et al. (2017) use pseudo-random assignment to judges of differing leniency in their application. Second, behavioral issues arise. Even when an algorithm can help, we must understand the factors that determine adoption of these tools (Dawes, Faust, and Meehl 1989; Dietvorst, Simmons, and Massey 2015; Yeomans, Shah, Mullainathan, and Kleinberg 2016). What factors determine faith in the algorithm? Would a simpler algorithm be more believed? How do we encourage judges to use their private information optimally? These questions combine problems of technology diffusion, information economics, and behavioral economics.

Testing Theories

A final application of supervised machine learning is to test directly theories that are inherently about predictability. Within the efficient markets theory in finance, for example, the inability to make predictions about the future is a key prediction. Moritz and Zimmermann (2016) adapt methods from machine learning to show that past returns of US firms do have significant predictive power over their future stock prices.

Machine learning can also be used to construct a benchmark for how well theories are performing. A common concern is that even if a theory is accurate, it may only clear up a little of the systematic variation it aims to explain. The R^2 alone does not address this question, as some of the total variance may not be explainable from what is measured. Kleinberg, Liang, and Mullainathan (2015) propose to compare the predictive power of a theory to that of an optimal predictor. Relatedly, Peysakhovich and Naecker (2015) compare the out-of-sample performance of behavioral economics models for choices under risk and ambiguity to an atheoretical machine-learning benchmark.

Conclusion

The problem of artificial intelligence has vexed researchers for decades. Even simple tasks such as digit recognition—challenges that we as humans overcome so effortlessly—proved extremely difficult to program. Introspection into how our mind solves these problems failed to translate into procedures. The real breakthrough came once we stopped trying to deduce these rules. Instead, the problem was turned into an inductive one: rather than hand-curating the rules, we simply let the data tell us which rules work best.

For empiricists, these theory- and data-driven modes of analysis have always coexisted. Many estimation approaches have been (often by necessity) based on top-down, theory-driven, deductive reasoning. At the same time, other approaches have aimed to simply let the data speak. Machine learning provides a powerful tool to hear, more clearly than ever, what the data have to say.

These approaches need not be in conflict. A structural model of poverty, for example, could be applied to satellite image data processed using machine learning. Theory could guide what variables to manipulate in an experiment; but in analyzing the results, machine learning could help manage multiple outcomes and estimate heterogeneous treatment effects.

In the long run, new empirical tools have also served to expand the kinds of problems we work on. The increased use of randomized control trials has also changed the kinds of questions empirical researchers work on. Ultimately, machine learning tools may also increase the scope of our work—not just by delivering new data or new methods but by focusing us on new questions.

■ *We are grateful to the editors Gordon Hanson, Enrico Moretti, and Timothy Taylor for their feedback and to Alberto Abadie, Peter Ganong, Lawrence Katz, Supreet Kaur, Jens Ludwig, and Andrei Shleifer for thoughtful and detailed comments.*

References

Abadie, Alberto, and Maximilian Kasy. 2017. *The Risk of Machine Learning.* https://ideas.repec.org/p/qsh/wpaper/383316.html.

Abelson, Brian, Kush R. Varshney, and Joy Sun. 2014. "Targeting Direct Cash Transfers to the Extremely Poor." In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining,* pp. 1563–72. ACM.

Angrist, Joshua D., Guido W. Imbens, and Alan B. Krueger. 1999. "Jackknife Instrumental Variables Estimation." *Journal of Applied Econometrics* 14(1): 57–67

Angrist, Joshua D., and Alan B. Krueger. 1995. "Split-Sample Instrumental Variables Estimates of the Return to Schooling." *Journal of Business and Economic Statistics* 13(2): 225–35.

Angrist, Joshua D., and Jörn-Steffen S. Pischke. 2008. *Mostly Harmless Econometrics: An Empiricist's Companion.* Princeton University Press.

Antweiler, Werner, and Murray Z. Frank. 2004. "Is All That Talk Just Noise? The Information Content of Internet Stock Message Boards." *Journal of Finance* 59(3): 1259–94.

Athey, Susan. 2015. "Machine Learning and Causal Inference for Policy Evaluation." In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining,* pp. 5–6. ACM.

Athey, Susan, and Guido Imbens. 2016. "Recursive Partitioning for Heterogeneous Causal Effects. *PNAS* 113(27): 7353–60.

Bekker, Paul A. 1994. "Alternative Approximations to the Distributions of Instrumental Variable Estimators." *Econometrica* 62(3): 657–81.

Belloni, Alexandre, Daniel Chen, Victor Chernozhukov, and Christian Hansen. 2012. "Sparse Models and Methods for Optimal Instruments with an Application to Eminent Domain." *Econometrica* 80(6): 2369–2429.

Belloni, Alexandre, Victor Chernozhukov, and Christian Hansen. 2011. "Inference for High-Dimensional Sparse Econometric Models." arXiv:1201.0220.

Belloni, Alexandre, Victor Chernozhukov,

and Christian Hansen. 2014. "Inference on Treatment Effects after Selection among High-Dimensional Controls." *Review of Economic Studies* 81(2): 608–650.

Bernheim, Douglas, Daniel Bjorkegren, Jeffrey Naecker, and Anatonio Rangel. 2013. "Non-Choice Evaluations Predict Behavioral Responses to Changes in Economic Conditions." NBER Working Paper 19269.

Blumenstock, Joshua E., Gabriel Cadamuro, and Robert On. 2015. "Predicting Poverty and Wealth from Mobile Phone Metadata." *Science* 350(6264): 1073–76.

Blumenstock, Joshua Evan. 2016. "Fighting Poverty with Data." *Science* 353(6301): 753–54.

Bound, John, David A. Jaeger, and Regina M. Baker. 1995. "Problems with Instrumental Variables Estimation When the Correlation between the Instruments and the Endogenous Explanatory Variable is Weak." *Journal of the American Statistical Association* 90(430): 443–50.

Card, David. 1999. "The Causal Effect of Education on Earnings." *Handbook of Labor Economics*, vol. 3A, edited by Orley C. Ashenfelter and David Card, 1801–1863. North-Holland, Elsevier.

Carrasco, Marine 2012. "A Regularization Approach to the Many Instruments Problem." *Journal of Econometrics* 170(2): 383–98.

Chalfin, Aaron, Oren Danieli, Andrew Hillis, Zubin Jelveh, Michael Luca, Jens Ludwig, and Sendhil Mullainathan. 2016. "Productivity and Selection of Human Capital with Machine Learning." *American Economic Review* 106(5): 124–27.

Chandler, Dana, Steven D. Levitt, and John A. List. 2011. "Predicting and Preventing Shootings among At-Risk Youth." *American Economic Review* 101(3): 288–92.

Chernozhukov, Victor, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, and Whitney Newey. 2016. "Double Machine Learning for Treatment and Causal Parameters." arXiv:1608.00060.

Dawes, Robyn M., David Faust, and Paul E. Meehl. 1989. "Clinical versus Actuarial Judgment." *Science* 243(4899): 1668–74.

Dietvorst, Berkeley J., Joseph P Simmons, and Cade Massey. 2015. "Algorithm Aversion: People Erroneously Avoid Algorithms after Seeing Them Err." *Journal of Experimental Psychology: General* 144(1): 114–26.

Dobbie, Will. 2011. "Teacher Characteristics and Student Achievement: Evidence from Teach For America." https://www.princeton.edu/~wdobbie/files/dobbie_tfa_2011.pdf.

Donaldson, Dave, and Adam Storeygard. 2016. "The View from Above: Applications of Satellite Data in Economics." *Journal of Economic Perspectives* 30(4): 171–98.

Engstrom, Ryan, Jonathan Hersh, and David Newhouse. 2016. "Poverty from Space: Using High Resolution Satellite Imagery for Estimating Economic Well-being and Geographic Targeting." Unpublished paper.

Einav, Liran, and Jonathan Levin. 2014. "Economics in the Age of Big Data." *Science* 346(6210): 1243089.

Feigenbaum, James J. 2015a. "Automated Census Record Linking." http://scholar.harvard.edu/jfeigenbaum/publications/automated-census-record-linking.

Feigenbaum, James J. 2015b. "Intergenerational Mobility during the Great Depression." http://scholar.harvard.edu/jfeigenbaum/publications/jmp.

Glaeser, Edward L., Scott Duke Kominers, Michael Luca, and Nakhil Naik. 2016. "Big Data and Big Cities: The Promises and Limitations of Improved Measures of Urban Life." *Economic Inquiry*, Early View Article, online July 12.

Grimmer, Justin, Solomon Messing, and Sean J. Westwood. 2016. "Estimating Heterogeneous Treatment Effects and the Effects of Heterogeneous Treatments with Ensemble Methods." https://stanford.edu/~jgrimmer/het.pdf.

Györfi, L., M. Kohler, A. Krzyzak, and H. Walk. 2002. *A Distribution-Free Theory of Nonparametric Regression*. Springer.

Hansen, Bruce E. 2014. "Nonparametric Sieve Regression: Least Squares, Averaging Least Squares, and Cross-Validation." In *The Oxford Handbook of Applied Nonparametric and Semiparametric Econometrics and Statistics,* edited by Jeffrey S. Racine, Liangjun Su, and Aman Ullah. Oxford Univerity Press.

Hansen, Christian, and Damian Kozbur. 2014. "Instrumental Variables Estimation with Many Weak instruments using Regularized JIVE." *Journal of Econometrics* 182(2): 290–308.

Hartford, Jason, Greg Lewis, Kevin Leyton-Brown, and Matt Taddy. 2016. "Counterfactual Prediction with Deep Instrumental Variables Networks." arXiv:1612.09596.

Hastie, Trevor, Robert Tibshirani, and Jerome Friedman. 2009. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction.* Second edition. New York, NY: Springer.

Henderson, J. Vernon, Adam Storeygard, and David N. Weil. 2012. "Measuring Economic Growth from Outer Space." *American Economic Review* 102(2): 994–1028.

Hoberg, Gerard, and Gordon Phillips. 2016. "Text-based Network Industries and Endogenous Product Differentiation. *Journal of Political Economy*

124(5): 1423–65.

Imai, Kosuke, and Aaron Strauss. 2011. "Estimation of Heterogeneous Treatment Effects from Randomized Experiments, with Application to the Optimal Planning of the Get-Out-the-Vote Campaign." *Political Analysis* 19(1): 1–19.

Jacob, Bryan, Jonah E. Rockoff, Eric S. Taylor, Benjamin Lindy, and Rachel Rosen. 2016. "Teacher Applicant Hiring and Teacher Performance: Evidence from DC Public Schools." NBER Working Paper 22054.

Jean, Neal, Marshall Burke, Michael Xie, W. Matthew Davis, David B. Lobell, and Stefano Ermon. 2016. "Combining Satellite Imagery and Machine Learning to Predict Poverty." *Science* 353(6301): 790–94.

Kane, Thomas J., and Douglas O. Staiger. 2008. "Estimating Teacher Impacts on Student Achievement: An Experimental Evaluation." NBER Working Paper 14607.

Kang, Jun Seok, Polina Kuznetsova, Michael Luca, and Yejin Choi. 2013. "Where *Not* to Eat? Improving Public Policy by Predicting Hygiene Inspections Using Online Reviews." EMNLP 2013: 2013 Conference on Empirical Methods in Natural Language.

Kleinberg, Jon, Annie Liang, and Sendhil Mullainathan. 2015. "The Theory is Predictive, But is It Complete? An Application to Human Perception of Randomness." Unpublished paper.

Kleinberg, Jon, Jens Ludwig, Sendhil Mullainathan, and Ziad Obermeyer. 2015. "Prediction Policy Problems." *American Economic Review* 105(5): 491–95.

Kleinberg, Jon, Himabindu Lakkaraju, Jure Leskovec, Jens Ludwig, and Sendhil Mullainathan. 2017. "Human Decisions and Machine Predictions." NBER Working Paper 23180.

Kogan, Shimon, Dimitry Levin, Byran R. Routledge, Jacob S. Sagi, and Noah A. Smith. 2009. "Predicting Risk from Financial Reports with Regression." In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics,* pp. 272–80. ACM.

Lee, Brian K., Justin Lessler, and Elizabeth A. Stuart. 2010. "Improving Propensity Score Weighting using Machine Learning." *Statistics in Medicine* 29(3): 337–46.

Leeb, Hannes, and Benedikt M. Pötscher. 2006. "Can One Estimate the Conditional Distribution of Post-Model-Selection Estimators?" *Annals of Statistics* 34(5): 2554–91.

Leeb, Hannes, and Bendikt M. Pötscher. 2008. "Can One Estimate the Unconditional Distribution of Post-Model-Selection Estimators?" *Econometric Theory* 24(2): 338–76.

Lobell, David B. 2013. "The Use of Satellite Data for Crop Yield Gap Analysis." *Field Crops Research* 143: 56–64.

Ludwig, Jens, Sendhil Mullainathan, and Jann Spiess. 2017. "Machine Learning Tests for Effects on Multiple Outcomes." Unpublished paper.

McBride, Linden, and Austin Nichols. 2016. "Retooling Poverty Targeting Using Out-of-Sample Validation and Machine Learning." *World Bank Economic Review,* lhw056.

Misra, Sanjog, and Jean-Pierre Dubé. 2016. "Scalable Price Targeting." Unpublished paper.

Moritz, Benjamin, and Tom Zimmermann. 2016. "Tree-Based Conditional Portfolio Sorts: The Relation between Past and Future Stock Returns." Available at SSRN: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2740751.

Peysakhovich, Alexander, and Jeffrey Naecker. 2015. "Using Methods from Machine Learning to Evaluate Models of Human Choice."

Staiger, Douglas, and James H. Stock. 1997. "Instrumental Variables Regression with Weak Instruments." *Econometrica* 65(3): 557–86.

van der Vaart, Aad W., Sandrine Dudoit, and Mark J. van der Laan. 2006. "Oracle Inequalities for Multi-fold Cross Validation." *Statistics and Decisions* 24(3): 351–71.

Varian, Hal R. 2014. "Big Data: New Tricks for Econometrics." *Journal of Economic Perspectives* 28(2): 3–28.

Wager, Stefan, and Susan Athey. 2015. "Estimation and Inference of Heterogeneous Treatment Effects using Random Forests." arXiv:1510.04342

Yeomans, Michael, Anuj K. Shah, Sendhil Mullainathan, and Jon Kleinberg. 2016. "Making Sense of Recommendations." Unpublished paper.

Zhao, Peng, and Bin Yu. 2006. "On Model Selection Consistency of Lasso." *Journal of Machine Learning Research* 7: 2541–63.

Journal of Economic Perspectives—Volume 31, Number 2—Spring 2017—Pages 107–124

Identification and Asymptotic Approximations: Three Examples of Progress in Econometric Theory

James L. Powell

Not long after I agreed to write an article for this issue about recent accomplishments in econometric theory, I said to myself, "Well, this will probably end badly." Like any academic endeavor, research in econometrics is more about "progress" and "themes" than "accomplishments," and it is difficult for any participant to give a balanced view of that progress without shortchanging the contributions outside his or her particular area of expertise. In addition, the last few decades have seen a growing emphasis on applied/empirical research in economics relative to theoretical/methodological research, and so it is hard to avoid appearing defensive. I recognize that a large segment of the discipline regards research contributions of theoretical econometrics as being increasingly divorced from practice. Although the ultimate aim of econometric theory is empirical application, the increasing mathematical background needed to make progress in these research areas can make the motivation and results inaccessible to those outside the field. But I think it has always been so, from the days of the Cowles Commission to the present, and I am far from convinced that the gap is widening. It can be hard to predict which strands of the theoretical literature in econometrics will ultimately be useful in practice, but the same can be said of most scholarship, and theoretical econometrics has been making progress in understanding what can be learned about models from the kind of data we observe in economics. Explaining why all economists—even practical econometricians—should care about seemingly arcane

■ *James L. Powell is the George Break and Helen Schnacke Break Distinguished Professor of Economics, University of California, Berkeley, California. His email address is powell@econ.berkeley.edu.*

† For supplementary materials such as appendices, datasets, and author disclosure statements, see the article page at
https://doi.org/10.1257/jep.31.2.107 doi=10.1257/jep.31.2.107

concerns is a worthy intellectual enterprise, even if the outcome is destined to be slanted, incomplete, and obscure.[1]

Instead of presenting a comprehensive survey of the state of the art in econometric theory, I will focus on some interrelated research areas that have received an increasing emphasis in recent years, both in theoretical econometrics and in the profession as a whole. These issues arise from some substantial changes in empirical economics: the size and quality of datasets and computational power has grown substantially, along with the size and complexity of the econometric models and the population parameters of interest. With more and better data, it is natural to expect to be able to answer more subtle questions about population relationships, and to pay more attention to the consequences of misspecification of the model for the empirical conclusions.

Much of the recent work in econometrics has emphasized two themes. The first is the fragility of statistical identification, that is, the issues that arise in seeking to characterize the possible values of model parameters when the true distribution of the observable random variables is assumed to be known. As Khan and Tamer (2010) demonstrate, "irregular identification" has strong consequences for the attainable precision of estimators and the size and power of tests of hypotheses about the key parameters in a model. When exact identification of these parameters fails, it may still be possible to make inferences about their sign or likely range of values, as is the goal of the recent contributions on partial identification.

The other, related theme in econometric theory involves the way economists make large-sample approximations to the distributions of estimators and test statistics. In standard asymptotic theory, approximate distributions of these statistics are derived assuming the model and parameters are held fixed as the sample size increases. However, it is often more realistic to use "nonstandard" or "alternative" asymptotics, which assume the model and parameters are allowed to vary with the number of observations. These issues have gained prominence as structural models have become more flexible in functional form and identification of the parameters of interest has become more delicate. In empirical practice, the number of parameters is no more "tending to infinity" than the sample size N, but it seems reasonable to expect that linking the two will give better large-sample approximations than taking limits with the number of parameters fixed while the sample size increases. Similarly, making approximations to sampling distributions where the parameters are assumed to be "moving" with the sample size toward pathological values can yield an approximate distribution theory that is less discontinuous between the regular and pathological cases.

Alternative asymptotics is also at the core of the derivation of sampling distributions of statistics for models with nonparametric components, where flexible

[1] When I was in graduate school, econometricians were compared to kickers on an (American) football team—you wanted good ones, but not too many, and they had trouble communicating with the rest of the team. An old departmental skit-party some years ago asserted that if I ever wrote a textbook, it would be titled *Econometrics, Without Applications.*

parametric forms or other "smoothing" devices are used to make approximations to general unknown functions in the model, and the number of parameters (say, K) in the approximating model is adjusted to achieve an optimal tradeoff between bias (which decreases with K) and variance (which increases in K). Thinking about this bias–variance tradeoff changes the way efficiency of estimation is characterized. In standard asymptotics, estimators are approximately unbiased with approximate variances that decline in inverse proportion to the sample size, and efficiency comparisons involve comparison of approximate variances. However, for estimators of nonparametric components, the relationship of precision to the sample size can vary across procedures, and efficiency comparisons involve relative convergence *rates* of (approximate) mean-squared error to zero, rather than the relative levels of variances. Models with many nonparametric components often have a complicated interplay of convergence rates for each, and keeping track of exponents can be a source of frustration to both casual readers and econometricians. The theoretical challenge is to find a combination of convergence rates that ensures consistency and a feasible distributional approximation for estimators of the parameters of the model that we care about.

In what follows, I will discuss how these issues of identification and alternative asymptotic approximations have been studied in three research areas: analysis of linear endogenous regressor models with many and/or weak instruments; nonparametric models with endogenous regressors; and estimation of partially identified parameters. The first illustrates the use of "alternative asymptotics" to get better approximations to the distributions of estimators, while the other two are different approaches to identification when strong functional form assumptions are relaxed. Of course, there is more to recent econometric theory than the work in these three areas, and I make no claim that the specific articles I've picked out for discussion are the most representative or most important contributions to econometric theory in recent years. However, I do think these research areas offer good examples of the progress that has been made in the field. To quote the renowned econometrician Arthur Goldberger, "econometrics is what econometricians do," and these are some things that econometricians have been doing.

Though my own research has touched on only one of these areas (nonparametric endogeneity models), I think that all three have been very active research areas and all are natural successors to prominent themes of earlier theoretical work in the 1980s and 1990s. In my view, the hot topics in theoretical econometrics back then were unit roots and cointegration, semiparametric estimation, and model misspecification. The first of these involved the pathological behavior of standard estimation procedures for time series data that were nonstationary or nearly so, and the work on "local to unity" asymptotic theory is an intellectual ancestor to the alternative asymptotic theory for models with weak instruments. Like the unit root literature, work on weak- and many-instrument problems was inspired by simulation and empirical evidence that the usual approaches to estimation and inference could be poorly behaved when the standard assumptions didn't fit well.

Similarly, the earlier research on semiparametric models and model misspecification addressed the concern that empirical results for nonlinear econometric models were sensitive to the strong parametric restrictions imposed to construct likelihood-based estimators and test statistics. Like the literature on semiparametric models, the work on nonparametric endogeneity models generally follows a "top down" approach that starts from a restrictive model and investigates the possibilities for relaxation of parametric assumptions (like a linear or other parametric form for a regression function). In contrast, the research on partial identification is more "bottom up," starting with minimal assumptions and investigating the consequences of adding identifying restrictions.

Again, the three examples considered in this article are far from exhaustive of the range of methodological problems studied by econometricians. Many other research areas in econometric theory are discussed in the other articles in this symposium, which consider issues of causality in applied econometrics, time series econometrics, structural microeconomics, machine learning, and econometric pedagogy; these review a number of substantial theoretical contributions to econometrics on topics that are not considered here.

Many Instruments and Weak Instruments

A canonical problem in econometrics is the identification and estimation of the direct effect of an endogenous regressor on an outcome variable in a linear model. For this purpose, instrumental variables estimation has become a workhorse of empirical economics. To build some intuition about the issues here, consider a simple "limited information" model with two equations. The first is an "outcome" equation for a dependent variable y in terms of a linear combination of a single endogenous regressor x and a (correlated) mean-zero error term u,

$$y = \gamma + \beta x + u.$$

The second equation is the "first stage" for the regressor x in terms of a K-dimensional vector z of instrumental variables that is independent of u and the first-stage error v,

$$x = \delta + z\pi + v.$$

A good instrumental variable (or vector of instrumental variables) will be correlated with the regressor x, but not correlated with the outcome variable y.

As is familiar from introductory econometrics classes, simply regressing y on x in the first equation will not provide a consistent estimate of the β parameter. Instead, a bias will arise as long as the covariance between the error terms u and v is nonzero, which will typically be true when x is truly endogenous. Under standard asymptotic theory, consistent estimation of β can be obtained by two-stage least squares, first regressing the x variables on the instruments z and then regressing y on the fitted value $\hat{x}$ estimated from this first-stage regression. When the number of instrumental variables in z is small and they explain a lot of the variation in x, the estimate of the β parameter will be consistent and approximately normally distributed (with the usual fine print, "under suitable regularity conditions").

However, problems arise when there are many instruments and/or if the instruments are weakly correlated with the regressors. Standard asymptotics treats the number of instruments K as fixed, which can be a poor approximation with many instruments. A large number of instrumental variables raises a risk of "overfitting"—essentially, that the correlation arising from a large number of instrumental variables may mechanically generate a high R^2, but standard asymptotic theory for instrumental variables estimation can be a poor approximation to its finite-sample behavior when the number of instruments is large relative to the sample size.

To get better large-sample approximations to the behavior of instrumental variable estimators with many or weak instruments, Bekker (1994) used nonstandard asymptotic approximations for the model under which the "first-stage" relation between the regressors and instruments was assumed to depend on the sample size. To model the first-stage overfitting phenomenon when the number of instruments K is large, Bekker assumed that the number of instruments K was proportional to the number of observations N, that is, $K/N = \alpha$ for some α between zero and one. Under this assumption, two-stage least squares for this simple regression model is also inconsistent. The key relationship can be expressed in this way in the form of a probability limit:

$$\hat{\beta}_{2SLS} = \frac{\widehat{\mathrm{Cov}}(\hat{x}, y)}{\widehat{\mathrm{Var}}(\hat{x})} \to \beta + \frac{\alpha \,\mathrm{Cov}(u, v)}{\mathrm{Var}(z\pi) + \alpha \,\mathrm{Var}(v)},$$

In this formulation, if K is small compared to the number of observations, then α is (nearly) zero and the two-stage least squares estimator is a (nearly) consistent estimate of β. However, if K is large compared to the number of observations, then the two-stage least squares estimator $\hat{\beta}_{2SLS}$ will be inconsistent as long as the error terms u and v have nonzero covariance.

Bekker (1994) shows that another standard estimator of β, the limited information maximum likelihood (LIML) estimator, is consistent even if α is not zero, but even for LIML the usual form for the asymptotic covariance matrix of the estimator changes when there are many instruments, so the standard errors for $\hat{\beta}_{LIML}$ must account for the estimation error in the first stage.

In addition to the pure overfitting issues with many instruments, inference when the instruments are weak—meaning, in the one-regressor example, that the correlation between the instrumental variables z and the regressors x is not very strong—is even more challenging, as the regression coefficient β may not be consistently estimable. A well-discussed empirical example is Angrist and Krueger's (1991) study of the earnings–education relationship. In part of their study, they used date of birth and state of residence as instrumental variables, based on the insight that the required years of school attendance and associated laws varied across states and that students could end up with one more or one less year of schooling depending on when their birthdate fell in the calendar year. But in their reanalysis of these data, Bound, Jaeger, and Baker (1995) showed, using simulation results, that the standard asymptotic theory for instrumental variables estimation can be a poor approximation to its actual finite-sample behavior when instruments are weak, as they are for the birthdate and state dummies.

Staiger and Stock (1997) analyzed nearly unidentified models, in which the instrumental variables are extremely weak, and also found two-stage least squares to be inconsistent with a nonstandard limiting distribution. The instrument z is "weak" when the first-stage slope coefficients π are close to zero, so the denominator of the two-stage least squares estimator (the variance of the fitted value $\hat{x}$) is also close to zero. In a seminal application of the "alternative asymptotics" approach, Staiger and Stock assume the first-stage slope coefficients "shrink to zero" as the sample size increases, and they show that the estimated $\hat{\beta}_{2SLS}$ no longer tends in probability to the true constant value β. Instead, it is variable even as the sample size N tends to infinity, behaving more like a ratio of two jointly normal random variables which have a covariance matrix equal to the products zu and zv of the instruments and error terms.[2]

As a check for problems of either weak or many instrumental variables, Staiger and Stock (1997) suggested what has become a popular "rule of thumb": that the usual F-statistic from the first-stage regression of x on the instruments z should exceed 10, rather than a more usual critical level from a table of the F-distribution. The F-statistic is approximately $(N-K)R^2/K(1-R^2) = (1-\alpha)R^2/\alpha(1-R^2)$. It increases as the first-stage R^2 increases and as the ratio α of number of instruments to observations declines. Staiger and Stock show that as one uses a higher F-statistic, the approximate bias of two-stage least squares relative to classical least squares diminishes.

Unlike the case with many (but not weak) instruments, where the limited information maximum likelihood (LIML) estimator is consistent and its standard errors can be corrected to obtain valid confidence intervals, in a weak instrument setting there is no way to "fix the standard errors" or to use the usual normal approximations to make inferences about the regression coefficients β using the instrumental variable, two-stage least squares, or LIML estimators.[3] The challenge in the weak instrument setting has been to construct confidence regions for β with the right coverage probability regardless of the strength of the instruments, while also making the confidence regions not too conservative, meaning they are not

[2] In earlier work, Phillips (1989) and Choi and Phillips (1992) showed, in the completely unidentified case ($\pi = 0$), that the two stage-least squares was inconsistent and has a nonstandard, mixed normal distribution, and in multiparameter problems where some regression coefficients are identified, their approximate distributions were also nonnormal.

[3] In the just-identified case, it is possible to derive the distribution of the ratio of normal variables characterizing the limiting distribution of the coefficient estimator, but that distribution depends on nuisance parameters in the covariance matrix of zu and zv that are not consistently estimable (because of the inability to consistently estimate β). Dufour (1997) shows that any confidence set with correct coverage probability in finite samples must be unbounded with positive probability when the true parameter β is nearly unidentified, so that the usual confidence intervals based on the standard asymptotic distribution of the t-statistic must have zero coverage probability when the instruments are weak.

so large as to be not useful. The primary approach to confidence set construction in the weak instrument literature has been "test statistic inversion," which means finding the set of values of β_0 for which a test of the null hypothesis H_0: $\beta = \beta_0$ fails to reject. This approach transforms the problem of finding good confidence regions into the problem of finding good hypothesis tests, where "good" means having a correct significance level under the null hypothesis, regardless of the number and strength of the instruments, and high power under the alternative H_A: $\beta \neq \beta_0$. The first priority is to get correct size (significance level), and tests with correct size can then be ranked on their power properties.

Staiger and Stock (1997) proposed two methods to obtain confidence sets for β. The first constructs confidence regions by inverting what is called the Anderson and Rubin (1949) test statistic: essentially, this approach checks whether the hypothesized residuals that emerge from an instrumental variables procedure $\varepsilon_0 = y - x\beta_0$ are linearly related to the instrument vector z using a standard F test for a regression of ε_0 on z. Staiger and Stock's second proposal combines a confidence interval for the F-statistic in the first-stage with a confidence interval using the distribution of the t-statistic for β, conditional on the first-stage F-statistic, to obtain a conservative confidence interval for β. Neither procedure was better in all cases, with the Anderson–Rubin approach dominating for smaller numbers of instruments and weaker instruments and vice-versa for the second approach. Wang and Zivot (1998) showed that the critical values for Anderson–Rubin are valid with weak instruments for all three test statistics, but confidence sets constructed using this result will generally be too conservative (larger than needed or useful), especially when the number of instruments is large.

A number of alternative procedures have been proposed for constructing confidence tests for β under weak instruments. Kleibergen (2002) proposes a modification of the Lagrange Multiplier (LM) test statistic for H_0: $\beta = \beta_0$ using an improved estimator of the first-stage coefficients π under the null hypothesis, and shows that this test has correct asymptotic size regardless of the strength of the instruments. Moreira (2003) describes a procedure to make the significance level of a test using a standard test statistic constant (termed "similar") across all possible values of the first-stage coefficient vector π, focusing on the conditional likelihood ratio (CLR) test. Andrews, Moreira, and Stock (2006) review a number of these efforts and consider the problem of finding a uniformly most powerful test of H_0: $\beta = \beta_0$ among tests that are similar. In the just-identified case with the number of instruments equaling the number of regressors, the tests proposed by Kleibergen (2002) and Moreira (2003) are equivalent to the Anderson–Rubin test, which is optimal in this setting. In the over-identified case, where the number of instruments exceeds the number of regressors, Andrews, Moreira, and Stock show that no uniformly most powerful invariant test exists, but they also found that Moreira's conditional likelihood ratio test was nearly optimal.[4]

[4] Stock and Wright (2000) and Kleibergen (2005) extend the treatment of weak identification to nonlinear models with moment restrictions on the error terms, while Andrews and Cheng (2012) give a

There are some complications in translating this theoretical result into empirical practice. The test inversion procedure described here gives a *joint* confidence set for all of the coefficients of the endogenous regressors, but individual confidence intervals for particular components can be quite conservative, and thus too broad to be useful. The method works best when the number of endogenous variables is small (like the single endogenous regressor x in the earlier example). Even for models with a single endogenous variable, the confidence region for β using the test inversion approach need not be a single interval, but may be a collection of intervals, possibly open-ended. While there have been some empirical applications of confidence sets that are valid with weak instruments, the main influence of the theoretical work on weak instruments is likely the routine use of the Staiger and Stock (1997) "rule of thumb" check of the first-stage F-statistic as a diagnostic for adequate performance of standard inference procedures for instrumental variables estimators.

Nonparametric Identification with Endogenous Regressors

While the weak instrument literature focused on identification problems in linear models, different problems arise when the linearity of the regression function is relaxed. Much research in theoretical econometrics in the 1980s and 1990s was devoted to estimation of models with nonparametric components. In these analyses, the linearity of the regression function is relaxed, and the object of interest was either the nonparametrically specified function or a parametric component of the model (a "semiparametric" problem). Existing results for nonparametric estimation of smooth and/or monotonic functions were refined and applied to econometric models, with the goal of producing estimators of any parametric components that had standard large-sample behavior (that is, their distribution was approximately normal with approximate variances shrinking to zero inversely with the sample size) and estimators of nonparametric components that had approximate mean-squared error shrinking to zero as quickly as possible. The regressors appearing as arguments of the nonparametric functions in these models were generally assumed to be exogenous—that is, the error terms were independent (at least in expectation) of the regressors.

More recently, research in theoretical econometrics has sought to extend the theoretical results available for identification and estimation of nonparametric and semiparametric models to allow for endogenous variables in the nonparametric components. Combining nonparametrics and endogeneity turns out to be tricky, and the various identification strategies for the nonparametric regression functions all face some limitations.

unified treatment of construction of nearly optimal confidence sets for a general class of weakly identified econometric models.

Early articles on nonparametric endogeneity considered a starting model that assumes a single dependent variable y and single regressor x are related by the structural equation

$$y = g(x) + \varepsilon$$

for some unobservable error term ε, where the regression function $g(x)$ is not restricted to be linear or parametric but is only required to be well-behaved, meaning that it is smooth and/or monotonic in x. If the regressor x was assumed to be exogenous, then the function $g(x)$ would be the conditional mean of y given x, that is, $g(x) = E[y|x]$, which is identified from the joint distribution of the observable variables y and x and can be consistently estimated using existing methods: for example, local averaging of values of y for nearby values of x or flexible parametric approximations to $g(x)$.

But when x is endogeneous, the conditional mean of the error term ε given x is nonzero. One traditional strategy for identification of the regression function would be to seek out some instrumental variable z, assuming the error ε has mean zero given z. This instrumental variable approach is fairly flexible, in that it does not require a full specification of a first-stage equation for the endogenous regressor x, but, as a result, identification and estimation under this restriction can be tenuous. The "reduced form" regression function $h(z)$ of the relation of y to the instruments z is identified and can be estimated using standard nonparametric methods, and the same holds for the conditional density $f(x|z)$ of the regressors given the instruments (both assumed to be continuously distributed). Alas, this implies that the unknown "structural" regression function $g(x)$ is the solution to a complicated (integral) equation involving $h(z)$ and $f(x|z)$.[5] Whether the true regression function $g(x)$ is an identified and thus a unique solution depends on how the density $f(x|z)$ depends on the instrumental variable z and how variable the instrument z is. Even if it is directly assumed that the solution $g(x)$ is identified and unique, in general the solution $\hat{g}(x)$ based on nonparametric estimators of $h(z)$ and $f(x|z)$ will not be consistent, because slight departures of the estimators of $h(z)$ and $f(x|z)$ from their true values can lead to very large departures in the estimator $\hat{g}(x)$ of the true function $g(x)$, a phenomenon known as the "ill-posed inverse problem."

Several articles have investigated the possibility of consistent estimation of the regression function $g(x)$ or similar nonparametric objects. For example, the papers Newey and Powell (2003), Ai and Chen (2003), and Blundell, Chen, and Kristensen (2007) used series approximations to estimate the unknown functions nonparametrically. Alternatively, Hall and Horowitz (2005) and Darolles, Fan,

[5] Formally, writing the moment restriction $E[\varepsilon|z] = 0$ in terms of $\varepsilon = y - g(x)$, the unknown regression function $g(x)$ is a solution to the "reduced form" integral equation

$$h(z) \equiv E[y|z] = E[g(x)|z] = \int g(x) f(x|z)\, dx,$$

where $f(x|z)$ is the density of the endogenous regressor x (assumed continuously distributed) given the instruments z.

Florens, and Renault (2011) used estimators based on kernel methods. The first two articles gave conditions under which $g(x)$ could be consistently estimated but did not derive rates of convergence of the proposed estimators to the true function, while the latter three articles provide convergence rates under different characterizations of the degree of smoothness of the regression function $g(x)$ and the conditional density function $f(x|z)$.[6] The convergence rates for the nonparametric estimators depended upon the extent to which the transformation between $g(x)$ and the reduced form function $h(x)$ is "ill-posed" or even "severely ill-posed." These theoretical results show the sensitivity of the estimators of $g(x)$ to the underlying smoothness and other features of the unknown functions, making one wonder how well the methods would perform in practice. However, Blundell, Chen, and Kristensen (2007) did find that their method gave plausible results in estimation of Engel curves for household expenditure categories when total expenditure was treated as an endogenous regressor. Also, the Ai and Chen (2003) and Blundell, Chen, and Kristensen (2007) articles considered more general econometric models with both parametric and nonparametric components and demonstrated that, even when the nonparametric components are imprecisely estimated under the instrumental variable condition, the parametric part of the model may be precisely estimated, with approximate variance inversely proportional to the sample size, so the nonparametric instrumental variables approach may be more appealing when the parametric part of the model is the main object of interest.

A different strategy for identification of the nonparametric regression function $g(x)$ with an endogenous regressor puts more structure on the first-stage relationship between the endogenous regressor x and the instrumental variables z, with the goal of constructing a "control variable" v that can be included to correct for endogeneity in a second-stage nonparametric regression. For example, suppose that the error ε is assumed to satisfy the "mean exclusion restriction" $E[\varepsilon|x, v] = E[x|v] \equiv h(v)$ for some identified variable v; under this condition, the mean of the dependent variable y given the endogenous regressor x and control variable v takes the additively separable form

$$E[y|x, v] = g(x) + h(v),$$

so the structural regression function $g(x)$ can be directly estimated from a nonparametric regression of y on x and v. In addition, the control function approach to modeling nonparametric endogeneity can also be generalized to models in which the dependent variable is not additively separable in the regression function and the error term, as discussed in Blundell and Powell (2003) and Imbens and Newey

[6] Chernozhukov and Hansen (2005) considered a "quantile" version of this regression problem by replacing the identifying restriction $E[y - g(x)|z] = 0$ with the condition $\Pr[y - g(z) \leq 0|z] = \tau$ for a known value of τ between zero and one, which yields a similar "ill-posed" integral equation.

(2009), who also considered estimation of the quantiles (percentiles) of the structural function $g(x)$.[7]

The tricky part is to find the right specification of the relationship of the regressors x and instruments z that yields the control variable v satisfying the required mean exclusion assumption. Using the control variable approach, the direct effect of x on y cannot be directly identified from features of the conditional distribution of y given x (like the conditional expectation) alone, but instead is identifiable from the conditional distribution of y given both x and a "control variable" v under the assumption of independence of the errors ε and the endogenous regressors x given v.[8] For example, the average structural function can be estimated by first estimating the nonparametric regression of y on x and v and then averaging it over the marginal distribution of v. The catch, of course, is to come up with a control variable v that is observable or estimable that satisfies this conditional independence assumption, which generally involves more assumptions on how the endogenous regressor x is related to the structural error ε. The simplest case has a first-stage equation for x that takes the form

$$x = q(z, v)$$

for some instrumental variable z and a first-stage error v; this is called a "triangular" model because x appears in the equation for the outcome variable y, but not vice-versa. Like the definition of the direct effect of x on y, the literature has a number of possible alternative definitions of the control variable v. The Newey, Powell, and Vella (1999), Pinkse (2000), and Das, Newey, and Vella (2003) papers assumed the first-stage had additive errors, $q(z, \varepsilon) = r(z) + v$, and estimated v using residuals from a nonparametric regression of x on z. This was also true for the empirical application in Blundell and Powell (2003), which estimated a model of labor force participation for which the endogenous regressor x includes "outside income" and a government benefit eligibility variable was used as an instrument. For panel data applications, Altonji and Matzkin (2005) show how symmetric functions of the regressors x over all time periods might be used to control for endogeneity of the regressors x that are specific to each time period. Imbens

[7] There is some ambiguity about the proper counterfactual summary measures of the "direct effect" of x on y when the relationship is nonseparable, that is, $y = H(x, \varepsilon)$. The literature on nonseparable endogeneity models includes various approaches: the average of the structural function $H(x, \varepsilon)$ over the marginal distribution of the errors ε (Blundell and Powell 2003), or the corresponding quantiles of the structural function (Imbens and Newey 2009); the derivative of the average structural function (in x), termed the "average partial effect" (Wooldridge 2005); the average derivative of $H(x, \varepsilon)$, over the conditional distribution of ε given x, or "local average response" (Altonji and Matzkin 2005) or derivatives of $H(x, \varepsilon)$ evaluated at particular quantiles of ε (Chesher 2003). Florens, Heckman, Meghir, and Vytlacil (2008) note that the average partial effect is the continuous-regressor analogue to the "average treatment effect" when x is a binary treatment variable, while the local average response function is the analogue of the "treatment on treated" effect.

[8] For some applications, a weaker form of independence such as conditional mean or quantile independence of ε and x given v is sufficient.

and Newey (2009) showed that an additive error was not needed as long as x was an increasing function of v, in which case the needed control variable v could be defined as the conditional distribution function of x given z evaluated at the observed random variables x and z.[9] Each of these specifications of the "control variable" v is based upon a correct specification of the relationship between x and the instrumental variables z (including a complete list of the relevant instruments, to ensure the assumption of independence of ε and x given v). The ability to specify this first-stage relationship greatly simplifies the identification of the direct effect of x on y, however that effect is defined.

Identification of structural relations for simultaneous equations systems that do not have a triangular structure (that is, when the "first-stage" equation for the endogenous variable x also depends on y) is far more challenging. Benkard and Berry (2006) noted that earlier results on nonparametric identification of systems of simultaneous equations were incorrect; corrected conditions for identification of such systems were derived by Matzkin (2008), conditions that involved a rank condition on a matrix of derivatives of the structural functions and error density that yielded the identification results for the triangular systems considered by Chesher (2003) and by Imbens and Newey (2009) as special cases. Nonparametric identification of structural equations is even more problematic when the endogenous regressor x is not continuously distributed, in which case it is generally impossible to find a control variable v that makes the error ε and endogenous regressor x conditionally independent. Chernozhukov and Hansen (2005) showed how the quantile structural function for y could be identified with a binary endogenous regressor x under the assumption that the rank ordering of the error ε was preserved conditional on instrumental variables z, but typically the theoretical results on nonparametric identification with discrete endogenous regressors involve the set identification concepts discussed in the next section.

Partial Identification and Inference

In some econometric models, the parameters of interest may not be uniquely determined by the distributions of observable data—that is, they are not "point" or "fully" identified. Instead, the population distribution may restrict the possible values of those parameters to some subset (which one hopes is relatively small) of their possible values, in which case the parameters are said to be "set" or "partially" identified.

The roots of much of the research on partial identification begin with Manski (1989, 1990), who provided examples that demonstrated how information on the identified components of an econometric model can be used to reduce the range

[9] Matzkin (2003) showed that this representation for v was observationally equivalent to a general invertible specification for the first-stage function h.

of possible values of parameters that are not fully identified. For a model with a nonrandomly missing outcome variable y that is bounded between known values y_L and y_U, Manski (1989) used the iterated expectations formula

$$E[y] = E[y|A]\,\Pr\{A\} + E[y|\,not\,A]\,\Pr\{not\,A\}$$

to show that knowledge of the conditional mean $E[y|A]$ of y for some subpopulation A, along with the proportion P_A of the population in A, would reduce the width $y_U - y_L$ of the range of possible values of the unconditional (population) mean $E[y]$ by a factor $1 - P_A$.[10] Variations on this idea were applied to obtain identification bounds for other problems with nonrandomly missing data, like treatment effects for programs with nonrandom assignment (Manski 1990) and regression functions for data with censored outcomes or regressors (Horowitz and Manski 1995, 1998).

It can be difficult to determine prior bounds for continuously distributed outcome variables. However, when the dependent variable is binary and its expectation, a probability, is the parameter of interest, then the bounds zero and one are automatic. As another useful restriction, monotonicity requirements on unknown functions can substantially tighten their identifiable regions, as those functions inherit the largest of the lower bounds (or smallest of the upper bounds) derived for the function at lower (or higher) values of its argument. Articles by Manski (1997) and Manski and Pepper (2000) demonstrated how monotonicity restrictions on unknown functions could sharpen identification bounds for parameters of an unobservable treatment response schedule when either that schedule is monotonic in the treatment variable or is monotonically related to an observable instrumental variable. Manski and Tamer (2002) derived bounds for regression functions when one of the regressors is interval-censored—that is, it is only known to lie in an interval with observable endpoints—when the regression is monotonic in the uncensored regressor.

Sometimes the bounds for the outcome variable of interest arise naturally from the economic model for the data-generating process. For example, Haile and Tamer (2003) show how bounds for the distribution function of independent private values in an English auction can be derived using the (fully identified) distribution function of order statistics of bids in those auctions. This relationship exploits the behavioral assumptions that bids never exceed valuations, and also that valuations never exceed the winning bid of a competitor. Furthermore, given bounds on the valuation distribution obtained from auctions with different numbers of bidders, tighter bounds can be obtained using the maximum of the lower bound and the minimum of the upper bound across auction sizes. Haile and Tamer are also able to obtain bounds on the optimal reserve price of the seller in this market, and discuss consistent estimation of the bounds using the empirical distribution of order statistics of bids in auctions of different sizes.

[10] The lower and upper bounds for $E[y]$ are $y_L + P_A\,(E[y|A] - y_L)$ and $y_U - P_A\,(y_U - E[y|A])$, respectively.

Another application of the partial identification strategy to a traditional econometrics problem was proposed by Honoré and Tamer (2006), who considered identification of regression parameters for a dynamic nonlinear panel data model in which the dependent variable is binary and depends upon several factors: an unknown linear combination of exogenous regressors; a lagged value of the dependent variable (to capture "state dependence," an effect emphasized by Heckman 1981); an individual-specific intercept term α (which works much like a "fixed effect"); and a time-varying error term. Identification of the underlying regression coefficients was known to be challenging for this model even when the error term has a parametric distribution, due to the difficulty in estimation of the unknown distribution of the dependent variable for the first time period, for which no lagged value is available (the "initial conditions" problem).[11] Honoré and Tamer treated the distribution of the initial value of the dependent variable as an unidentified nuisance parameter and showed how to compute the resulting identified sets of the regression parameters for common variants of the binary panel data model. Their numerical results suggested that the identification regions for the key parameters of interest were quite small for the cases they considered, suggesting that the lack of full identification for these parameters would be a secondary consideration in empirical applications of these models. A similar set-identification strategy for the regression parameters in dynamic panel data models was proposed by Chernozhukov, Fernándex-Val, Hahn, and Newey (2013); this strategy treated the conditional distribution of the "fixed effect" given the regressors as the unidentified nuisance parameter and derived estimators of bounds for the regression parameters and for average and quantile effects, in this case exploiting the boundedness of the dependent variable between zero and one.

Most of the early papers on set identification (like Haile and Tamer 2003) proposed estimation of the identified sets using nonparametric estimators of the identified components, but they less frequently derived inference procedures for the partially identified parameters in a way that accounted for the estimation error in the nonparametric components. An exception, from Horowitz and Manski (2000), considered construction of a confidence set when the estimated identified set is an interval with estimated endpoints that are approximately normally distributed around the true values. Imbens and Manski (2004) noted that a conservative 95 percent confidence interval for the entire identified set (that is, the confidence interval would cover the true identified set with probability at least equal to 95 percent) could be constructed by adding and subtracting two standard errors to the estimated upper and lower bounds (respectively). However, as Imbens and Manski noted, construction of a confidence set for the entire identification region is a more conservative objective than construction of a traditional confidence set for

[11] In a related paper, Honoré and Kyriazidou (2000) showed how the regression parameters could be (point) identified for panels with four or more time periods, by restricting attention to the subsample of observations in which the dependent variable changes in the middle two time periods while the exogenous regressors do not change, but this "regressor matching" approach can be problematic if the subsample of observations with unchanging regressors is small or empty because some regressors are continuously distributed or time specific (for example, time dummies).

the single true parameter, which can assume only one of the values in the identified set; for the latter goal, a 95 percent confidence interval would add and subtract an estimated critical value $\hat{c}_N$ times the standard error from the estimated upper and lower bounds of the identified set, where $\hat{c}_N$ depends on the sample size and the width of the identification interval and tends to either 1.645 (the one-sided normal critical value) or 1.96 (the two-sided critical value) depending on whether the true identification interval is has positive length or is a single point.

Chernozhukov, Hong, and Tamer (2007) proposed a general method to construct confidence sets for the identification regions for a parameter vector θ that is (partially) identified as the solution to a collection of *moment inequalities*—that is, all values of θ that satisfy conditions of the form

$$\mu(\theta) \equiv E[m(w, \theta)] \leq 0$$

for some known vector of functions $m(w, \theta)$ of the observable data vector w (outcome variables, regressors, and instruments) and unknown parameter θ. Moment inequalities can be used to represent the partial identification problems described earlier, and they often arise in economic models of strategic behavior, as illustrated by examples given by Chernozhukov, Hong, and Tamer and by Pakes, Porter, Ho, and Ishii (2015). The identified set of possible true values of θ is the set of values for which this equality is satisfied, and Chernozhukov, Hong, and Tamer propose a confidence region for this identified set.[12] A number of articles have proposed different methods to construct confidence sets for classes of moment inequality and other partial identification problems. Some, like Stoye (2009) and Andrews and Soares (2010), proposed confidence regions guaranteed to cover only the true values of the parameter and not the entire identified set (as in Imbens and Manski's approach). Others, including Beresteanu and Molinari (2008), Rosen (2008), and Romano and Shaikh (2010), constructed confidence sets intended to cover the entire identified set with high probability.

In addition to these general methods for statistical inference, the partial identification approach continues to find applications to thorny identification problems in structural econometric models, including triangular models with endogenous binary regressors (Chesher 2005; Shaikh and Vytlacil 2011), nonparametric regression models with endogeneity (Santos 2012), and nonseparable dynamic panel data models (Chernozhukov, Fernández-Val, Hahn, and Newey 2013). Indeed, articles in econometrics now regularly include sections discussing bounds for the parameters of interest when the assumptions for point identification fail to hold, and application of partial identification methods in econometrics remains a growth area in the field.

[12] Moment inequalities are a generalization of the familiar moment restrictions $E[m(w, \theta)] = 0$ that are the basis for generalized method-of-moment (GMM) estimation. The equality $E[m(w, \theta)] = 0$ can always be expressed as the pair of inequalities $E[m(w, \theta)] \leq 0$ and $E[-m(w, \theta)] \leq 0$. The procedure discussed here is analogous to construction of confidence regions for moment equalities using inversion of the *J*-test statistic proposed by Hansen (1982) to test the validity of over-identifying restrictions in generalized method of moments estimation.

Conclusions

The three specific research areas discussed here give a glimpse of some trends in theoretical econometrics, but they are of course not exhaustive of the progress made in the field. Browsing through outlets for econometric theory like *Econometrica*, the *Review of Economic Studies*, the *Journal of Econometrics*, and *Econometric Theory*, among others, I came across many other "wish I'd thought of that" articles. And in the past few the years, econometric theorists have worked to extend the foundational concepts of endogeneity and causal inference to increasingly complex problems in statistical inference.

In my view, the biggest current growth areas in econometrics involve analysis of "high-dimensional" models, in which, like the "many instrument" literature, the number of parameters K may be as large, or larger than, the sample size N. Such phenomena arise naturally in economic models of networks, where the number of potential links between agents in a network grows quadratically with the number of agents, and the object is to flexibly model the link probabilities or exchanges among groups of agents. The traditional "selection of regressors" problem in econometrics is another high-dimensional model when the number of potential regressors is large, and ongoing research is investigating the benefits and pitfalls of different model selection schemes. Some of these schemes are adapted from the research in statistics and computer science on "machine learning" (surveyed elsewhere in this symposium), and adapting these large-scale predictive methods to answer the causal questions of interest to economists is and will be a hot topic for econometric theory. It is hard to guess what the next "big idea" in econometrics will be, but I think that, when viewed in retrospect, it will be a logical successor to the problems considered in the three research areas discussed above.

■ *I am grateful to Bryan Graham, Bo Honoré, Whitney Newey, Michael Jansson, Demian Pouzo, and the editors for their helpful comments.*

References

Ai, Chunrong, and Xiaohong Chen. 2003. "Efficient Estimation of Models with Conditional Moment Restrictions Containing Unknown Functions." *Econometrica* 71(6): 1795–1843.

Altonji, Joseph G., and Rosa L. Matzkin. 2005. "Cross Section and Panel Data Estimators for Nonseparable Models with Endogenous Regressors." *Econometrica* 73(4): 1053–1102.

Anderson, Theodore W., and Herman Rubin. 1949. "Estimation of the Parameters of a Single Equation in a Complete System of Stochastic Equations." *Annals of Mathematical Statistics* 20(1): 46–63.

Andrews, Donald W. K., and Xu Cheng. 2012. "Estimation and Inference with Weak, Semi-Strong, and Strong Identification." *Econometrica*

80(5): 2153–2211.

Andrews, Donald W. K., Marcelo J. Moreira, and James H. Stock. 2006. "Optimal Two-Sided Invariant Similar Tests for Instrumental Variables Regression." *Econometrica* 74(3): 715–52.

Andrews, Donald W. K., and Gustavo Soares. 2010. "Inference for Parameters Defined by Moment Inequalities Using Generalized Moment Selection." *Econometrica* 78(1): 119–57.

Angrist, Joshua D., and Alan B. Krueger. 1991. "Does Compulsory School Attendance Affect Schooling and Earnings?" *Quarterly Journal of Economics* 106(4): 979–1014.

Bekker, Paul A. 1994. "Alternative Approximations to the Distributions of Instrumental Variable Estimators." *Econometrica* 62(3): 657–81.

Benkard, C. Lanier, and Steven Berry. 2006. "On the Nonparametric Identification of Nonlinear Simulataneous Equations Models: Comment on Brown (1983) and Roehrig (1988)." *Econometrica* 74(5): 1429–40.

Beresteanu, Arie, and Francesca Molinari. 2008. "Asymptotic Properties for a Class of Partially Identified Models." *Econometrica* 76(4): 763–814.

Blundell, Richard, Xiaohong Chen, and Dennis Kristensen. 2007. "Semi-Nonparametric IV Estimation of Shape-Invariant Engel Curves." *Econometrica* 75(6): 1613–69.

Blundell, Richard, and James L. Powell. 2003. "Endogeneity in Nonparametric and Semiparametric Regression Models." Chap. 8 in *Advances in Economics and Econometrics: Theory and Applications*, vol. 2, edited by Matias Dewatripont, Lars P. Hansen, and Stephen Turnovsky. Cambridge University Press.

Bound, John, David A. Jaeger, and Regina M. Baker. 1995. "Problems with Instrumental Variables Estimation when the Correlation between the Instruments and the Endogenous Explanatory Variable is Weak." *Journal of the American Statistical Association* 90(430): 443–50.

Chernozhukov, Victor, Iván Fernández-Val, Jinyong Hahn, and Whitney Newey. 2013. "Average and Quantile Effects in Nonseparable Panel Models." *Econometrica* 81(2): 535–80.

Chernozhukov, Victor, and Christian Hansen. 2005. "An IV Model of Quantile Treatment Effects." *Econometrica* 73(1): 245–61.

Chernozhukov, Victor, Han Hong, and Elie Tamer. 2007. "Estimation and Confidence Regions for Parameter Sets in Econometric Models." *Econometrica* 75(5): 1243–84.

Chesher, Andrew. 2003. "Identification in Nonseparable Models." *Econometrica* 71(5): 1405–41.

Chesher, Andrew. 2005. "Nonparametric Identification under Discrete Variation." *Econometrica* 73(5): 1525–50.

Choi, In, and Peter C. B. Phillips. 1992. "Asymptotic and Finite Sample Distribution Theory for IV Estimators and Tests in Partially Identified Structural Equations." *Journal of Econometrics* 51(1–2): 113–50.

Darolles, Serge, Yanqin Fan, Jean-Pierre Florens, and Eric Renault. 2011. "Nonparametric Instrumental Regression." *Econometrica* 79(5): 1541–65.

Das, Mitali, Whitney K. Newey, and Francis Vella. 2003. "Nonparametric Estimation of Sample Selection Models." *Review of Economic Studies* 70(1): 33–58.

Dufour, Jean-Marie. 1997. "Some Impossibility Theorems in Econometrics with Applications to Structural and Dynamic Models." *Econometrica* 65(6): 1365–87.

Florens, Jean-Paul, James J. Heckman, Costas Meghir, and Edward Vytlacil. 2008. "Identification of Treatment Effects Using Control Functions in Models with Continuous, Endogenous Treatment and Heterogeneous Effects." *Econometrica* 76(5): 1191–1206.

Haile, Philip A., and Elie Tamer. 2003. "Inference with an Incomplete Model of English Auctions." *Journal of Political Economy* 111(1): 1–51.

Hall, Peter, and Joel L. Horowitz. 2005. "Nonparametric Methods for Inference in the Presence of Instrumental Variables." *Annals of Statistics* 33(6): 2904–29.

Hansen, Lars Peter. 1982. "Large Sample Properties of Generalized Method of Moments Estimators." *Econometrica* 50(4): 1029–54.

Heckman, James J. 1981. "The Incidental Parameters Problem and the Problem of Initial Conditions in Estimating a Discrete Time–Discrete Data Stochastic Process." In *Structural Analysis of Discrete Panel Data with Econometric Applications*, edited by Charles F. Manski and Daniel McFadden, pp. 114–78. Cambridge, MA: MIT Press.

Honoré, Bo E., and Ekaterini Kyriazidou. 2000. "Panel Data Discrete Choice Models with Lagged Dependent Variables." *Econometrica* 68(4): 839–74.

Honoré, Bo E., and Elie Tamer. 2006. "Bounds on Parameters in Panel Dynamic Discrete Choice Models." *Econometrica* 74(3): 611–29.

Horowitz, Joel L., and Charles F. Manski. 1995. "Identification and Robustness with Contaminated and Corrupted Data." *Econometrica* 63(2): 281–302.

Horowitz, Joel L., and Charles F. Manski. 1998. "Censoring of Outcomes and Regressors Due to Survey Nonresponse: Identification and Estimation Using Weights and Imputations." *Journal of Econometrics* 84(1): 37–58.

Horowitz, Joel L., and Charles F. Manski. 2000. "Nonparametric Analysis of Randomized

Experiments with Missing Covariate and Outcome Data." *Journal of the American Statistical Association* 95(449): 77–84.

Imbens, Guido W., and Charles F. Manski. 2004. "Confidence Intervals for Partially Identified Parameters." *Econometrica* 72(6): 1845–57.

Imbens, Guido W., and Whitney K. Newey. 2009. "Identification and Estimation of Triangular Simultaneous Equations Models without Additivity." *Econometrica* 77(5): 1481–1512.

Khan, Shakeeb, and Elie Tamer. 2010. "Irregular Identification, Support Conditions, and Inverse Weight Estimation." *Econometrica* 78(6): 2021–42.

Kleibergen, Frank. 2002. "Pivotal Statistics for Testing Structural Parameters in Instrumental Variables Regression." *Econometrica* 70(5): 1781–1803.

Kleibergen, Frank. 2005. "Testing Parameters in GMM Without Assuming that They Are Identified." *Econometrica* 73(4): 1103–23.

Manski, Charles F. 1997. "Monotone Treatment Response." *Econometrica* 65(6): 1311–44.

Manski, Charles F. 1989. "Anatomy of the Selection Problem." *Journal of Human Resources* 24(3): 343–60.

Manski, Charles F. 1990. "Nonparametric Bounds on Treatment Effects." *American Economic Review* 80(2): 319–23.

Manski, Charles F., and John V. Pepper. 2000. "Monotone Instrumental Variables: With an Application to the Returns to Schooling." *Econometrica* 68(4): 997–1010.

Manski, Charles F., and Elie Tamer. 2002. "Inference on Regressions with Interval Data on a Regressor or Outcome." *Econometrica* 70(2): 519–46.

Matzkin, Rosa L. 2003. "Nonparametric Estimation of Nonadditive Random Functions." *Econometrica* 71(5): 1339–75.

Matzkin, Rosa L. 2008. "Identification in Nonparametric Simultaneous Equations Models." *Econometrica* 76(5): 945–78.

Moreira, Marcelo J. 2003. "A Conditional Likelihood Ratio Test for Structural Models." *Econometrica* 71(4): 1027–48.

Newey, Whitney K., and James L. Powell. 2003. "Instrumental Variable Estimation of Nonparametric Models." *Econometrica* 71(5): 1565–78.

Newey, Whitney K., James L. Powell, and Francis Vella. 1999. "Nonparametric Estimation of Triangular Simultaneous Equations Models." *Econometrica* 67(3): 565–603.

Pakes, Ariel, Jack Porter, Kate Ho, and Joy Ishii. 2015. "Moment Inequalities and Their Application." *Econometrica* 83(1): 315–34.

Phillips, Peter C. B. 1989. "Partially Identified Econometric Models." *Econometric Theory* 5(2): 181–240.

Pinkse, Joris. 2000. "Nonparametric Two-Step Regression Estimation When Regressors and Error Are Dependent." *Canadian Journal of Statistics* 28(2): 289–300.

Romano, Joseph P., and Azeem M. Shaikh. 2010**.** "Inference for the Identified Set in Partially Identified Econometric Models." *Econometrica* 78(1): 169–211.

Rosen, Adam M. 2008. "Confidence Sets for Partially Identified Parameters that Satisfy a Finite Number of Moment Inequalities." *Journal of Econometrics* 146(1): 107–117.

Santos, Andres. 2012. "Inference in Nonparametric Instrumental Variables with Partial Identification." *Econometrica* 80(1): 213–75.

Shaikh, Azeem M., and Edward J. Vytlacil. 2011. "Partial Identification in Triangular Systems of Equations with Binary Dependent Variables." *Econometrica* 79(3): 949–55.

Staiger, Douglas, and James H. Stock. 1997. "Instrumental Variables Regression with Weak Instruments." *Econometrica* 65(3): 557–86.

Stock, James H., and Jonathan H. Wright. 2000. "GMM with Weak Identification." *Econometrica* 68(5): 1055–96.

Stoye, Jörg. 2009. "More on Confidence Intervals for Partially Identified Parameters." *Econometrica* 77(4): 1299–1315.

Wang, Jiahui, and Eric Zivot. 1998. "Inference on Structural Parameters in Instrumental Variables Regression with Weak Instruments." *Econometrica* 66(6): 1389–1404.

Wooldridge, Jeffrey M. 2005. "Unobserved Heterogeneity and Estimation of Average Partial Effects." In *Identification and Inference for Econometric Models,* edited by Donald W. K. Andrews and James H. Stock. Cambridge University Press.

Journal of Economic Perspectives—Volume 31, Number 2—Spring 2017—Pages 125–144

Undergraduate Econometrics Instruction: Through Our Classes, Darkly

Joshua D. Angrist and Jörn-Steffen Pischke

As the Stones' Age gave way to the computer age, applied econometrics was mostly concerned with estimating the parameters governing broadly targeted theoretical description of the economy. Canonical examples include multi-equation macro models describing economy-wide variables like unemployment and output, and micro models characterizing the choices of individual agents or market-level equilibria. The empirical framework of the 1960s and 1970s typically sought to explain economic outcomes with the aid of a long and diverse list of explanatory variables, but no single variable of special interest.

Much of the contemporary empirical agenda looks to answer specific questions, rather than provide a general understanding of, say, GDP growth. This agenda targets the causal effects of a single factor, such as the effects of immigration on wages or the effects of democracy on GDP growth, often focusing on policy questions like the employment effects of subsidies for small business or the effects of monetary policy. Applied researchers today look for credible strategies to answer such questions.

Empirical economics has changed markedly in recent decades, but, as we document below, econometric instruction has changed little. Market-leading econometrics texts still focus on assumptions and concerns motivated by a model-driven approach to regression, aimed at helping students produce a statistically precise

■ *Joshua D. Angrist is Ford Professor of Economics, Massachusetts Institute of Technology, Cambridge, Massachusetts. Jörn-Steffen Pischke is Professor of Economics, London School of Economics, London, United Kingdom. Their email addresses are angrist@mit.edu and s.pischke@lse.ac.uk.*

[†]For supplementary materials such as appendices, datasets, and author disclosure statements, see the article page at
https://doi.org/10.1257/jep.31.2.125 doi=10.1257/jep.31.2.125

account of the processes generating economic outcomes. Much of this material prioritizes technical concerns over conceptual matters. We still see, for example, extended textbook discussions of functional form, distributional assumptions, and how to correct for serial correlation and heteroskedasticity. Yet this instructional edifice is not of primary importance for the modern empirical agenda. At the same time, newer and widely-used tools for causal analysis, like differences-in-differences and regression discontinuity methods, get cursory textbook treatment if they're mentioned at all.

How should changes in our *use* of econometrics change the way we *teach* econometrics?

Our take on this is simple. We start with empirical strategies based on randomized trials and quasi-experimental methods because they provide a template that reveals the challenges of causal inference, and the manner in which econometric tools meet these challenges. We call this framework the *design-based* approach to econometrics because the skills and strategies required to use it successfully are related to research design. This viewpoint leads to our first concrete prescription for instructional change: a revision in the manner in which we teach regression.

Regression should be taught the way it is now most often used: as a tool to control for confounding factors. This approach abandons the traditional regression framework in which all regressors are treated equally. The pedagogical emphasis on statistical efficiency and functional form, along with the sophomoric narrative that sets students off in search of "true models" as defined by a seemingly precise statistical fit, is ready for retirement. Instead, the focus should be on the set of control variables needed to insure that the regression-estimated effect of the economic variable of interest has a causal interpretation.

In addition to a radical revision of regression pedagogy, the exponential growth in economists' use of quasi-experimental methods and randomized trials in pursuit of causal effects should move these tools to center stage in the classroom. The design-based approach emphasizes single-equation instrumental variables estimators, regression-discontinuity methods, and variations on differences-in-differences strategies, while focusing on specific threats to a causal interpretation of the estimates generated by these fundamental tools.

Finally, real empirical work plays a central role in our classes. Econometrics is better taught by example than abstraction.

Causal questions and research design are not the only sort of econometric work that remains relevant. But our experience as teachers and researchers leads us to emphasize these skills in the classroom. For one thing, such skills are now much in demand: Google and Netflix post positions flagged by keywords like causal inference, experimental design, and advertising effectiveness; Facebook's data science team focuses on randomized controlled trials and causal inference; Amazon offers prospective employees a reduced form/causal/program evaluation *track*.[1]

[1] See also the descriptions of modern private sector econometric work in Ayres (2007), Brynjolfsson and McAfee (2011), Christian (2012), and Kohavi (2015).

Of course, there's econometrics to be done beyond the applied micro applications of interest to Silicon Valley and the empirical labor economics with which we're personally most engaged. But the tools we favor are foundational for almost any empirical agenda. Professional discussions of signal economic events like the Great Recession and important telecommunications mergers are almost always arguments over causal effects. Likewise, Janet Yellen and the hundreds of researchers who support her at the Fed crave reliable evidence on whether *X* causes *Y*. Purely descriptive research remains important, and there's a role for data-driven forecasting. Applied econometricians have long been engaged in these areas, but these valuable skills are the bread-and-butter of disciplines like statistics and, increasingly, computer science. These endeavors are not where our comparative advantage as economists lies. Econometrics at its best is distinguished from other data sciences by clear causal thinking. This sort of thinking is therefore what we emphasize in our classes.

Following a brief description of the shift toward design-based empirical work, we flesh out the argument for change by considering the foundations of econometric instruction, focusing on old and new approaches to regression. We then look at a collection of classic and contemporary textbooks, and a sample of contemporary reading lists and course outlines. Reading lists in our sample are more likely to cover modern empirical methods than are today's market-leading books. But most courses remain bogged down in boring and obsolete technical material.

Good Times, Bad Times

The exponential growth in economists' use of quasi-experimental methods and randomized trials is documented in Panhans and Singleton (forthcoming). Angrist and Krueger (1999) described an earlier empirical trend for labor economics, but this trend is now seen in applied microeconomic fields more broadly. In an essay on changing empirical work (Angrist and Pischke 2010), we complained about the modern macro research agenda, so we're happy to see recent design-based inroads even in empirical macroeconomics (as described in Fuchs-Schündeln and Hassan 2016). Bowen, Frésard, and Taillard (forthcoming) report on the accelerating adoption of quasi-experimental methods in empirical corporate finance.

Design-based empirical analysis naturally focuses the analyst's attention on the econometric tools featured in this work. A less obvious intellectual consequence of the shift towards design-driven research is a change in the way we use our linear regression workhorse.

Yesterday's Papers (and Today's)

The changed interpretation of regression estimates is exemplified in the contrast between two studies of education production, Summers and Wolfe (1977) and Dale and Krueger (2002). Both papers are concerned with the role of schools in generating human capital: Summers and Wolfe with the effects of elementary school characteristics on student achievement; Dale and Krueger with the effects of college characteristics on post-graduates' earnings. These questions are similar in nature, but the analyses in the two papers differ sharply.

Summers and Wolfe (1977) interpret their mission to be one of modeling the complex process that generates student achievement. They begin with a general model of education production that includes unspecified student characteristics, teacher characteristics, school inputs, and peer composition. The model is loosely motivated by an appeal to the theory of human capital, but the authors acknowledge that the specifics of how achievement is produced remain mysterious. What stands out in this framework is lack of specificity: the Summers and Wolfe regression puts the change in test scores from 3rd to 6th grade on the left-hand side, with a list of 29 student and school characteristics on the right. This list includes family income, student IQ, sex, and race; the quality of the college attended by the teacher and teacher experience; class size and school enrollment; and measures of peer composition and behavior.

The Summers and Wolfe (1977) paper is true to the 1970s empirical mission, the search for a true model with a large number of explanatory variables:

> We are confident that the coefficients describe in a reasonable way the relationship between achieving and GSES [genetic endowment and socioeconomic status], TQ [teacher quality], SQ [non-teacher school quality], and PG [peer group characteristics], for this collection of 627 elementary school students.

In the spirit of the wide-ranging regression analyses of their times, Summers and Wolfe offer no pride of place to any particular set of variables. At the same time, their narrative interprets regression estimates as capturing causal effects. They draw policy conclusions from empirical results, suggesting, for example, that schools not use the National Teacher Exam score to guide hiring decisions.

This interpretation of regression is in the spirit of Stones' Age econometrics, which typically begins with a linear regression equation meant to describe an economic process, what some would call a "structural relation." Many authors of this Age go on to say that in order to obtain unbiased or consistent estimates, the analyst must assume that regression errors are mean-independent of regressors. But since *all* regressions produce a residual with this orthogonality property, for *any* regressor included in the model, it's hard to see how this statement promotes clear thinking about causal effects.

The Dale and Krueger (2002) investigation likewise begins with a question about schools, asking whether students who attend a more selective college earn more as a result, and, like Summers and Wolfe (1977), uses ordinary least squares regression methods to construct an answer. Yet the analysis here differs in three important ways. The first is a focus on specific causal effects: there's no effort to "explain wages." The Dale and Krueger study compares students who attend more- and less-selective colleges. College quality (measured by schools' average SAT score) is but one factor that might change wages, surely minor in an R^2 sense. This highly focused inquiry is justified by the fact that the analysis aspires to answer a causal question of concern to students, parents, and policymakers.

The second distinguishing feature is a research strategy meant to eliminate selection bias: Graduates of elite schools undoubtedly earn more (on average) than those who went elsewhere. Given that elite schools select their students carefully, however, it's clear that this difference may reflect selection bias. The Dale and Krueger (2002) paper outlines a selection-on-observables research strategy meant to overcome this central problem.

The Dale and Krueger (2002) research design compares individuals who sent applications to the same set of colleges and received the same admission decisions. Within groups defined by application and admission decisions, students who attend different sorts of schools are far more similar than they would be in an unrestricted sample. The Dale and Krueger study argues that any remaining within-group variation in the selectivity of the school attended is essentially serendipitous—as good as randomly assigned—and therefore unrelated to ability, motivation, family background, and other factors related to intrinsic earnings potential. This argument constitutes the most important *econometric content* of the Dale and Krueger paper.

A third important characteristic of the Dale and Krueger (2002) study is a clear distinction between causes and controls on the right hand side of the regressions at the heart of their study. In the modern paradigm, regressors are not all created equal. Rather, only one variable at a time is seen as having causal effects. All others are controls included in service of this focused causal agenda.[2]

In education production, for example, coefficients on demographic variables and other student characteristics are unlikely to have a clear economic interpretation. For example, what should we make of the coefficient on IQ in the earlier Summers–Wolfe regression? This coefficient reveals only that two measures of intellectual ability—IQ and the dependent variable—are positively correlated after regression-adjusting for other factors. On the other hand, features of the school environment, like class sizes, can sometimes be changed by school administrators. We might indeed want to consider the implications of class size coefficients for education policy.

The modern distinction between causal and control variables on the right-hand side of a regression equation requires more nuanced assumptions than the blanket statement of regressor-error orthogonality that's emblematic of the traditional econometric presentation of regression. This difference in roles between right-hand variables that might be causal and those that are just controls should emerge clearly in the regression stories we tell our students.

Out of Control

The modern econometric paradigm exemplified by Dale and Krueger (2002) treats regression as an empirical control strategy designed to capture causal effects. Specifically, regression is an automated matchmaker that produces within-group

[2] We say "one variable at a time," because some of the Dale and Krueger (2002) models replace college selectivity with tuition as the causal variable of interest.

comparisons: there's a single causal variable of interest, while other regressors measure conditions and circumstances that we would like to hold fixed when studying the effects of this cause. By holding the control variables fixed—that is, by including them in a multivariate regression model—we hope to give the regression coefficient on the causal variable a *ceteris paribus*, apples-to-apples interpretation. We tell this story to undergraduates without elaborate mathematics, but the ideas are subtle and our students find them challenging. Detailed empirical examples showing how regression can be used to generate interesting, useful, and surprising causal conclusions help make these ideas clear.

Our instructional version of the Dale and Krueger (2002) application asks whether it pays to attend a private university, Duke, say, instead of a state school like the University of North Carolina. This converts college selectivity into a simpler, binary treatment, so that we can cast the effects of interest as generated by simple on/off comparisons. Specifically, we ask whether the money spent on private college tuition is justified by future earnings gains. This leads to the question of how to use regression to estimate the causal effect of private college attendance on earnings.

For starters, we use notation that distinguishes between cause and control. In this case, the causal regressor is P_i, a dummy variable that indicates attendance at a private college for individual i. Control variables are denoted by X_i, or given other names when specific controls are noteworthy, but in all cases distinct from the privileged causal variable, P_i. The outcome of interest, Y_i, is a measure of earnings roughly 20 years post-enrollment.

The causal relationship between private college attendance and earnings is described in terms of potential outcomes: Y_{1i}, representing the earnings of individual i were he or she to go private ($P_i = 1$), and Y_{0i}, representing i's earnings after a public education ($P_i = 0$). The causal effect of attending a private college for individual i is the difference, $Y_{1i} - Y_{0i}$. This difference can never be seen; rather, we see only Y_{1i} or Y_{0i}, depending on the value of P_i. The analyst's goal is therefore to measure an average causal effect, like $\mathrm{E}(Y_{1i} - Y_{0i})$.

At MIT (where we have both taught), we ask our private-college econometrics students to consider their personal counterfactual had they made a public-school choice instead of coming to MIT. Some of our students are seniors who have lined up jobs with the likes of Google and Goldman. Many of the people they work with at these firms—perhaps the majority—will have gone to state schools. In view of this fact, we ask our students to consider whether MIT-style private colleges really make a difference when it comes to career success.

The first contribution of a causal framework based on potential outcomes is to explain why naive comparisons of public and private college graduates are likely to be misleading. The second is to explain how an appropriately constructed regression strategy leads us to something better.

Naive comparisons between alumni of private and public universities will confound the average causal effect of private attendance with selection bias. The selection bias here reflects the fact that students who go to private colleges are, on average, from stronger family backgrounds and probably more motivated and better prepared for college. These characteristics are reflected in their *potential* earnings, that is, in how much they could earn without the benefit of a private college degree. If those who end up attending private schools had instead attended public schools, they probably would have had higher incomes anyway. This reflects the fact that public and private students have different Y_{0i}'s, on average.

To us, the most natural and useful presentation of regression is as a model of potential outcomes. Write potential earnings in the public college scenario as $Y_{0i} = \alpha + \eta_i$, where α is the mean of Y_{0i}, and η_i is the difference between this potential outcome and its mean. Suppose further that the difference in potential outcomes is a constant, β, so we can write $\beta = Y_{1i} - Y_{0i}$. Putting the pieces together gives a causal model for observed earnings

$$Y_i = \alpha + \beta P_i + \eta_i.$$

Selection bias amounts to the statement that Y_{0i} (potential earnings after going to a public college) and hence η_i depends (in a statistical sense) on P_i, that is, on the type of school one chooses.

The road to a regression-based solution to the problem of selection bias begins with the claim that the analyst has information that can be used to eliminate selection bias, that is, to purge Y_{0i} of its correlation with P_i. In particular, the modern regression modeler postulates a control variable X_i (or perhaps a set of controls). Conditional on this control variable, the private and public earnings comparison is apples-to-apples, at least on average, so those being compared have the same average Y_{0i}'s or η_i's. This *ceteris paribus*-type claim is embodied in the *conditional independence assumption* that ultimately gives regression estimates a causal interpretation:

$$\mathrm{E}(\eta_i | P_i, X_i) = \mathrm{E}(\eta_i | X_i).$$

Notice that this is a weaker and more focused assumption than the traditional presentation, which says that the error term is mean-independent of *all* regressors, that is, $\mathrm{E}(\eta_i | P_i, X_i) = 0$.

In the Dale and Krueger (2002) study, the variable X_i identifies the schools to which the college graduates in the sample had applied and were admitted. The conditional independence assumption says that, having applied to Duke and UNC and having been admitted to both, those who chose to attend Duke have the same earnings potential as those who went to the state school. Although such conditioning does not turn college attendance into a randomized trial, it provides a compelling source of control for the major forces confounding causal inference. Applicants target schools in view of their ambition and willingness to do the required work; admissions offices look carefully at applicant ability.

We close the loop linking causal inference with linear regression by introducing a functional form hypothesis, specifically that the conditional mean of potential earnings when attending a public school is a linear function of X_i. This can be written formally as $\mathrm{E}(\eta_i | X_i) = \gamma X_i$. Econometrics texts fret at length about linearity and its limitations, but we see such hand-wringing as misplaced. In the Dale and Krueger research design, the controls are a large set of dummies for all possible applicant groups. The key controls in this case come in the form of a saturated model, that is, an exhaustive set of dummies for all possible values of the conditioning variable. Such models are inherently linear. In other cases, we can come as close as we like to the underlying conditional mean function by adding polynomial terms and interactions. When samples are small, we happily use linearity to interpolate, thereby using the data at hand more efficiently. In some of the Dale and Krueger models, for example, dummies for groups of schools are replaced by a linear control for the schools' average selectivity (that is, the average SAT scores of their students).

Combining these three ingredients, constant causal effects, conditional independence, and a linear model for potential outcomes conditional on controls, produces the regression model

$$Y_i = \alpha + \beta P_i + \gamma X_i + e_i,$$

which can be used to construct unbiased and consistent estimates of the causal effect of private school attendance, β. The causal story that brings us to this point reveals what we mean by β and why we're using regression to estimate it.

This final equation looks like many seen in market-leading texts. But this apparent similarity is less helpful than a source of confusion. In our experience, to present this equation and recite assumptions about the correlation of regressors and e_i clouds more than clarifies the basis for causal inference. As far as the control variables go, regressor-residual orthogonality is *assured* rather than assumed; that is, regression algebra makes this happen. At the same time, while the controls are surely uncorrelated with the residuals, it's unlikely that the regression coefficients multiplying the controls have a causal interpretation. We don't imagine that the controls are as good as randomly assigned and we needn't care whether they are. The controls have a job to do: *they are the foundation for the conditional independence claim* that's central to the modern regression framework. Provided the controls make this claim plausible, the coefficient β can be seen as a causal effect.

The modern regression paradigm turns on the notion that the analyst has data on control variables that generate apples-to-apples comparisons for the variable of interest. Dale and Krueger (2002) explain what this means in their study:

> If, conditional on gaining admission, students choose to attend schools for reasons that are independent of [unobserved determinants of earnings] then students who were accepted and rejected by the same set of schools would have the same expected value of [these determinants, the error term in their model]. Consequently, our proposed solution to the school selection problem is to include an unrestricted set of dummy variables indicating groups of students who received the same admissions decisions (i.e., the same combination of acceptances and rejections) from the same set of colleges.

In our analysis of the Dale and Krueger data (reported in Chapter 2 of Angrist and Pischke 2015), estimates from a regression with no controls show a large private school effect of 13.5 log points. This effect shrinks to 8.6 log points after controlling for the student's own SAT scores, his or her family income, and a few more demographic variables. But controlling for the schools to which a student applied and was admitted (using many dummy variables) yields a small and statistically insignificant private school effect of less than 1 percent.

Comparing regression results with increasing numbers of controls in this way—that is, comparing uncontrolled results, results with crude controls, and results with a control variable that more plausibly addresses the issue of selection bias—offers powerful insights. These insights help students understand why the last model is more likely to have a causal interpretation than the first two.

First, we note in discussing these results that the large uncontrolled private differential in wages is apparently driven by selection bias. We learn this from the fact that the raw effect vanishes after controlling for students' precollege attributes, in this case, ambition and ability as reflected in the set of schools a student applies to and qualifies for. Of course, there may still be selection bias in the private–public contrast conditional on these controls. But because the controls are coded from application and admissions decisions that predate college enrollment decisions, they cannot themselves be a consequence of private school attendance. They must be associated with differences in Y_{0i} that generate selection bias. Eliminating these differences, that is, comparing students with similar Y_{0i}'s, is therefore likely to generate private school effects that are less misleading than simpler models omitting these controls.

We also show our students that after conditioning on the application and admissions variables, ability and family background variables in the form of SAT scores and family income are uncorrelated with private school attendance. The finding of a zero private-school return is therefore remarkably insensitive to further control beyond a core set. This argument uses the omitted variables bias formula, which we see as a kind of golden rule for the modern regression practitioner. Our regression estimates reveal robustness to further control that we'd expect to see in a well-run randomized trial.

Using a similar omitted-variables-type argument, we note that even if there are other confounders that we haven't controlled for, those that are positively correlated with private school attendance are likely to be positively correlated with earnings as well. Even if these variables remain omitted, their omission leads the estimates computed with the variables at hand to *overestimate* the private school premium, small as it already is.

Empirical applications like this demonstrate the modern approach to regression, highlighting the nuanced assumptions needed for a causal interpretation of regression parameters.[3] If the conditional independence assumption is violated,

[3] In a recent publication, Arcidiacono, Aucejo, and Hotz (2016) use the Dale and Krueger conditioning strategy to estimate causal effects of enrolling at different University of California campuses on graduation and college major.

regression methods fail to uncover causal effects and are likely to be misleading. Otherwise, there's hope for causal inference. Alas, the regression topics that dominate econometrics teaching, including extensive discussions of classical regression assumptions, functional form, multicollinearity, and matters related to statistical inference and efficiency, pale in importance next to this live-or-die fact about regression-based research designs.

Which is not to say that causal inference using regression methods has now been made easy. The question of what makes a good control variable is one of the most challenging in empirical practice. Candidate control variables should be judged by whether they make the conditional independence assumption more plausible, and it's often hard to tell. We therefore discuss many regression examples with our students, all interesting, but some more convincing than others. A particular worry is that not all controls are good controls, even if they're related to both P_i and Y_i. Specific examples and discussion questions—"Should you control for occupation in a wage equation meant to measure the economic returns to schooling?"—illuminate the bad-control issue and therefore warrant time in the classroom (and in our books, Angrist and Pischke 2009, 2015).

Take It or Leave It: Classical Regression Concerns

It is easiest to use the conditional independence assumption to derive a causal regression model when the causal effect is the same for everyone, as assumed above. While this is an attractive simplification for expository purposes, the key result is remarkably general. As long as the regression function is suitably flexible, the regression parameter capturing the causal effect of interest is a weighted average of underlying covariate-specific causal effects. In fact, with discrete controls, regression can be viewed as a matching estimator that automates the estimation of many possibly heterogeneous covariate-specific treatment effects, producing a single weighted average in one easy step.

More generally, linearity of the regression function is best seen as a convenient approximation to possibly nonlinear functional forms. This claim is supported by pioneering theoretical studies such as White (1980a) and Chamberlain (1982). To the best of our knowledge, the first textbook to highlight these properties is Goldberger (1991), a graduate text never in wide use and one rarely seen in undergraduate courses. Angrist (1998), Angrist and Krueger (1999), and our graduate text (Angrist and Pishke 2009) develop the theoretical argument that regression is a matching estimator for average treatment effects (see also Yitzhaki 1996).

An important consequence of this approximation and matchmaking view of regression is that the assumptions behind the textbook linear regression model are both implausible and irrelevant. Heteroskedasticity arises naturally as a result of variation in the closeness between a regression fit and the underlying conditional mean function it approximates. But the fact that the quality of the fit may vary does not obviate the value of regression as a summarizer of economically meaningful causal relationships.

Classical regression assumptions are helpful for the derivation of regression standard errors. They simplify the math and the resulting formula reveals the features of the data that determine statistical precision. This derivation takes little of our class time, however. We don't dwell on statistical tests for the validity of classical assumptions or on generalized least squares fix-ups for their failures. It seems to us that most of what is usually taught on inference in an introductory undergraduate class can be replaced with the phrase "use robust standard errors." With a caution about blind reliance on asymptotic approximations, we suggest our students follow current research practice. As noted by White (1980b) and others, the robust formula addresses the statistical consequences of heteroskedasticity and nonlinearity in cross-sectional data. Autocorrelation in time-series data can similarly be handled by Newey and West (1987) standard errors, while cluster methods address correlation across cross-sectional units or in panel data (Moulton 1986; Arellano 1987; Bertrand, Duflo, and Mullainathan 2004).

In Another Land: Econometrics Texts and Teaching

Traditional econometrics textbooks are thin on empirical examples. In Johnston's (1972) classic text, the first empirical application is a bivariate regression linking road casualties to the number of licensed vehicles. This example focuses on computation, an understandable concern at the time, but Johnston doesn't explain why the relationship between casualties and licenses is interesting or what the estimates might mean. Gujarati's (1978) first empirical example is more substantive, a Cobb–Douglas production function estimated with a few annual observations. Production functions, implicitly causal relationships, are a fundamental building block of economic theory. Gujarati's discussion helpfully interprets magnitudes and considers whether the estimates might be consistent with constant returns to scale. But this application doesn't appear until page 107.

Decades later, real empirical work was still sparse in the leading texts, and the presentation of empirical examples often remained focused on mathematical and statistical technicalities. In an essay published 16 years ago in this journal, Becker and Greene (2001) surveyed econometrics texts and teaching at the turn of the millennium:

> Econometrics and statistics are often taught as branches of mathematics, even when taught in business schools ... the focus in the textbooks and teaching materials is on presenting and explaining theory and technical details with secondary attention given to applications, which are often manufactured to fit the procedure at hand ... applications are rarely based on events reported in financial newspapers, business magazines or scholarly journals in economics.

Following a broader trend towards empiricism in economic research (documented in Hammermesh 2013 and Angrist, Azoulay, Ellison, Hill, and Lu forthcoming), today's texts are more empirical than those they've replaced. In particular, modern econometrics texts are more likely than those described by Becker and Greene to integrate empirical examples throughout, and often come with access to websites where students can find real economic data for problem sets and practice.

But the news on the textbook front is not all good. Many of today's textbook examples are still contrived or poorly motivated. More disappointing to us than the uneven quality of empirical applications in the contemporary econometrics library is the failure to discuss modern empirical tools. Other than Stock and Watson (2015), which comes closest to embracing the modern agenda, none of the modern undergraduate econometrics texts surveyed below mentions regression-discontinuity methods, for example. Likewise, we see little or no discussion of the threats to validity that might confound differences-in-differences–style policy analysis, even though empirical work of this sort is now ubiquitous. Econometrics texts remain focused on material that's increasingly irrelevant to empirical practice.

To put these and other claims about textbook content on a firmer empirical foundation, we classified the content of 12 books (listed in online Appendix Table A1), six from the 1970s and six currently in wide use. Our list of classics was constructed by identifying 1970s-era editions of the volumes included in Table 1 of Becker and Green (2001), which lists undergraduate textbooks in wide use when they wrote their essay. We bought copies of these older first or second edition books. Our list of classic texts contains Kmenta (1971), Johnston (1972), Pindyck and Rubinfeld (1976), Gujarati (1978), Intriligator (1978), and Kennedy (1979). The divide between graduate and undergraduate books was murkier in the 1970s: unlike today's undergraduate books, some of these older texts use linear algebra. Intriligator (1978), Johnston (1972), and Kmenta (1971) are noticeably more advanced than the other three. We therefore summarize 1970s book content with and without these three included.

Our contemporary texts are the six most often listed books on reading lists found on the Open Syllabus Project website (http://opensyllabusproject.org/). Specifically, our modern market leaders are those found at the top of a list generated by filtering the Project's "syllabus explorer" search engine for "Economics" and then searching for "Econometrics." The resulting list consists of Kennedy (2008), Gujarati and Porter (2009), Stock and Watson (2015), Wooldridge (2016), Dougherty (2016), and Studenmund (2017).[4]

Recognizing that such an endeavor will always be imperfect, we classified book content into the categories shown in Table 1. This scheme covers the vast majority of the material in the books on our list, as well as in many others we've used or read. Our classification scheme also covers three of the tools for which growth in usage appears most impressive in the bibliometric data tabulated by Panhans and Singleton (forthcoming), specifically, instrumental variables, regression-discontinuity methods,

[4] These books are also ranked highly in Amazon's econometrics category and (at one edition removed) are market leaders in sales data from Nielsen for 2013 and 2014. Dougherty (2016) is number eight on the list yielded by Open Syllabus, but the sixth book, Hayashi (2000), is clearly a graduate text, and the seventh, Maddala (1977), is not particularly recent.

Table 1
Topic Descriptions

Topic	*Which includes …*
Bivariate regression	Basic exposition of the bivariate regression model, interpretation of bivariate model parameters
Regression properties	Derivation of estimators, classical linear regression assumptions, mathematical properties of regression estimators like unbiasedness and regression anatomy, the Gauss–Markov Theorem
Regression inference	Derivation of standard errors for coefficients and predicted values, hypothesis testing and confidence intervals, R^2, analysis of variance, discussion and illustration of inferential reasoning
Multivariate regression	General discussion of the multivariate regression model, interpretation of multivariate parameters
Omitted variables bias	Omitted variables bias in regression models
Assumption failures and fix-ups	Discussion of classical assumption failures including heteroskedasticity, serial correlation, non-normality, and stochastic regressors; multicollinearity, inclusion of irrelevant variables, generalized least squares (GLS) fix-ups
Functional form	Discussion of functional form and model parametrization issues including the use of dummy variables, logs on the left and right, limited dependent variable models, other nonlinear regression models
Instrumental variables	Instrumental variables (IV), two-stage least squares (2SLS), and other single equation IV-estimators like limited information maximium likelihood (LIML) and *k*-class estimators, the use of IV for omitted variables and errors-in-variables problems
Simultaneous equations models	Discussion of multi-equation models and estimators, including identification of simultaneous equation systems and system estimators like seemingly unrelated regressions (SUR) and three-stage least squares (3SLS)
Panel data	Panel techniques and topics, including the definition and estimation of models with fixed and random effects, pooling time series and cross section data, and grouped data
Time series	Time series issues, including distributed lag models, stochastic processes, autoregressive integrated moving average (ARIMA) modeling, vector autoregressions, and unit root tests. This category omits narrow discussions of serial correlation as a violation of classical assumptions
Causal effects	Discussion of causal effects and the causal interpretation of econometric estimates, the purpose and interpretation of randomized experiments, and threats to a causal interpretation of econometric estimates including sample selection issues
Differences-in-differences	Differences-in-differences assumptions and estimators
Regression discontinuity methods	Sharp and fuzzy regression discontinuity designs and estimators

and differences–in-differences estimators.[5] Our classification strategy counts pages devoted to each topic, omitting material in appendices and exercises, and omitting remedial material on mathematics and statistics. Independently, we also counted pages devoted to real empirical examples, that is, presentations of econometric results computed using genuine economic data. This scheme for counting examples omits the many textbook illustrations that use made-up numbers.

Not Fade Away

For the most part, legacy texts have a uniform structure: they begin by introducing a linear model for an economic outcome variable, followed closely by stating that the error term is assumed to be either mean-independent of, or uncorrelated with, regressors. The purpose of this model—whether it is a causal relationship in the sense of describing the consequences of regressor manipulation, a statistical forecasting tool, or a parameterized conditional expectation function—is usually unclear.

The textbook introduction of a linear model with orthogonal or mean-independent errors is typically followed by a list of technical assumptions like homoskedasticity, variable (yet nonstochastic!) regressors, and lack of multicollinearity. These assumptions are used to derive the good statistical properties of the ordinary least squares estimator in the classical linear model: unbiasedness, simple formulas for standard errors, and the Gauss–Markov Theorem, (in which ordinary least squares is shown to be a best linear unbiased estimator, or BLUE). As we report in Table 2, this initial discussion of *Regression properties* consumes an average of 11 to 12 percent of the classic textbooks. *Regression inference*, which usually comes next, gets an average of roughly 13 percent of page space in these traditional books.

The most deeply covered topic in our taxonomy, accounting for about 20 percent of material in the classic textbooks, is *Assumption failures and fix-ups*. This includes diagnostics and first aid for problems like autocorrelation, heteroskedasticity, and multicollinearity. Relief for most of these maladies comes in the form of generalized least squares. Another important topic in legacy texts is *Simultaneous equations models*, consuming 14 percent of page space in the more elementary texts. The percentage given over to orthodox simultaneous equations models rises to 18 percent when the sample includes more advanced texts. Ironically, perhaps, *Assumption failures and fix-ups* claims an even larger share of the classics when more advanced books are excluded. These older books also devote considerable space to *Time series*, while *Panel data* get little attention across the board.

A striking feature of Table 2 is how similar the distribution of topic coverage in contemporary market leading econometrics texts is to the distribution in the classics. As in the Stones' Age, well over half of the material in contemporary texts is concerned with *Regression properties*, *Regression inference*, *Functional form*, and

[5] Panhans and Singleton (forthcoming) also document growth in the number of articles using the terms "natural experiment" and "randomized control trial."

Table 2
Topics Coverage in Econometrics Texts, Classic and Contemporary
(page counts as percentage)

	1970s	*1970s excluding more-advanced texts*	*Contemporary*
	(1)	*(2)*	*(3)*
Topic			
Bivariate regression	2.5	3.6	2.8
Regression properties	10.9	11.9	9.9
Regression inference	13.2	13.3	14.6
Multivariate regression	3.7	3.7	6.4
Omitted variables bias	0.6	0.5	1.8
Assumption failures and fix-ups	18.4	22.2	16.0
Functional form	10.2	9.3	15.0
Instrumental variables	7.4	5.1	6.2
Simultaneous equations models	17.5	13.9	3.6
Panel data	2.7	0.7	4.4
Time series	12.3	15.2	15.6
Causal effects	0.7	0.7	3.0
Differences-in-differences	–	–	0.5
Regression discontinuity methods	–	–	0.1
Empirical examples	14.0	15.0	24.4

Note: We classified the content of 12 econometrics texts, six from the 1970s and six currently in wide use (see text for details): Our classic texts are Kmenta (1971), Johnston (1972), Pindyck and Rubinfeld (1976), Gujarati (1978), Intriligator (1978), and Kennedy (1979). Our contemporary texts are Kennedy (2008), Gujarati and Porter (2009), Stock and Watson (2015), Wooldridge (2016), Dougherty (2016), and Studenmund (2017). We report percentages of page counts by topic. All topics sum to 100 percent. Empirical examples are as a percentage of the whole book. Column 2 excludes Kmenta (1971), Johnston (1972), and Intriligator (1978), the more advanced classic econometrics texts. Dashes indicate no coverage.

Assumption failures and fix-ups. The clearest change across book generations is the reduced space allocated to *Simultaneous equations models.* This presumably reflects declining use of an orthodox multi-equation framework, especially in macroeconomics. The reduced coverage of *Simultaneous equations* has made space for modest attention to *Panel data* and *Causal effects*, but the biggest single expansion has been in the coverage of *Functional form* (mostly discrete choice and limited dependent variable models).

Some of the volumes on our current book list have been through many editions, with first editions published in the Stones' Age. It's perhaps unsurprising that the topic distribution in Gujarati and Porter (2009) looks a lot like that in Gujarati (1978). But more recent entrants to the textbook market also deviate little from the classic template. On the positive side, recent market entrants are more likely to at least mention modern topics.

The bottom row of Table 2 reveals the moderate use of empirical examples in the Stones' Age: about 15 percent of pages in the classics are devoted to illustrations involving real data. This average conceals a fair bit of variation, ranging from zero (no examples at all) to more than one-third of page space covering applications. Remarkably, the most empirically oriented textbook in our 12-book sample remains Pindyck and Rubinfeld (1976), one of the classics. Although the field has moved to an average empirical content of over 24 percent, no contemporary text on this list quite matches their coverage of examples.[6]

BLUE Turns to Grey: Econometrics Course Coverage

Many econometrics instructors rely heavily on their lecture notes, using textbooks only as a supplement or a source of exercises. We might therefore see more of the modern empirical paradigm in course outlines and reading lists than we see in textbooks. To explore this possibility, we collected syllabuses and lecture schedules for undergraduate econometrics courses from a wide variety of colleges and universities.[7]

Our sampling frame for the syllabus study covers the ten largest campuses in each of eight types of institutions. The eight groups are research universities (very high activity), research universities (high activity), doctoral/research universities, and baccalaureate colleges, with each of these four split into public and private schools. The resulting sample includes diverse institutions like Ohio State University, New York University, Harvard University, East Carolina University, American University, US Military Academy, Texas Christian University, Calvin College, and Hope College. We managed to collect syllabuses from 38 of these 80 schools. Each of the eight types of schools we targeted is represented in the sample, but larger and more prestigious institutions are overrepresented. Most syllabuses are for courses taught since 2014, but the oldest is from 2009. A few schools contribute more than one syllabus, but these are averaged so each school contributes only one observation to our tabulations. The appendix available with this paper at http://e-jep.org lists the 38 schools included in the syllabus dataset.

For each school contributing course information, we recorded whether the topics listed in Table 1 are covered. A subset of schools also provided detailed lecture-by-lecture schedules that show the time devoted to each topic. It's worth noting that the amount of information that can be gleaned from reading lists and course schedules varies across courses. For example, most syllabuses cover material we've classified as *Multivariate regression*, but some don't list *Regression inference* separately, presumably covering inference as part of the regression module without spelling this out on the reading list. As a result, broader topics appear to get more coverage.

With this caveat in mind, the first column of Table 3 suggests a distribution of econometric lecture time that has much in common with the topic distribution in textbooks. In particular, well over half of class time goes to lectures on *Regression*

[6] The average is pulled down by the fact that one book on the list has no empirical content. Our view of how a contemporary undergraduate econometrics text can be structured around empirical examples is reflected in our book, Angrist and Pischke (2015).

[7] Our thanks to Enrico Moretti for suggesting a syllabus inquiry.

Table 3
Course Coverage

	Lecture time (percent)	*Courses covering topic (percent)*
Topic		
Bivariate regression	11.7	100.0
Regression properties	8.7	43.4
Regression inference	12.4	92.1
Multivariate regression	10.5	94.7
Omitted variables bias	1.9	28.5
Assumption failures and fix-ups	20.2	73.7
Functional form	15.7	92.1
Instrumental variables	3.9	51.8
Simultaneous equations models	0.4	19.3
Panel data	3.6	36.8
Time series	5.0	45.6
Causal effects	2.5	25.4
Differences-in-differences	2.0	27.2
Regression discontinuity methods	1.4	16.7
Number of institutions	15	38

Notes: The first column reports the percentage of class time devoted to each topic listed at left for the 15 schools for which we obtained a detailed schedule. This column sums to 100 percent. Column 2 reports the percentage of courses covering particular topics for the 38 schools for which we obtained a reading list.

properties, *Regression inference*, *Assumption failures and fix-ups*, and *Functional form.* Consistent with this distribution, the second column in the table reveals that, except for *Regression properties*, these topics are covered by most reading lists. The *Regression properties* topic is very likely covered under other regression headings.

Also paralleling the textbook material described in Table 2, our tabulation of lecture time shows that just under 6 percent of course schedules is devoted to coverage of topics related to *Causal effects*, *Differences-in-differences*, and *Regression discontinuity methods*. This is only a modest step beyond the modern textbook average of 3.6 percent for this set of topics. Single-equation *Instrumental variables* methods get only 3.9 percent of lecture time, less than we see in the average for textbooks, both old and new.

Always looking on the bright side of life, we happily note that Table 3 shows that over a quarter of our sampled instructors allocate at least some lecture time to *Causal effects* and *Differences-in-differences.* A healthy minority (nearly 17 percent) also find time for at least some discussion of *Regression discontinuity methods.* This suggests that econometric instructors are ahead of the econometrics book market. Many younger instructors will have used modern empirical methods in their PhD work, so they probably want to share this material with their students. Textbook authors are probably older, on average, than instructors, and therefore less likely to have personal experience with tools emphasized by the modern causal agenda.

Out of Time

Undergraduate econometrics instructions is overdue for a paradigm shift in three directions. One is a focus on causal questions and empirical examples, rather than models and math. Another is a revision of the anachronistic classical regression framework away from the multivariate modeling of economic processes and towards controlled statistical comparisons. The third is an emphasis on modern quasi-experimental tools.

We recognize that change is hard. Our own reading lists of a decade or so ago look much like those we've summarized here. But our approach to instruction has evolved as we've confronted the disturbing gap between what we do and what we teach. The econometrics we use in our research is interesting, relevant, and satisfying. Why shouldn't our students get some satisfaction too?

■ *Our thanks to Jasper Clarkberg, Gina Li, Beata Shuster, and Carolyn Stein for expert research assistance, to the editors Mark Gertler, Gordon Hanson, Enrico Moretti, and Timothy Taylor, and to Alberto Abadie, Daron Acemoglu, David Autor, Dan Fetter, Jon Gruber, Bruce Hansen, Derek Neal, Parak Pathak, and Jeffrey Wooldridge for comments.*

References

Angrist, Joshua D. 1998. "Estimating the Labor Market Impact of Voluntary Military Service Using Social Security Data on Military Applicants." *Econometrica* 66(2): 249–88.

Angrist, Joshua D., Pierre Azoulay, Glenn Ellison, Ryan Hill, and Susan Lu. Forthcoming. "Economic Research Evolves: Citations Fields and Styles." *American Economic Review.*

Angrist, Joshua D., and Alan B. Krueger. 1999. "Empirical Strategies in Labor Economics." Chap. 23 in *Handbook of Labor Economics*, vol. 3, edited by Orley Ashenfelter and David Card, 1277–1366. Elsevier.

Angrist, Joshua D., and Jörn-Steffen Pischke. 2009. *Mostly Harmless Econometrics: An Empiricist's Companion.* Princeton University Press.

Angrist, Joshua D., and Jörn-Steffen Pischke. 2010. "The Credibility Revolution in Empirical Economics: How Better Research Design Is Taking the Con out of Econometrics." *Journal of Economic Perspectives* 24(2): 3–30.

Angrist, Joshua D., and Jörn-Steffen Pischke. 2015. *Mastering 'Metrics: The Path from Cause to Effect.* Princeton University Press.

Arcidiacono, Peter, Esteban M. Aucejo, and V. Joseph Hotz. 2016. "University Differences in the Graduation of Minorities in STEM Fields: Evidence from California." *American Economic Review* 106(3): 525–62.

Arellano, Manuel. 1987. "Computing Robust Standard Errors for Within-Groups Estimators." *Oxford Bulletin of Economics and Statistics* 49(4): 431–34.

Ayres, Ian. 2007. *Super Crunchers.* Bantam Books.

Becker, William E., and William H. Greene. 2001. "Teaching Statistics and Econometrics to Undergraduates." *Journal of Economic Perspectives* 15(4): 169–82.

Bertrand, Marianne, Esther Duflo, and Sendhil Mullainathan. 2004. "How Much Should We Trust Differences-in-Differences Estimates?" *Quarterly*

Journal of Economics 119(1): 249–75.

Bowen, Donald E., III, Laurent Frésard, and Jérôme P. Taillard. *Forthcoming.* "What's Your Identification Strategy? Innovation in Corporate Finance Research." *Management Science.*

Brynjolfsson, Erik, and Andrew McAfee. 2011. "The Big Data Boom Is the Innovation Story of Our Time." *The Atlantic,* November 21.

Chamberlain, Gary. 1982. "Multivariate Regression Models for Panel Data." *Journal of Econometrics* 18(1): 5–46.

Christian, Brian. 2012. "The A/B Test: Inside the Technology That's Changing the Rules of Business." *Wired,* April 25.

Dale, Stacy Berg, and Alan B. Krueger. 2002. "Estimating the Payoff to Attending a More Selective College: An Application of Selection on Observables and Unobservables." *Quarterly Journal of Economics* 117(4): 1491–1527.

Dougherty, Christopher. 2016. *Introduction to Econometrics.* 5th edition. Oxford University Press.

Fuchs-Schündeln, Nicola, and Tarek A. Hassan. 2016. "Natural Experiments in Macroeconomics." Chap. 12 in *Handbook of Macroeconomics,* vol. 2, edited by John B. Taylor and Harald Uhlig, 923–1012. Elsevier.

Goldberger, Arthur S. 1991. *A Course in Econometrics.* Harvard University Press.

Gujarati, Damodar. 1978. *Basic Econometrics.* New York: McGraw-Hill.

Gujarati, Damodar N., and Dawn C. Porter. 2009. *Basic Econometrics.* 5th Edition. Boston: McGraw-Hill.

Hamermesh, Daniel S. 2013. "Six Decades of Top Economics Publishing: Who and How?" *Journal of Economic Literature* 51(1): 162–72.

Hayashi, Fumio. 2000. *Econometrics.* Princeton University Press.

Intriligator, Michael D. 1978. *Econometric Models, Techniques, and Applications.* Englewood Cliffs, NJ: Prentice Hall.

Johnston, J. 1972. *Econometric Methods,* 2nd Edition. New York: McGraw-Hill.

Kennedy, Peter. 1979. *A Guide to Econometrics.* Cambridge, MA: The MIT Press.

Kennedy, Peter. 2008. *A Guide to Econometrics.* 6th Edition, Malden, MA: Blackwell Publishing.

Kmenta, Jan. 1971. *Elements of Econometrics.* New York: The Macmillan Company.

Kohavi, Ron. 2015. "Online Controlled Experiments: Lessons from Running A/B/n Tests for 12 Years." *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.* ACM.

Maddala, G. S. 1977. *Econometrics.* McGraw-Hill.

Moulton, Brent R. 1986. "Random Group Effects and the Precision of Regression Estimates." *Journal of Econometrics* 32(3): 385–97.

Newey, Whitney K., and Kenneth D. West. 1987. "A Simple, Positive Semi-Definite, Heteroskedasticity and Autocorrelation Consistent Covariance Matrix." *Econometrica* 55(3): 703–08.

Panhans, Matthew T., and John D. Singleton. Forthcoming. "The Empirical Economist's Toolkit: From Models to Methods." *History of Political Economy.*

Pindyck, Robert S., and Daniel L. Rubinfeld. 1976. *Econometric Models and Economic Forecasts.* New York: McGraw-Hill.

Stock, James H., and Mark M. Watson. 2015. *Introduction to Econometrics.* 3rd Edition. Boston: Pearson.

Studenmund, A. H. 2017. *Using Econometrics: A Practical Guide.* 7th Edition, Boston: Pearson.

Summers, Anita A., and Barbara L. Wolfe. 1977. "Do Schools Make a Difference?" *American Economic Review* 67(4): 639–52.

White, Halbert. 1980a. "Using Least Squares to Approximate Unknown Regression Functions." *International Economic Review* 21(1): 149–70.

White, Halbert. 1980b. "A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity." *Econometrica* 48(4): 817–38.

Wooldridge, Jeffrey M. 2016. *Introductory Econometrics: A Modern Approach.* 6th edition. Boston: Cengage Learning.

Yitzhaki, Shlomo. 1996. "On Using Linear Regressions in Welfare Economics." *Journal of Business & Economic Statistics* 14(4): 478–86.

Journal of Economic Perspectives—Volume 31, Number 2—Spring 2017—Pages 145–164

Underestimating the Real Growth of GDP, Personal Income, and Productivity

Martin Feldstein

Economists have long recognized that changes in the quality of existing goods and services, along with the introduction of new goods and services, can raise grave difficulties in measuring changes in the real output of the economy. Prominent economists have led and served on government commissions to analyze and report on the subject, including the Stigler Commission in 1961, the Boskin Commission in 1996, discussed in a symposium in the Winter 1998 issue of this journal, and the Schultze Commission in 2002, discussed in a symposium in the Winter 2003 issue of this journal (Stigler 1961; Boskin et al. 1996; National Research Council 2002). But despite the attention to this subject in the professional literature, there remains insufficient understanding of just how imperfect the existing official estimates actually are.

After studying the methods used by the US government statistical agencies as well as the extensive previous academic literature on this subject, I have concluded that, despite the various improvements to statistical methods that have been made through the years, the official data understate the changes of real output and productivity. The measurement problem has become increasingly difficult with the rising share of services that has grown from about 50 percent of private sector GDP in 1950 to about 70 percent of private GDP now. The official measures provide at best a lower bound on the true real growth rate with no indication of the size of the underestimation. Thus, Coyle (2014, p. 125) concludes her useful history of GDP

■ *Martin Feldstein is George F. Baker Professor of Economics, Harvard University, Cambridge, Massachusetts. His email address is mfeldstein39@gmail.com.*

† For supplementary materials such as appendices, datasets, and author disclosure statements, see the article page at
https://doi.org/10.1257/jep.31.2.145 doi=10.1257/jep.31.2.145

by saying: "Gross domestic product is a measure of the economy best suited to an earlier era."

In considering these issues, I have been struck by the difference between the official statistics about economic growth and how people judge whether their own economic condition has improved. The official figures tell us that real GDP per capita grew at an average rate of just 1.4 percent during the past 20 years. It is common to read in the press that because of this slow overall growth and changes in the distribution of income, the real income of the median household did not rise at all between 1995 and 2013 (for example, in the Council of Economic Advisers' 2015 *Economic Report of the President*, p. 30). When polls ask how the economy is doing, a majority of respondents say the country is doing badly; for example, 57 percent of respondents to a CNN–ORC poll in January 2016 said that the country is "doing poorly" (as reported in Long 2016) and in a Gallup poll in October 2016, 29 percent of respondents said the US economy is "poor" while only 29 percent said it was good or excellent (as reported in Dugan 2016). But in a Federal Reserve (2014) survey of household attitudes, two-thirds of households reported that they were doing as well or better than they had been five years earlier and that they were either "living comfortably" or "doing OK."

The contrast is revealing. People know their personal experience directly, but they depend on news media, politicians, and official statistics to judge how the economy as a whole is doing. And while the government statisticians are careful to say that GDP doesn't measure how well we are doing, there is a strong temptation on the part of the press, the politicians, and the public to think that it measures changes in the real standard of living. In this way, when the official statistics on economic growth understate real economic growth, it reduces public faith in the political and economic system. For example, the low measured growth of incomes probably exacerbates concerns about mobility, with people worrying that they and their children are "stuck" at low income levels: in a CNN/ORC poll, 56 percent of respondents said they think their children will be worse off than they are (as reported in Long 2016), and in a Pew Research Center poll, 60 percent of Americans said their children will be financially worse off than their parents (at http://www.pewglobal.org/database/indicator/74/survey/all/response/Worse+off/). Moreover, I think it creates a pessimism that contributes to political attitudes that are against free trade and critical of our market economy more generally.

I begin this essay by briefly reviewing the age-old question of why national income should not be considered a measure of well-being. I then turn to a description of what the government statisticians actually do in their attempt to measure improvements in the quality of goods and services. Next, I consider the problem of new products and the various attempts by economists to take new products into account in measuring overall price and output changes.

Although the officially measured rates of output growth have slowed substantially in recent years, the problem of understating real economic growth is not a new one.[1] It reflects the enormous difficulty of dealing with quality change and the even greater difficulty of measuring the value created by the introduction of new goods and services. This paper is not about the recent productivity slowdown, but I return to that issue later in this paper and discuss the implications of these measurement issues for the measurement of productivity and the recent slowdown in the rate of productivity growth.

The final section of this paper discusses how the mismeasurement of real output and of prices might be taken into account in considering various questions of economic policy. Fortunately, there are important uses of nominal GDP that do not require conversion to real GDP.

Not Even Measuring Output, and Certainly Not Well-being

There is a long-running debate about the extent to which national income estimates should be designed to measure the well-being of the population or just the output of the economy. But in practice, national income concepts have been intentionally defined in ways that fall far short of measuring even *economic* well-being, let alone the broader well-being of individuals as influenced by matters like the environment and crime.

Even if we focus just on economic output, the concept of national output has been explicitly defined ever since the initial work of Kuznets (1934) and Kuznets, Epstein, and Jenks (1941) to exclude goods and services produced within the home. An earlier National Bureau of Economic Research study by Mitchell, King, and Macaulay (1921) offered a conjectural value of housewives services equal to about 30 percent of their estimate of the more narrowly defined traditional national income. Franzis and Stewart (2011) estimate that household production, under various assumptions, ranges from 31 to 47 percent of money earnings. The official statistics also exclude services that are provided outside the home but not sold. This omission has probably had a larger effect in recent years with the provision of such services as Google and Facebook and the vast expansion of publicly available videos and music, together with written commentary, stories, reports, and information, all of which are now available to web-connected users for essentially zero marginal payment.

Similarly, national income estimates focus on the positive value of the goods and services that households consume, not on the time and effort involved in earning the funds to buy those goods and services. The average workweek has declined but the number of two-earner households has increased. Working conditions have

[1] The vast literature bearing on the measurement of changes in the real output of the economy reaches back to Sidgwick (1883), Marshall (1887), Kuznets (1934), and Kuznets, Epstein, and Jenks (1941) and includes, more recently among others, Griliches (1992), Nordhaus (1997), Hausman (1996, 1999), and Gordon (2016). The NBER Conference on Research in Income and Wealth has focused work on this issue for more than 80 years (as discussed in Hulten 2015).

improved as employment has moved from factories and farms to offices. All of this affects economic well-being, but there is (by agreement) no attempt to take it into account in our measures of national income.

I mention these issues not to criticize the official definition of national income, but to stress that it is intended by design to be a measure of national output, not a measure of well-being. The public clearly wants a description of changes in well-being and inappropriately uses the official measures of real GDP and real personal incomes for that purpose. It might be useful to develop a formal array of well-being indicators and perhaps some summary index. These indicators might include measures of health, air pollution in cities, crime, and other matters that are not measured in the official economic statistics: Coyle (2014, chap. 5) discusses some previous attempts to provide such additional indicators. Alternatively, more attention might be focused on the Federal Reserve's Survey of the Economic Well-Being of US Households and its frequency might be increased from an annual survey to quarterly to increase its public saliency.

However, in this essay I will set aside the issues concerning what economic and noneconomic factors are left out of GDP, and how a broader measure of well-being might be constructed. Instead, I will argue that the official measure of real GDP does not even achieve its stated goal of measuring real national output on its own terms.

Measuring Quality Change

The government's calculation of real GDP growth begins with the estimation of nominal GDP, which is the market value of the millions of goods and services sold in the market to households, firms, governments, and foreign buyers. The government statisticians do a remarkable and prodigious job of collecting and then updating data from a wide array of sources.[2]

But for comparisons between one time period and the next, it is necessary to convert nominal GDP to real GDP. That process requires dividing the rise in nominal quantities into a real component and an inflation component, though the use of an appropriate price index. The overall GDP price deflator uses components based on the Consumer Price Index (CPI) and the Producer Price Index (PPI), requiring estimates done by the Bureau of Labor Statistics of the US Department of Labor and by the Bureau of Economic Analysis of the US Department of Commerce.

For each good and service, there are three possibilities when one compares one year with the next: 1) it is the same good or service with the same quality as in

[2] For a detailed analysis of the sources used to estimate these sales/purchases, see "Updated Summary of NIPA Methodologies" (Bureau of Economic Analysis 2015). Boskin (2000) shows that these estimates are subject to substantial revisions, with nearly all revisions from 1959 to 1998 in the upward direction, and some of those revisions being quite large. In this journal, Landefeld, Seskin, and Fraumeni (2008) provide a very useful description of how nominal GDP and related measures are estimated from a variety of primary sources.

the previous period; 2) it is essentially the same good, but of a different quality; or 3) it is a wholly new good. Each category receives a different treatment in the official US statistics.

Fortunately, most goods and services fall in the first category of "no (significant) change in quality." For those products, it is possible to collect the number of physical units sold and the total revenue. The percentage increase in revenue in excess of the percentage increase in physical volume is pure inflation, and the rest is the rise in real output. When exactly the same good is not available in the second period, the US Bureau of Labor Statistics tries to find a very similar good that does exist in the two successive periods and compares the revenue growth and physical quantity growth for that good. The BLS calls this procedure the "matched model" method.

Although much of the growth in the real value of economic output reflects substantial quality change and the introduction of wholly new products, the official procedures do not adequately reflect these sources of increased value. For products that experience quality change, the official methods tell us more about the increase in the value of inputs, in other words about the change in the cost of production, and not much about the increased value to the consumer or other ultimate user. This is true for goods as well as for services, although measuring quality improvement for services is even more difficult than it is for goods.

The government statisticians divide the period-to-period increase in total spending on each unit of product into a part due to a pure price increase ("inflation") and a part due to an increase in quality. The part attributed to a quality increase is considered an increase in the quantity of output although, as I will explain, the method used by the BLS means that it is generally a measure of the quantity of inputs.

The Bureau of Labor Statistics is responsible for creating the Consumer Price Index and the Producer Price Indexes, as well as a number of subsidiary indexes for various categories.[3] One main difference between the CPI and the PPI is that many of the PPI indexes are used primarily to deflate the prices of intermediate products, rather than to deflate output for final demand. The Bureau of Economic Analysis uses those price indexes and other data to create measures of real output. These estimates are also used for measuring the output of the nonfarm business sector and are used by the Department of Commerce to calculate the GDP deflator and real GDP. The same underlying data are also used to calculate the Personal Consumption Expenditures price index that the Federal Reserve uses for its price stability target.[4]

[3] For a clear description of the methods of the US Bureau of Labor Statistics, see *BLS Handbook of Methods*. Chapter 14 discusses the PPI indexes (available at http://www.bls.gov/opub/hom/pdf/homch14.pdf), and chapter 17 spells out the CPI indexes (available at http://www.bls.gov/opub/hom/pdf/homch17.pdf).

[4] A list of the price indexes used to create specific output numbers is available at Bureau of Economic Analysis (2015). For details, see also Bureau of Economic Analysis (2014).

The key question is how the Bureau of Labor Statistics estimates the change in price when there is a change in the quality of the good or service. The BLS asks the producer of each good or service whether there has been a change in the product made by that producer. If there has been no change in the product, any change in its price is considered to be pure inflation as called for in the "matched model method."

If a change has occurred, one approach to estimating the quality change is the "hedonic regression" method originally developed by Griliches (1961). The basic idea, which was used extensively for computers, is to regress the prices of computers in year *t* on a variety of the computers' capacity and performance measures. This gives an implicit price for each of these features (if the linearity assumption of the model is correct). Applying these implicit prices to a computer model in year $t + 1$ generates a price that would apply for that computer if the values of the individual features at time *t* had continued to prevail.

For example, a variety of econometric studies showed that the true price of mainframe computers assessed in this way declined at an annual rate of more than 20 percent per year during the period from 1950 to 1980 (Chow 1967; Baily and Gordon 1989; Triplett 1989). For personal computers, Berndt, Griliches, and Rappaport (1995) found a 28 percent annual rate of quality-adjusted price decline during a more recent period. The lack of use of hedonic regressions in these earlier decades may be part of the explanation for Robert Solow's (1987) comment that "you can see the computer age everywhere but in the productivity statistics."

Hedonic regressions are used for a variety of categories in the Consumer Price Index and the Producer Price Index. In the CPI, hedonic regressions are used in categories of goods that account for about one-third of the value in the basket of goods in the Consumer Price Index, including several categories of apparel, appliances, and electronics, but the main effect of hedonic analysis on the price index is in the analysis of housing, which by itself is more than 30 percent of the basket of goods represented in the CPI. The Bureau of Economic Analysis incorporates these estimates, and also uses hedonic price indexes to deflate nominal output, but for only about 20 percent of GDP.

The use of hedonics is no doubt very difficult to apply for many of these products and services for which, unlike computers, there is not a clear list of measured technical product attributes. There is also a problem of assuming that the attributes affect willingness to pay in a linear or log-linear way. According to the government, extensions of hedonics to even more products and services is limited by the lack of detailed data and staff resources required to build and maintain the hedonic models. In this journal, Hausman (2003) discusses the limitations of hedonic pricing.

When a producer indicates that a quality change has occurred and a hedonic regression is not used, the Bureau of Labor Statistics (2014, 2015a) typically uses the "resource cost method of quality adjustment," which is based on information about the cost of production supplied by the producer. If the producer says there has been a change in the product, the BLS asks about the "marginal cost of new input requirements that are directly tied to changes in product quality." The rationale relied on by the BLS for this input cost as a method for defining the "quality adjustment" or, equivalently, the measure of the increased output, is described in Triplett (1983).

When the resource cost method is used, the Bureau of Labor Statistics concludes that there has been a quality improvement if and only if there is such an increase in the cost of making the product or service. The government statisticians then use the marginal cost of the product change, measured as a percentage of the previous cost of the product, to calculate a share of the price rise that is due to a quality improvement and that is therefore deemed to be an increase in the output of the product. The rest is regarded as inflation. The resource cost method can also treat a decline in production cost as evidence of a decline in quality.

This resource cost method of defining an improvement in a product or service is remarkably narrow and misleading. For the very specific case where a quality improvement is exclusively the result of adding an input, it will work. But according to this method, a pure technological innovation that makes the product or service better for the consumer doesn't count as a product improvement unless it involves an increased cost of production! In reality, product improvements generally occur because of new ideas about how to redesign or modify an existing product or service. Those changes need not involve an increased cost of production.

Government services provide an extreme version of treating costs of inputs as equivalent to the value of outputs. Government services are valued in the GDP at their cost, and so there is no possibility of reflecting changes in government productivity or the value created by the introduction of new government services.

Although the "resource cost method" may be the most common approach for quality adjustment, it and the hedonic procedure are not the only ones. The Bureau of Economic Analysis also uses what it calls the "quantity extrapolation method" and the "direct valuation method" for a few types of output. For example, the real quantity of bank services is derived from volume data on consumers' deposits and loans (for discussion, see Bureau of Economic Analysis 2015).

When government statisticians deal with quality change in services, they use a variety of different methods, but none of them attempts to capture changes in the true output of the service. For some services, like legal services provided to households, the Bureau of Labor Statistics creates a price index for a variety of specific services, like writing a will, and uses that price index and total expenditure to calculate the increase in real output.

The official GDP statistics for the healthcare industry, which accounts for more than 17 percent of US GDP, focus on costs of providing various categories of health services but do not seek to capture the effect of the health products and services on the health of the patient. For example, the "output" measure for hospitals recently shifted from a day of in-patient care to an episode of hospital treatment for a particular condition. Changes in the cost-per-episode-of-treatment is the corresponding price for the Producer Price Index, which is then used to deflate expenditure to get a measure of the quantity of output. Triplett (2012, p. 17), a careful analyst of the statistical health debate, concluded that there is a "very large error in measuring output generated in the medical care sector."

More generally, as Triplett and Bosworth (2004) note, the official data imply that productivity in the health industry, as measured by the ratio of output to the number of employee hours involved in production, declined year after year between 1987 and 2001. They conclude (p. 265) that such a decline in true productivity is unlikely, but that officially measured productivity declines because "the traditional price index procedures for handling product and service improvements do not work for most medical improvements." More recent data show that health sector productivity has continued to decline since 2001.

None of these measures of productivity attempt to value the improved patient outcomes. As one concrete example, when Triplett and Bosworth (2004, p. 335) wrote about the remarkable improvement in treating cataracts—from more than a week as an immobilized hospital inpatient to a quick outpatient procedure—they questioned whether accounting for medical improvements like that would cross over the traditional "production boundary in national accounts" and asked whether "the increased value to the patient of improvement in surgery … belongs in national accounts if no additional charges are made."

The Department of Commerce is experimenting with health sector "satellite accounts" that calculate the cost of treating a patient with a particular diagnosis for a calendar year, including the cost of hospital care, physicians, and pharmaceuticals. But these accounts also do not try to capture the value of improved health outcomes. There are some research studies that attempt to measure the effect of a certain treatment on such health outcomes as Quality Adjusted Life Years (QALYs) or Disability Adjusted Life years (DALYs).

For another example of the difficulties of adjusting for quality in a service, consider mutual fund management. The Bureau of Labor Statistics (2015b) has noted a substantial expansion over time in the types of funds that are available (including exchange-traded funds, fund-of-funds, long-short funds, a large number of emerging market funds, and more), but it ignores this increase in diversity of products and focuses only on the measuring output of mutual fund providers based on a percentage of all assets, concluding: "Under the current methodology, no special procedures are necessary for adjusting for the changes in the quality of portfolio management transactions" (p. 13).

To study the growth of output and productivity for individual industries, the Bureau of Labor Statistics sometimes measures real output at the industry level by the quantity of services provided. For passenger air travel, output of the industry is the number of passenger miles and productivity is defined as passenger miles per employee hour. The analysis of output "does not account for changes in service quality such as flight delays and route circuitry …" (Duke and Torres 2005).

From time to time the Bureau of Labor Statistics re-examines its approach to a particular industry. When the productivity program re-examined its measure of the commercial banking industry in 2012, it revised the activities of commercial banks and raised the estimated annual output growth from 1987 to 2010 by 58 percent, from 2.4 percent a year to 3.9 percent a year (Royster 2012, p. 8).

My own judgment is that, for most goods and services, the official estimate of quality change contains very little information about the value of the output to consumers and other final purchasers. As a result, the corresponding official measures of total real output growth are underestimates, and there is a substantial but unknown upward bias in the measure of price inflation. We don't know what the true values are, and we don't know how wide a margin of error there is around the official estimates.

Dealing with New Products

Although the sales of new products become immediately a part of *nominal* GDP, the extent to which they increase the real incomes of consumers is underestimated. Similarly, the effects of new products are not well reflected in the measures of real output and in price indexes. Moreover, the resource cost method and other government procedures for valuing changes in quality do not provide an approach to dealing with the value to consumers of new goods and services.

Instead, new products and services are not even reflected in the price indexes used to calculate real incomes and output until they represent a significant level of expenditures. They are then rotated into the sample of products used for price index calculations, and subsequent changes in their price are taken into account in the usual way. It is only at that secondary stage, sometime long after the new product has been introduced, that it affects officially measured changes in real output.

As an example to clarify how this works in practice, consider statins, the remarkable class of drugs that lowers cholesterol and reduces deaths from heart attacks and strokes. By 2003, statins were the best-selling pharmaceutical product in history and had become part of the basket of goods and services measured for the Consumer Price Index. When patents on early versions of statins then expired and generic forms became available, their prices fell. The Bureau of Labor Statistics recorded those price declines, implying a rise in real incomes. But the official statistics never estimated any value for the improvement in health that came about as a result of the introduction of statins.

To understand the magnitude of the effect of omitting the value of that single healthcare innovation, here is a quick history of the impact of statins. In 1994, researchers published a five-year study of 4,000-plus patients. They found that taking a statin caused a 35 percent reduction in cholesterol and a 42 percent reduction in the probability of dying of a heart attack. It didn't take long for statins to become a best-selling product with dramatic effects on cholesterol and heart attacks. According to the US Department of Health and Human Services (2011, pg. 26, fig. 17), between 1999–2002 and 2005–2008, the percentage of men aged 65–74 taking a statin doubled to about 50 percent. High cholesterol levels declined by more than half among men and women over age 75, and the death rate from heart disease among those over 65 fell by one-third. Grabowski et al. (2012) calculated that the combination of reduced mortality and lower hospital costs associated with heart attacks and strokes in the year 2008 alone was some $400 billion, which was almost 3 percent of GDP in that year. None of this value produced by statins is included in the government's estimate of increased real income or real GDP.

This example of how statins have been treated in the national income statistics is representative of how all new products and services are treated. The value to consumers of a new good or service is ignored when the new product is at first introduced. Its price level becomes part of the Consumer Price Index when spending on that good or service is large enough to warrant inclusion. Subsequent declines in the price of the product are treated as real income gains, while price increases are part of inflationary real income losses. In short, the basic value to the consumer of the new good is completely ignored.

Ignoring what happens at the time of introduction of new products is therefore a serious further source of understating the real growth of output, incomes, and productivity. In addition, new products and services are not only valuable in themselves but are also valued by consumers because they add to the variety of available options. In an economy in which new goods and services are continually created, their omission in the current method of valuing aggregate real output makes the existing measure of real output even more of a continually increasing underestimate of true output. Hulten (2015, p. 2) summarizes decades of research on dealing with new products done by the Conference on Research in Income and Wealth with the conclusion that "the current practice for incorporating new goods are complicated but may miss much of the value of these innovations."

The introduction of new products into the official price indexes has historically also been subject to remarkably long delays. The Boskin Commission (Boskin et al. 1996) noted that at the time of their report in 1996 there were 36 million cellular phones in the United States, but their existence had not yet been included in the Consumer Price Index. The earlier Stigler Commission (Stigler 1961) found that decade-long delays were also noted for things like room air conditioners. Autos were only introduced to the Consumer Price Index in 1940 and refrigerators in 1934. More recently, the Bureau of Labor Statistics has introduced procedures that cause new products to be rotated into the analysis more quickly, but only after they have achieved substantial scale in spending. These delays cause the price index to miss the gains from introducing the product in the first place as well as the declines in prices that often happen early in product cycles.

But these delays in the introduction of new products to the price indexes are not the key problem. Much more important is the fact that the official statistics ignore the very substantial direct benefit to consumers of new products per se, causing an underestimate of the rate of increase in real output and an overestimate of the corresponding rate of increase of the price index.

There is great uncertainty about the size of these potential biases. For example, the Boskin Commission (Boskin et al. 1996) was charged by the US Senate with calculating the bias in the Consumer Price Index that is used for adjusting Social Security for changes in retirees' cost of living. The Commission considered several sources of bias in the existing Consumer Price Index, including the bias caused by changes in quality and by the omission of new products and provided estimates of each type of bias in the CPI (see also the discussion of the report in the Winter 1998 issue of this journal).

But because the Boskin Commission was not able to do new research on the issue of quality change and innovation bias, it drew on existing research and on personal perceptions. For example, for "food and beverage," which accounts for 15 percent of the CPI, the commission members asked themselves how much a consumer would be willing to pay "for the privilege of choosing from the variety of items available in today's supermarket instead of being constrained to the much more limited variety available 30 years ago." They concluded, based on pure introspection, that "a conservative estimate … might be 10 percent for food consumed at home other than produce, 20 percent for produce where the increased variety in winter (as well as summer farmers' markets) has been so notable, and 5 percent for alcoholic beverages …" They used these numbers for 30 years and converted them to annual average rates of change for the 30-year period. This may be plausible, or not, but there is no real basis for believing that any of these estimates is even vaguely accurate.

Housing is the most heavily weighted component of the Consumer Price Index with a weight of nearly one-third. The Boskin Commission (Boskin et al. 1996) concluded that "a conservative estimate is that the total increase in apartment quality per square foot, including the rental value of all appliances, central air conditioning, and improved bathroom plumbing, and other amenities amounted to 10 percent over the past 40 years, or 0.25 percent per year." Maybe that is right, or maybe a better estimate would be 1 percent per year. There is nothing in the commission's report that helps to choose between differences of this magnitude.

In the end, the Boskin Commission concluded that the weighted average of these individual biases implied a total bias from product innovation and quality change in the annual CPI inflation rate for 1996 of 0.6 percentage points. I have no idea how much margin of error should be attached to that estimate. It served to satisfy the background political purpose for the Boskin Commission of providing a politically acceptable basis for reducing the rate of increase of Social Security benefits.

A formal analytic approach to the problem of valuing new products was developed by Hausman (1996, 2003). He showed how the value to consumers of a single new product could be measured by estimating the value of introducing a new brand of breakfast cereal—specifically Apple-Cinnamon Cheerios. His approach, following the theory presented by Hicks (1940), was to estimate the "virtual price," that is the price that would prevail when the good is just introduced at zero quantity. The consumer gains an amount of real income when the good is introduced implied by the decline in its price from the virtual price to the actual market price. He concluded that the Consumer Price Index component for cereals may be overstated by about 20 percent because of its neglect of new cereal brands. The Hausman estimates were controversial, but if the magnitude is even roughly indicative of the overstatement of the Consumer Price Index from a failure to reflect the introduction of new varieties of cereal brands, then surely the overstatement of the Consumer Price Index and the understatement of real income that result from failing to take into account new products like statins and new anti-cancer drugs must be substantially larger.

Broda and Weinstein (2010) and Redding and Weinstein (2016) extend the Hausman (1996) approach and present a new method for valuing new products as well as the value to consumers of changes in product quality. They analyze a very large set of data on bar-coded package goods for which prices and quantities are available over time. By studying these data in the framework of a demand system based on constant-elasticity-of-substitution utility functions, they find that conventional price indexes overstate inflation for this set of goods by as much as 5 percentage points because the conventional measure ignores quality and new goods biases. Of course, this method is limited to goods and services for which the bar-coded price and quantity data are available and requires accepting a specific theoretical demand specification for these products. But as the availability of data on prices and quantity grows, it provides a starting point for improving the overall measurement of consumer prices and the corresponding estimates of real income.

The creation of new products also means an increased variety of choice, a form of quality improvement in itself, as Hausman (1996) noted. The value to consumers of access to an increased variety of options, which allows individuals to make choices that conform to their personal taste, can be substantial. Coyle (2014) noted that in the 30 years after 1970, the number of commonly available television channels rose from five to 185, and the number of soft drink brands climbed from 20 to 87.

The failure to take new products into account in a way that reflects their value to consumers may be an even greater distortion in the estimate of real growth than the failure to reflect changes in the quality of individual goods and services. At present, there is no way to know.

Productivity Change and Its Recent Slowdown

Labor productivity is defined as the ratio of real output to the number of hours worked by all employed persons. The Bureau of Labor Statistics estimates labor productivity for the nonfarm business sector, as well as for some parts of that sector, using output estimates provided by the Bureau of Economic Analysis.[5]

The key problem in measuring labor productivity is in the numerator—that is, in measuring output. The failure to measure quality changes adequately and to incorporate the value of new products means that true output has grown faster than

[5] In contrast, multifactor productivity is the ratio of real output to a combination of labor and capital input services. It is intended to measure the increase in output that is not attributable to either labor inputs or capital inputs. A good deal of research has been devoted to the very difficult problem of measuring the input of capital services and to the correct way to combine labor and capital inputs. Here, I will sidestep these issues by focusing on labor productivity.

measured output and therefore that the pace of productivity growth has been underestimated. This problem is particularly difficult in service industries. Bosworth and Triplett (2000, p. 6; Triplett and Bosworth 2004, p. 331) note that the official data imply that productivity has *declined* in several major service industries—including health care, hotels, education, entertainment, and recreation—and concluded that this apparent decline was "unlikely" and probably reflected measurement problems.

While the understatement of productivity growth is a chronic problem, there has been a sharp decline in the officially measured rate of productivity growth in the last decade. That sharp decline remains a puzzle that is yet to be resolved, as Syverson discusses in this issue. His work, along with papers by Fernald (2014) and Byrne, Fernald, and Reinsdorf (2016) show that the recent productivity slowdown cannot be attributed to the effects of the recession of 2008–2009, to changes in the labor force demographics in recent years, or to the growth of unmeasured internet services. One possible explanation of the recent downturn in productivity growth may be that the unusually rapid *increase* in the productivity growth in the prior few years was an anomaly and the recent decline is just a return to earlier productivity patterns.

A further hypothesis for explaining the recent downturn in productivity growth that has not yet been fully explored involves the mismeasurement of official estimates of output and productivity. Any attempt to explain the recent decline in the estimated productivity growth rate must attempt to understand not just the aggregate behavior for the nonfarm business sector as a whole, but also what happened at the disaggregated level. (Official estimates of productivity by industry, are available from the Bureau of Labor Statistics ("Industry Productivity" 1987–2015), although it should be noted that the overall productivity measure is not calculated by combining the individual industry numbers but is estimated separately based on a measure of real value added.)

The recent decline in the official measure of overall labor productivity growth in the nonfarm business sector reflects an enormous diversity of changes of productivity in specific industry groups. For the nonfarm business sector as a whole, the rate of productivity growth fell from 3.2 percent a year in the decade from 1995 to 2004 to just 1.5 percent in the decade from 2004 to 2013. The decline of 1.7 percentage points in the overall productivity change reflects an enormous range of changes in various industries. Even if attention is limited to the relatively aggregate three-digit level, the official productivity data show that productivity in apparel manufacturing went from annual growth at 1 percent in the earlier decade to an annual productivity decline of 5 percent in the later period, a drop of 6 percentage points. For manufacturing of computers and electronic products, productivity growth fell from a 15 percent annual rate to a 4 percent annual rate, a fall of 11 percentage points. Some industries experienced faster productivity growth, with productivity in the manufacturing of wood products increasing from a 2 percent annual rise in the early period to a 2.4 percent rise in the later period.

The differences are even greater at a more disaggregated level. At the four-digit level, for example, productivity growth increased by 5 percentage points annually for radio and TV broadcasting but declined by 18 percent for semiconductors and electronic components. The deflation of output for disaggregated industries is even harder than for the economy as a whole because nominal outputs must be deflated by quality-adjusted prices for the disaggregated industries (Dennison 1989).

It would be intriguing, although difficult, to explore how or whether productivity differences across industries might be correlated with the problems of dealing with product change and the introduction of new goods and services in those industries.

Using Our Imperfect Data

What can be learned from the imperfect measures of real output and from the corresponding overstatement of price inflation? How should our understanding of the mismeasurement affect the making of monetary and fiscal policies?

Assessing Cyclical Economic Conditions

Consider first the assessment of short-run business cycle conditions. Policymakers and financial markets often focus on short-term fluctuations of real GDP as an indication of the state of the business cycle. Although measuring the size of fluctuations of real GDP is flawed by the difficulty of dealing with new products and quality changes, the official measure of real GDP fluctuations can in principle capture the short-term up or down changes in the pace of economic activity. Of course, it is important to recognize the substantial uncertainty about the estimated short-run fluctuations in GDP and the subsequent revisions.[6]

But it is interesting to note that when the Business Cycle Dating Committee of the National Bureau of Economic Research meets to consider appropriate dates for the start and end of a recession, it places relatively little emphasis on GDP. Contrary to popular belief, the NBER Committee has never used two quarters of decline in real GDP as its definition of a recession. Instead, it has traditionally looked at employment, industrial production, wholesale-retail sales, as well as real income. In recent years, the NBER Committee has also looked at monthly GDP when Macro Advisers began creating monthly estimates of GDP.

All data involve problems of interpretation in judging the state of economic activity, but employment, industrial production, and nominal sales are relatively free from the problem of quality adjustment and price measurement that affect measures of real GDP. Employment data are available monthly with substantial detail based on a large survey of employers. Industrial production is estimated by

[6] The Federal Reserve Banks of New York and Atlanta have recently begun using official data to produce preliminary estimates of changes in real GDP even before the corresponding quarter is over, but with some variability in results. In April 2016, the New York Federal Reserve estimated that real GDP increased by 1.1 percent in the recently completed first quarter of 2016, while the Atlanta Federal Reserve estimated that the increase in the same quarter was only 0.1 percent.

the Federal Reserve based primarily on data on physical production (such as tons of steel and barrels of oil) obtained from trade associations and government agencies, supplemented when necessary with data on production-worker hours and for some high-tech products by using nominal output and a price index (for details, see the Federal Reserve Board data https://www.federalreserve.gov/releases/G17/). These measures of industrial production as well as wholesale-retail sales deal with economic activity without having to impute value in large amounts, as must be done for the services of owner-occupied homes that are involved in the estimate of GDP.

Assessing Longer-Term Growth and Inflation

For the longer term, the official measures of changes in real output are misleading because they essentially ignore the value created by the introduction of new goods and services and underestimate changes in the quality of these products. It follows therefore that "true" real output is growing faster than the official estimates imply and that the corresponding "true" GDP price index is rising more slowly than the official one—or is actually declining.

The economics profession should educate the general public and the policy officials that "true" real incomes are rising faster than the official data imply. We can reassure people that it is very unlikely that the real incomes of future generations will be lower than real incomes today. Even if the future will not see the "epochal innovations" of the type that Kuznets (1971) referred to or such fundamental changes as electricity and indoor plumbing that caused jumps in living standards (as emphasized by Gordon 2016), current and future generations can continue to experience rising real incomes due to technological changes, improvements in education, and increases in healthcare technology.

One can only speculate about whether the bias in the officially measured pace of real output change is greater now than in the past. One reason to think that the gap between true output growth and measured growth is greater now than in the past is that services now represent about 70 percent of private value added, up from about 50 percent of private value added back in 1950, and the degree of underestimation of quality change and product innovation may be greater for services. Within services, health occupies a larger share of output—and quality improvements there may be greater than in other parts of the service sector. The internet and services through the internet have become much more important, and are also harder to measure.

Poverty and Distribution

Trends in the overestimation of inflation and therefore in the underestimation of real incomes may vary among demographic groups and income groups because of differences in the mix of goods and services consumed by these different groups. For example, are the goods and services bought by older people improving relatively faster than the goods and services bought by younger households? Health care is an obvious example, although most of the consumption of health care by the elderly is financed by government transfers.

Implications for Fiscal and Monetary Policy

Policy issues that depend on nominal measures of output are unaffected by the problems discussed in this essay. The most obvious of these is the ratio of debt to GDP, since both the numerator and the denominator are nominal values. Similarly, the rate of change of the debt-to-GDP ratio depends only on the nominal value of the annual deficit and the annual rate of nominal GDP growth. If the debt-to-GDP ratio is not on an explosive path, its long-run equilibrium value is equal to the annual nominal deficit ratio divided by the rate of nominal GDP growth.

The evidence that the true inflation rate is less than the measured inflation rate may imply that the true inflation rate is now less than zero. Fortunately, this does not imply that the US economy is experiencing the traditional problem of debt deflation (Fisher 1933) that occurs when a declining price level reduces aggregate demand by increasing the value of household debt relative to current incomes. The traditional problem of debt deflation does not arise under current conditions because the nominal value of wage income is not declining and the real monthly wage is rising more rapidly.

Overestimating the true rate of inflation does imply that the real rate of interest is higher than the conventionally measured rate. If households recognize that their dollars will buy relatively more in the future, this could alter the household saving rate—either increasing saving in response to the greater reward for saving or decreasing saving because a given volume of assets will buy more in the future, depending on whether substitution or income effects dominate. Because many factors affect the household saving rate, it is not clear which of these effects now dominates.

Uncertainty about the true rate of inflation should affect the optimal monetary policy. There seems little point in having a precise inflation target when the true rate of inflation is measured with a great deal of uncertainty. The goal of price stability also takes on a new meaning if true inflation is substantially negative while measured inflation is low but positive. Would it be better to have a target range for measured inflation as the Federal Reserve does now? Or to have a target range for measured inflation that is higher and further from the zero bound, thus leaving more room for larger changes in nominal interest rates while recognizing that the actual inflation rate is lower than the officially measured one? Or to restate the inflation goal of monetary policy as reacting when there is a rapid movement in measured inflation either up or down?

The underestimation of real growth has affected Federal Reserve decision-making in the past. Back in 1996, Fed chairman Alan Greenspan persuaded members of the Federal Open Market Committee that the official data underestimated productivity growth, so that maintaining strong demand would not cause a rise in inflation and there was no reason to raise interest rates (Mallaby 2016). In the last few years, the perception of slow real growth is often mentioned in support of a Federal Reserve policy of exceptionally low interest rates, but if real growth rates are actually higher (or if real growth rates have not dipped as much as the official statistics seem to show), then the Fed's policy of ultra-low interest rates has been providing little gain while contributing to certain risks of potential financial instability.

A great deal of effort and talent has been applied over past decades to the measurement of real income and inflation. These problems are extremely difficult. In my judgement, they are far from being resolved, and as a result, substantial errors of unknown size remain in our ability to measure both real output and inflation. It is important for economists to recognize the limits of our knowledge and to adjust public statements and policies to what we can know.

■ *I am grateful for extensive help with this paper to Anna Stansbury and for comments on earlier drafts from Katherine Abraham, Graham Allison, Michael Boskin, Erica Groshen, Jim Stock, David Weinstein, members of the staffs of the Council of Economic Advisers, the Bureau of Labor Statistics (especially David Friedman), and the Bureau of Economic Analysis, as well as participants at a meeting of the Group of Thirty and at a Harvard seminar.*

References

Baily, Martin, and Robert J. Gordon. 1989. "Measurement Issues, the Productivity Slowdown and the Explosion of Computer Power." CEPR Discussion Paper 305.

Berndt, Ernst R., Zvi Griliches, and Neal J. Rappaport. 1995. "Econometric Estimates of Price Indexes for Personal Computers in the 1990s." *Journal of Econometrics* 68(1): 243–68.

Boskin, Michael J., Ellen R. Dulberger, Zvi Griliches, Robert J. Gordon, and Dale Jorgensen. 1996. *Toward a More Accurate Measure of the Cost of Living*. Final report to the Senate Finance Committee for the Advisory Commission to Study the Consumer Price Index.

Boskin, Michael J. 2000. "Economic Measurement: Progress and Challenges." *American Economic Review* 90(2): 247–52.

Bosworth, Barry, and Jack E. Triplett. 2000. "Numbers Matter: The US Statistical System and a Rapidly Changing Economy." *Brookings Policy Brief 63*.

Broda, Christian, and David E. Weinstein. 2010. "Product Creation and Destruction: Evidence and Price Implications." *American Economic Review* 100(3): 691–723.

Bureau of Economic Analysis. 2014. *Concepts and Methods of the U.S. National Income and Product Accounts*. February. https://www.bea.gov/national/pdf/allchapters.pdf.

Bureau of Economic Analysis. 2015. "Updated Summary of NIPA Methodologies." *Survey of Current Business*, November, 95(11).

Bureau of Labor Statistics, US Department of Labor. 1987–2015. "Industry Productivity." Database, available at https://www.bls.gov/lpc/data.htm (accessed April 3, 2016).

Bureau of Labor Statistics, US Department of Labor. 1987–2015. "Major Sector Productivity and Costs." Database, available at https://www.bls.gov/lpc/data.htm (accessed April 3, 2016).

Bureau of Labor Statistics. No date. "Producer Prices." Chap. 14 in *Handbook of Methods*. http://www.bls.gov/opub/hom/pdf/homch14.pdf.

Bureau of Labor Statistics. No date. "The Consumer Price Index." Chap. 17 In *Handbook of Methods*. http://www.bls.gov/opub/hom/pdf/homch17.pdf.

Bureau of Labor Statistics. 2014. "Quality Adjustment in the Producer Price Index." Last modified August 2014. http://www.bls.gov/ppi/qualityadjust.pdf.

Bureau of Labor Statistics. 2015a. "Hedonic Models in the Producer Price Index (PPI)." Last modified October 21, 2015. http://www.bls.gov/ppi/ppicomqa.htm.

Bureau of Labor Statistics. 2015b. "Synopsis: NAICS 523920. Portfolio Management." Producer Price Index Program. Revised March 2, 2015.

Byrne, David M., John G. Fernald, and Marshall B. Reinsdorf. 2016. "Does the United States have a Productivity Slowdown or a Measurement Problem?" Brookings Papers on Economic Activity Conference Draft, March 10–11.

Chow, Gregory C. 1967. "Technological Change and the Demand for Computers." *American Economic Review* 57(5): 1117–30.

Council of Economic Advisors, The. 2015. *Economic Report of the President.* Transmitted to the Congress, February 2015. Washington, DC: US Government Printing Office.

Coyle, Diane. 2014. *GDP: A Brief But Affectionate History*. Princeton University Press.

Dennison, Edward F. 1989. *Estimates of Productivity Change by Industry: An Evaluation and an Alternative.* Washington, DC: Brookings Institution Press.

Dugan, Andrew. 2016. "U.S. Economic Confidence Index Flat at -10." Gallup, October 18. http://www.gallup.com/poll/196526/economic-confidence-index-flat.aspx.

Duke, John, and Victor Torres. 2005. "Multifactor Productivity Change in the Air Transportation Industry." *Bureau of Labor Statistics Monthly Labor Review,* March, pp. 32–45.

Federal Reserve, Board of Governors. 2014. *Report on the Economic Well-Being of U.S. Households in 2013.* July.

Federal Reserve, Board of Governors. "Industrial Production and Capacity Utilization – G.17." http://www.federalreserve.gov/releases/g17/current/.

Federal Reserve Bank of Atlanta. No date. "GDP Now". https://www.frbatlanta.org/cqer/research/gdpnow.aspx?panel=1 (accessed April 2016).

Federal Reserve Bank of New York. 2016. "Nowcasting Report: April 8, 2016." https://www.newyorkfed.org/medialibrary/media/research/policy/nowcast/nowcast_2016_0408.pdf?la=en.

Fernald, John G. 2014. "Productivity and Potential Output Before, During and After the Great Recession." NBER Working Paper 20248.

Fisher, Irving. 1933. "The Debt-Deflation Theory of Great Depressions." *Econometrica* 1(4): 337–57.

Franzis, Henry, and Jay Stewart. 2011. "How Does Nonmarket Production Affect Measured Earnings Inequality." *Journal of Population Economics* 24(1): 3–22.

Gordon, Robert J. 2016. *The Rise and Fall of American Growth: The US Standard of Living since the Civil War.* Princeton University Press.

Grabowski, David, Darius N. Lakdawalla, Dana P. Goldman, Michael Eber, Larry Z. Liu, Tamer Abdelgawad, Andreas Kuznik, Michael E. Chernew, and Tomas Philipson. 2012. "The Large Social Value Resulting from Use of Statins Warrants Steps to Improve Adherence and Broader Treatment." *Health Affairs* 31(10): 2276–85.

Griliches, Zvi. 1961. "Hedonic Price Indexes for Automobiles: An Econometric Analysis of Quality Change." In *The Price Statistics of the Federal Government: Review, Appraisal, and Recommendations,* General Series 73, pp. 173–96. A Report to the Offices of Statistical Standards Bureau of the Budget prepared by the Price Statistics Review Committee of the National Bureau of Economic Research. New York: National Bureau of Economic Research.

Griliches, Zvi, ed. 1992. *Output Measurement in the Service Sectors.* National Bureau of Economic Research Studies in Income and Wealth. University of Chicago Press.

Hausman, Jerry A. 1996. "Valuation of New Goods under Perfect and Imperfect Competition." In *The Economics of New Goods,* edited by Timothy F. Bresnahan and Robert J. Gordon, 207–248. University of Chicago Press.

Hausman, Jerry A. 1999. "Cellular Telephone, New Products, and the CPI." *Journal of Business and Economic Statistics* 17(2): 188–94.

Hausman, Jerry A. 2003. "Sources of Bias and Solutions to Bias in the Consumer Price Index." *Journal of Economic Perspectives* 17(1): 23–44.

Hicks, John R. 1940. "The Valuation of the Social Income." *Economica* 7(26): 105–24.

Hulten, Charles R. 2015. "Measuring the Economy of the 21st Century." *NBER Reporter* no. 4, pp. 1–7.

Kuznets, Simon. 1934. "National Income, 1929–1932." Bulletin 49, National Bureau of Economic Research. July 7.

Kuznets, Simon, Lillian Epstein, and Elizabeth Jenks. 1941. *National Income and Its Composition, 1919–1938,* vol. 1. National Bureau of Economic Research.

Kuznets, Simon. 1971. "Modern Economic Growth: Findings and Reflections". Sveriges Riksbank Prize in Economic Sciences in Memory of Alfred Nobel, Prize Lecture, December 11, 1971.

Landefeld, J. Steven, Eugene P. Seskin, and Barbara M. Fraumeni. 2008. "Taking the Pulse of the Economy: Measuring GDP." *Journal of Economic Perspectives* 22(2): 193–216.

Long, Heather. 2016. "56% of Americans Think Their Kids Will Be Worse Off." *CNN Money,* January 28. http://money.cnn.com/2016/01/28/news/economy/donald-trump-bernie-sanders-us-economy.

Mallaby, Sebastian. 2016. *The Man Who Knew: The Life and Times of Alan Greenspan.* Bloomsbury.

Marshall, Alfred. 1887 [1925]. "Remedies for Fluctuations of General Prices" (1887) in *Memorials of Alfred Marshall,* edited by A. C. Pigou (1925). London: MacMillan.

Mitchell, Wesley Clair, Wilford Isbell King, and Frederick R. Macaulay. 1921. *Income in the United States: Its Amount and Distribution, 1909–1919,* Vol. 1: *Summary*. New York: National Bureau of Economic Research.

National Research Council. 2002. *At What Price? Conceptualizing and Measuring Cost-of-Living and Price Indexes,* edited by Charles Schultze and Christopher Mackie. Washington, DC: National Academies Press

Nordhaus, William D. 1997. "Traditional Productivity Estimates Are Asleep at the (Technological) Switch." *Economic Journal* 107(444): 1548–59.

Redding, Stephen J., and David E. Weinstein. 2016. "A Unified Approach to Estimating Demand and Welfare." http://www.columbia.edu/~dew35/Papers/Estimating-Demand-and-Welfare.pdf.

Royster, Sara E. 2012. "Improved Measures of Commercial Banking Output and Productivity." *Monthly Labor Review,* July, 135(7): 3–12.

Sidgwick, Henry. 1883. *The Principles of Political Economy*. Macmillan and Co.

Solow, Robert. 1987. "We'd Better Watch Out." *New York Times Book Review,* July 12. p. 36.

Stigler, George, ed. 1961. "The Price Statistics of the Federal Government: Review, Appraisal, and Recommendations." Report to the Office of Statistical Standards, Bureau of the Budget. New York: National Bureau of Economic Research.

Triplett, Jack E. 1983. "Concepts of Quality in Input and Output Price Measures: A Resolution of the User-Value Resource-Cost Debate." Chap. 5 in *The U.S. National Income and Product Accounts: Selected Topics*, edited by Murray F. Foss. University of Chicago Press.

Triplett, Jack E. 1989. "Price and Technological Change in a Capital Good: A Survey of Research on Computers." In *Technology and Capital Formation,* edited by Dale W. Jorgenson and Ralph Landau, 127–213. Cambridge, MA: MIT Press.

Triplett, Jack E. 2012. "Health System Productivity." Chap. 30 in *The Oxford Handbook of Health Economics*, edited by Sherry Glied and Peter C. Smith. Oxford University Press.

Triplett, Jack E., and Barry P. Bosworth. 2004. *Productivity in the US Services Sector: New Sources of Economic Growth*. Washington, DC: Brookings Institution Press.

US Department of Health and Human Services. 2011. "Health, United States, 2010: With Special Feature on Death and Dying." https://www.cdc.gov/nchs/data/hus/hus10.pdf.

Journal of Economic Perspectives—Volume 31, Number 2—Spring 2017—Pages 165–186

Challenges to Mismeasurement Explanations for the US Productivity Slowdown

Chad Syverson

The flow and ebb of US productivity growth since World War II is commonly divided into four periods: 1947–1973, 1974–1994, 1995–2004, and 2004–2015. After labor productivity growth averaged 2.7 percent per year from 1947–1973, it fell in a much-studied-but-still-debated slowdown to 1.5 percent per year over 1974–1994. Another fast/slow cycle has followed. Productivity growth rose to a trajectory of 2.8 percent average annual growth sustained over 1995–2004. But since then, the US economy has been experiencing a slowdown in measured labor productivity growth. From 2005 through 2015, labor productivity growth has averaged 1.3 percent per year (as measured by the nonfarm private business labor productivity series compiled by the US Bureau of Labor Statistics).

This slowdown is statistically and economically significant. A t-test comparing average quarterly labor productivity growth rates over 1995–2004 to those for 2005–2015 rejects equality with a p-value of 0.008. If the annualized 1.5 percentage point drop in labor productivity growth were to be sustained for 25 years, it would compound to an almost 50 percent difference in income per capita.

The productivity slowdown does not appear to be due to cyclical phenomena. Fernald (2014a) shows that the slowdown started before the onset of the Great Recession and is not tied to "bubble economy" phenomena in housing or finance. This work, along with the analysis in Byrne, Oliner, and Sichel (2013), ties the slowdown to a reversal of the productivity accelerations in the manufacturing and utilization

■ *Chad Syverson is J. Baum Harris Professor of Economics, University of Chicago Booth School of Business, Chicago, Illinois, and Research Associate, National Bureau of Economic Research, Cambridge, Massachusetts. His email is chad.syverson@chicagobooth.edu.*

† For supplementary materials such as appendices, datasets, and author disclosure statements, see the article page at
https://doi.org/10.1257/jep.31.2.165 doi=10.1257/jep.31.2.165

of information and communication technologies that drove the more rapid pace of productivity from 1995–2004. While one cannot rule out persistent, less-direct channels through which the Great Recession might have long-lived influences on productivity growth, it is clear that measured labor productivity in the United States has not awakened from its slowdown as the Great Recession recedes.

The debate about the causes of the productivity slowdown is ongoing. Gordon (2016) points to multiple possible explanations and ties the current slowdown to the one in 1974–1994, viewing the 1995–2004 acceleration as a one-off aberration. Cowen (2011) shares these views and enumerates multiple reasons why innovation—at least the kind that leads to changes in measured productivity and income—may slow. Tarullo (2014) suggests that the slowdown in US business dynamism documented by Decker, Haltiwanger, Jarmin, and Miranda (2014) and Davis and Haltiwanger (2014) may have a role. Some have argued that there are reasons to be optimistic that the slowdown may reverse itself. Baily, Manyika, and Gupta (2013) point to potential innovation opportunities in multiple sectors. Syverson (2013) notes that the productivity growth from electrification and the internal combustion engine—a prior diffusion of a general purpose technology—came in multiple waves, implying that the 1995–2004 acceleration need not be a one-time event.

However, these arguments all accept that the measured decline in productivity growth is meaningful. A separate set of explanations for the slowdown in measured productivity put forward by several parties is that it is substantially illusory (for example, Brynjolfsson and McAfee, 2011, 2014; Mokyr 2014; Alloway 2015; Byrne, Oliner, and Sichel 2015; Feldstein 2015; Hatzius and Dawsey 2015; Smith 2015). The theme of these arguments is that true productivity growth since 2004 has not slowed as much as official statistics may suggest—and perhaps productivity growth has even accelerated—but that due to measurement problems, the new and better products of the past decade are not being captured in official productivity metrics.

There is a prima facie case for this assertion, which for brevity I refer to as the "mismeasurement hypothesis." Many of the fastest-diffusing technologies since 2004—like smartphones, online social networks, and downloadable media—involve consumption of products that are time-intensive but do not impose a large direct monetary cost on consumers. If one considers the total expenditure on such products to be both the monetary price *and* the value of time spent consuming them, a revealed preference argument would suggest they deliver substantial utility (Becker 1965). At the same time, the fact that these new products are not particularly expensive (at least relative to consumers' supposed interest in them) could result in a relatively modest portion of their delivered consumption benefit to be reflected in GDP.

This mismeasurement hypothesis could take one of two related forms. One possibility is that a smaller share of the utility that these products provide is embodied in their prices than was the case for products made before 2004. If this were true, measured output growth would slow even as growth of total surplus continued apace. The second possibility is that if the price deflators of these new technology products are rising too fast (or falling too slowly) relative to their pre-2004 changes, the result would be that quantity growth as backed out from nominal sales is understated.[1]

In this study, I explore the quantitative plausibility of the mismeasurement hypothesis. One fact dominates the discussion: had the measured productivity slowdown not happened, measured GDP in 2015 would have been, conservatively, $3 trillion (17 percent) higher than it was. This is $9,300 for every person or $24,100 for every household in the United States. For the mismeasurement hypothesis to explain the productivity slowdown, the losses in measured incremental gains from the new technologies would need to be at or around this level. Thus, to explain even a substantial fraction of the productivity slowdown, current GDP measures must be missing hundreds of billions of dollars of incremental output (and moreover with no accompanying employment growth).

I start with a computation of the missing output lost to the productivity slowdown. I then turn to discussion of four patterns in the data, each looking at the mismeasurement hypothesis from different directions, which pose challenges for the hypothesis.

First, the productivity slowdown is not unique to the United States. It has occurred with similar timing across at least two dozen other advanced economies. However, the magnitude of the productivity slowdown across countries (of which there is nontrivial variation) is unrelated to the relative size of information and communication technologies (ICT) in the country's economy, whether this "ICT intensity" is measured in consumption or production terms.

Second, a research literature has attempted to measure the consumer surplus of the internet. These efforts are based on the notion that many of the newer technologies that could create large surplus with little revenue require internet access, which makes purchase and use of internet access a metric for the gains from such technologies. However, most of the estimates of the value of internet-linked technologies are at least an order of magnitude smaller than the trillions of dollars of measured output lost to the productivity slowdown. As I will discuss, even the largest estimate, which explicitly accounts for the time people spend online and is computed with very generous assumptions about the value of that time, totals only about one-third of the missing output.

Third, if the mismeasurement hypothesis were to account entirely (or almost so) for the productivity slowdown, and if the source of this mismeasurement is predominantly in certain industries that make and service digital and information

[1] These issues have arisen before. Diewert and Fox (1999) discuss related productivity measurement problems in the context of an earlier slowdown, arguing that there were several plausible sources of mismeasurement. The price-deflator-based interpretation of the measurement problem evokes the Boskin Commission report (US Congress 1996), which argued that the Consumer Price Index methodology at the time overstated inflation and therefore understated growth. Many of the commission's suggested changes, including those specifically aimed at better measurement of new products and technologies, were implemented before 2004 (Klenow 2003). The issues raised by the Boskin Commission report were discussed in a six-paper symposium on "Measuring the CPI" in the Winter 1998 issue of this journal, and a follow-up report by the National Academy of Sciences was discussed in a three-paper symposium on the "Consumer Price Index" in the Winter 2003 issue.

and communication technologies, then the implied change in real revenues of these industries would be five times their measured revenue change. Incremental real value added would have been six times the observed change, and true labor productivity in these industries would have risen 363 percent over 11 years.

Fourth, gross domestic income (GDI) and gross domestic product (GDP) are conceptually equivalent, but because they are computed with different source data, they are not actually equal. Since 2004, GDI has outstripped GDP by an average of 0.4 percent of GDP per year. This pattern is consistent with workers being paid to produce goods that are being given away for free or sold at steep discounts, which is consistent with the mechanism behind the mismeasurement hypothesis. However, I show that GDI began to be larger than GDP in 1998—several years before the productivity slowdown and, indeed, in the midst of a well-documented productivity acceleration. Additionally, a breakdown of GDI by income type shows that GDI growth over the period has been driven by historically high capital income (like corporate profits), while labor income has actually fallen. This is opposite the implication of a "workers paid to make products sold free" story.

In isolation, none of these four patterns are dispositive. But taken together, they challenge the ability of the mismeasurement hypothesis to explain a substantial part of the productivity slowdown.

Calculating the Missing Output

Whether the mismeasurement of productivity hypothesis is presumed to act through output gains disproportionately flowing into consumer surplus rather than GDP or through incorrect price deflators, the implication is the same: US consumers benefited from this missing output, but it just was not reflected in measured GDP. Any evaluation of the hypothesis needs to put estimates of productivity mismeasurement in the context of measures of this hypothetically missing output.

I first compute the implied lost output due to the productivity slowdown. Using quarterly labor productivity data from the US Bureau of Labor Statistics for the entire nonfarm business sector, I calculate average quarterly productivity growth over four post-WWII periods: 1947–1973, 1974–1994, 1995–2004, and 2005–2015 (period averages are inclusive of endpoint years). Past research has shown that average productivity growth has inflection points at or around the transitions between these periods, and work on both the most recent and prior productivity slowdowns has used these periods (for example, Byrne, Oliner, and Sichel 2013). Table 1 shows average productivity growth rates along with their annualized values for each period. As is clear in the table, measured labor productivity growth after 2004 fell by more than half from its 1995–2004 average.[2]

[2] Related productivity measures testify to the spread and depth of the slowdown. Sector-specific labor productivity growth slowed over the same period for each of the six two-digit NAICS industries with available data (mining, utilities, manufacturing, wholesale, retail, and accommodation and food services).

Table 1
Average Quarterly Labor Productivity (LP) Growth by Period

Period	*Average quarterly LP growth (%)*	*Annualized LP growth (%)*
1947–1973	0.681	2.73
1974–1994	0.386	1.54
1995–2004	0.712	2.85
2005–2015	0.317	1.27

Note: These values are taken from the Bureau of Labor Statistics nonfarm private industry labor productivity growth series. Annualized growth values are simply four times quarterly growth.

Labor productivity is defined as the ratio of real output to labor inputs, so it is straightforward to compute what counterfactual output would have been after 2004 had productivity growth not slowed. The drop in average quarterly labor productivity growth between 1995–2004 and 2005–2015 is 0.395 percentage points (= 0.712 – 0.317). Thus, counterfactual output in 2015 would thus have been 19 percent higher ($1.00395^{44} = 1.189$) than observed output in that period. Note that this exercise does not change labor inputs. Counterfactual output still reflects the observed movements in labor inputs over the period, like the considerable decline during the Great Recession. This exercise therefore does not assume away the employment downturn of the slowdown period.[3]

Nominal GDP in 2015 was $18.037 trillion. If I apply the counterfactual extra productivity growth of 19 percent to this value, the amount of output “lost” due to the productivity slowdown is $3.43 trillion per year.[4]

Notably, these sectors might vary in their inherent “measurability.” Total factor productivity growth also slowed. The Bureau of Labor Statistics measure of multifactor productivity fell from 1.4 percent per year during 1995–2004 to 0.5 percent per year over 2005–2015. The utilization-corrected total factor productivity measures of Fernald (2014b) also saw similar decelerations, by 2.5 percent per year in the equipment and consumer durables producing sectors and 1.1 percent per year for makers of other outputs.

[3] An implication of the mismeasurement hypothesis is that the reported output deflator does not reflect true price changes and should have grown more slowly than what was measured. It is therefore instructive to compare the average growth rates of the implicit price deflator for the Bureau of Labor Statistics productivity series in the 1995–2004 and 2005–2015 periods. The deflator grew an average of 0.36 percent per quarter from 1995–2004 and 0.41 percent per quarter from 2005–2015. Compounded over the 44 quarters of the latter period, the deflator grew a cumulative 2.3 percent more than had it remained at its earlier trajectory. To the extent that this acceleration might reflect real output mismeasurement (and the fact that it did accelerate does not imply that it shouldn’t have), it would only explain about one-eighth of the measured slowdown.

[4] The calculations here and throughout this paper use 2015 as an endpoint because several of the data sources I use extend only through that year. The implied “lost” output would be even larger than the reported values if I used the labor productivity data through 2016 (the latest available numbers as of this writing). This is for two reasons. First, average labor productivity growth during 2016 was even slower than the 2005–2015 average. Second, the slowdown would be compounded over another year of GDP growth. Conducting similar calculations to those above using the 2016 data imply values of lost output that are 14 percent larger than those reported here.

However, it is not immediately obvious if GDP is the correct base to which to apply the counterfactual growth rate. The Bureau of Labor Statistics labor productivity series that I use here applies to nonfarm business activity, which excludes farming, government, nonprofits, and paid employees of private households. The reason given is that the outputs of these sectors in GDP "are based largely on the incomes of input factors. In other words, the measure is constructed by making an implicit assumption of negligible productivity change" (http://www.bls.gov/lpc/faqs.htm). The value of owner-occupied dwellings is left out "because this sector lacks a measure of the hours homeowners spend maintaining their home." Together, these factors jointly account for about one-quarter of GDP. If labor productivity growth in the excluded activities didn't slow as much as in nonfarm business productivity growth, then the "lost" output could be smaller than $3.43 trillion per year; conversely, if productivity in the excluded activities slowed more, then the "lost" output could be larger. As long as productivity growth did not actually accelerate in these excluded sectors—which seems a fair assumption—a very conservative estimate of lost output would apply the 19 percent slowdown only to the three-fourths of GDP that the labor productivity series covers directly. This lower bound implies at least $2.57 trillion of lost output.

Some additional data can refine this lower bound estimate. First, the Bureau of Labor Statistics does compute a productivity series that adds the farming sector (which accounts for about 1 percent of GDP) to the set of covered industries. This series experienced an even larger productivity slowdown than the nonfarm business series, falling from an average growth per quarter of 0.741 percent over 1995–2004 to 0.310 percent for 2005–2015. This implies a larger amount of "missing" output—$3.80 trillion applied to GDP or a lower bound of about $2.89 trillion when applied only to the directly covered sectors. Second, I combined an unpublished Bureau of Labor Statistics series of total economy aggregate hours through 2015 with the real GDP index from the Bureau of Economic Analysis to compute a total economy labor productivity measure.[5] This metric indicates a drop in productivity growth between 1995–2004 and 2005–2015 of 0.369 percentage points per quarter. Applying this to all of GDP (which, here, the productivity metric spans) implies lost output due to the productivity slowdown of $3.21 trillion per year.

Thus, the amount of output lost to the productivity slowdown ranges somewhere between $2.57 trillion and $3.80 trillion per year. Going forward, I will analyze the case for the mismeasurement hypothesis using $3 trillion as the implied value of output "lost" because of the productivity slowdown. This measure is conservative in the sense that it leaves less total lost output for the hypothesis to explain than would applying the BLS measured productivity slowdown to all of GDP. Based on 2015 US Census estimates of a US population of 321 million living in 125 million households, this works out to output that is lower because of the productivity slowdown by $9,300 per capita and $24,100 per household.

[5] I thank Robert Gordon for sharing the hours data.

Thus, to explain the entire productivity slowdown as a figment of measurement problems implies that every person in the United States in 2015 enjoyed an average additional surplus of $9,300 that did not exist in 2004.

It is important to recognize that the question is *not* whether the average consumer surplus in 2015 is $9,300 per capita. GDP does not measure, nor ever has measured, consumer surplus. Nominal GDP values output at its market price; consumer surplus is the extent to which willingness to pay is above the market price. There surely was consumer surplus in both 2004 and 2015, and it was probably substantial in both years. The question instead is whether it is plausible that technological growth between 2004 and 2015—and in particular the advent and diffusion of digitally oriented technologies like smartphones, downloadable media, and social networks that have been the most cited examples—created $9,300 per person in *incremental and unmeasured* value above and beyond any consumer surplus that already existed in goods and services present in 2004 and was brought forward to 2015.

The Extent of the Productivity Slowdown Is Not Related to Digital Technology Intensity

Several studies have noted recent productivity slowdowns in economically advanced countries (for example, Mas and Stehrer 2012; Connolly and Gustafsson 2013; Pessoa and Van Reenen 2014; Goodridge, Haskel, and Wallis 2015). As in the US economy, these slowdowns began before the 2008–2009 financial crisis and recession (Cette, Fernald, and Mojon 2015).

Given the relatively technology-heavy profile of US production (and citation of digital technologies produced by US-based multinationals as prime examples of the sources of mismeasurement), one might argue that the fact that a productivity slowdown has occurred across a number of economies makes a measurement-based explanation for the slowdown less likely. Still, similar measurement problems could have arisen in multiple advanced economies. I test if there is any systematic relationship between the extent of a slowdown in a country and the importance of information and communications technology (ICT), whether on the production or consumption side, to that country's economy. The logic of this test is, if information and communication technologies have caused measured productivity to understate true productivity, the mismeasurement hypothesis would imply that the *measured* slowdown in productivity growth should be larger in countries with greater "ICT intensity."

I conduct this test using OECD labor productivity growth data, which contains yearly percentage changes in real GDP per worker-hour. Growth rates are reported for about three dozen countries in 2015—the latest year for which data are available—but only 30 have data going back to 1995 as needed to directly compare to the US slowdown. I combine this productivity growth data with two measures, also from the OECD, of the intensity of an economy in information and communications technology. The consumption-side measure is the fraction of a country's households with broadband internet access. My data are taken from 2007, the year in which this data was most widely available, and cover 28 countries, 25 of which overlap with those for which I can compute the change in average annual productivity growth between 1995–2004 and 2005–2015.[6] Obviously broadband access has increased since this time, but here I am interested in the much more stable cross-sectional variation. The production-side intensity metric is the share of the country's value added accounted for by industries related to information and communications technology. This data is only available for 2011. It spans 28 countries, 24 of which overlap with my productivity slowdown sample.

The ubiquity of the productivity slowdown is readily apparent in the data. Labor productivity growth decelerated between 1995–2004 and 2005–2015 in 29 of the 30 countries in the sample (Spain is the only exception). Labor productivity growth across the sample's countries fell on average by 1.2 percentage points per year between the periods, from 2.3 percent during 1995–2004 to 1.1 percent over 2005–2015. There was substantial variation in the magnitude of the slowdown, with a standard deviation of 0.9 percent per year across countries. While the crisis years of 2008–09 saw unusually weak productivity growth—these were the only two years with negative average productivity growth across the sample—the slowdown does not merely reflect the crisis years. Calculating later-period average productivity growth excluding 2008–2009 still reveals slowdowns in measured productivity growth in 28 of 30 countries (excepting Spain and Israel), with an average drop of 0.9 percentage points per year (a decline in annual rates from 2.3 to 1.4 percent). Similarly, computing the prior period average productivity growth using only 1996–2004 data in order to allow for an expanded sample gives the same results: productivity growth slows between the periods in 35 of 36 countries (Spain is again the exception).

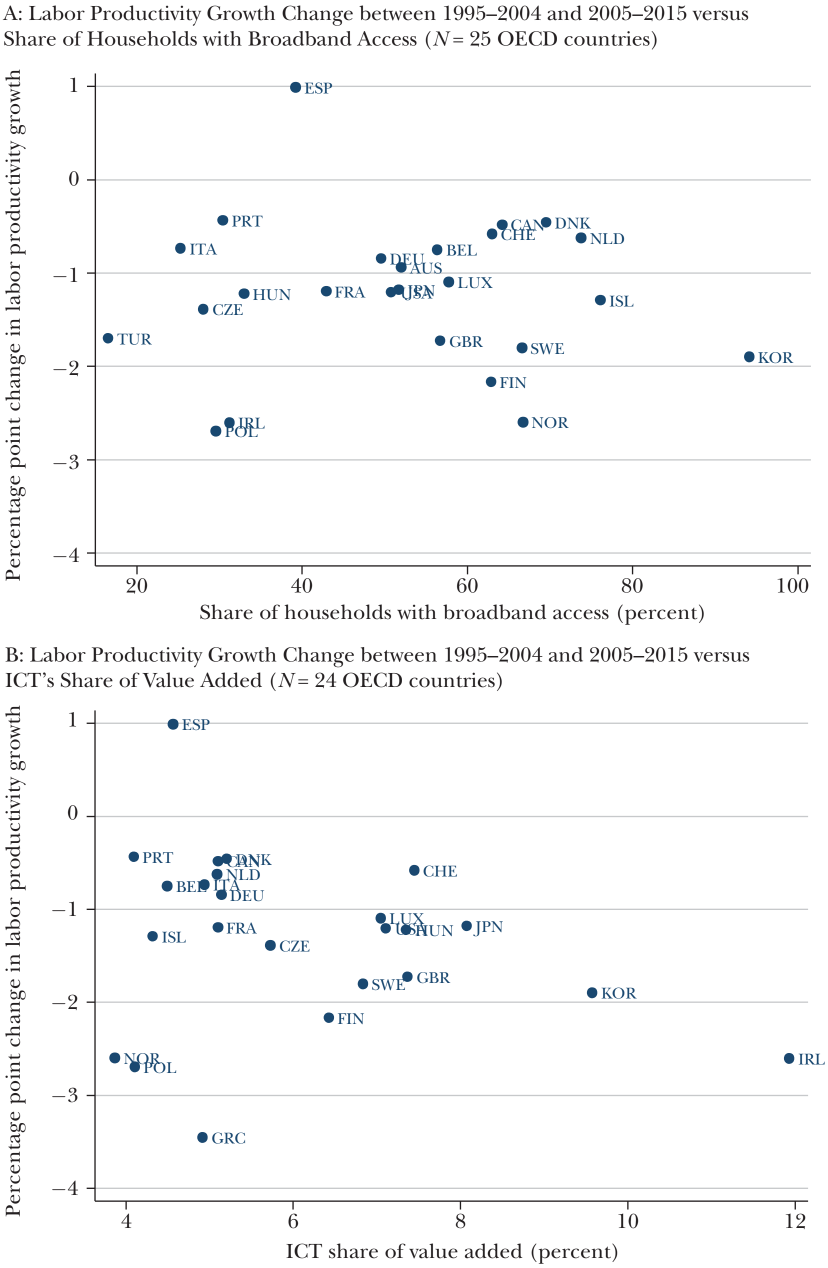
To consider the covariance between the size of a country's slowdown and its information and communications technology (ICT) intensity, Figure 1A plots each country's change in average annual labor productivity growth between 1995–2004 and 2005–2015 against the share of the country's households that have broadband access. There is no obvious relationship to the eye, and this is confirmed statistically. Regressing the change in labor productivity growth on broadband penetration yields a coefficient on broadband of -0.0003 (s.e. = 0.009). The point estimate implies that a one standard deviation difference in broadband penetration is associated with less than a one-hundredth of a standard deviation difference in the magnitude of the slowdown.

On the production side, Figure 1B plots the change in average annual labor productivity versus the share of a country's value added due to its ICT industries. Here the visual is less obvious, but as with the previous panel, a regression yields a statistically insignificant relationship. The coefficient on intensity of production

[6] Two countries, Iceland and Turkey, did not have 2015 data available, so I instead use 2005–2014 as the later period. I also use 2005–2014 for Ireland because reported labor productivity growth in 2015 was 22.5 percent, an astonishing number and one that is likely due to tax-driven corporate inversions (for example, Doyle 2016). That said, the results are not sensitive to these substitutions.

Figure 1

Change in Labor Productivity Growth versus Information and Communication Technology (ICT) Intensity



A: Labor Productivity Growth Change between 1995–2004 and 2005–2015 versus Share of Households with Broadband Access (*N* = 25 OECD countries)

B: Labor Productivity Growth Change between 1995–2004 and 2005–2015 versus ICT's Share of Value Added (*N* = 24 OECD countries)

Source: Data for both figures are from OECD. See text for details.

in information and communications technology is –0.123 (s.e. = 0.101). To the extent any relationship exists, it is due completely to the outlier Ireland, which has a value-added share in information and communications technology of 11.9 percent, double the sample average. Removing Ireland from the sample yields a statistically insignificant coefficient of –0.054 (s.e. = 0.133). This point estimate correlates a one standard deviation difference in share of value added from information and communications technology to one-eleventh of a standard deviation change in the magnitude of the productivity slowdown.

Similar results obtain both qualitatively and quantitatively if I instead measure the productivity slowdown using later-period growth rates that exclude 2008–2009 or the larger sample with 1996–2004 as the early period. This is not surprising given that the correlations between the three productivity slowdown measures are all above 0.9.

Overall, the size of the productivity slowdown in a country does not seem to be systematically related to measures of the intensity of consumption or production of information and communications technology in that country. These results echo and complement the findings of Cardarelli and Lusinyan (2015), who show that differences in the slowdown in total factor productivity growth across US states are uncorrelated with measures of state-level intensity of information and communication technologies, both as inputs and outputs in production.

Estimates of Surplus from Internet-Linked Technologies

Several researchers have attempted to measure the consumer surplus of newer technologies like those discussed in the context of the mismeasurement hypothesis. While not always explicitly motivated by the post-2004 measured productivity slowdown (some of these studies predated the recognition of the productivity slowdown among scholars), these analyses were impelled by a similar notion: certain newer technologies, those tied to internet access in particular, may have an exceptionally high ratio of consumer surplus to observed expenditure. Several studies that seek estimates of these values, which I update here, offer insight into the potential for such technologies to explain the productivity slowdown.

Greenstein and McDevitt (2009) estimate the consumer surplus created by broadband access. They choose broadband because, as an access channel, its price at least partially embodies the surplus created by otherwise unpriced technologies (for example, internet search, some downloadable media, social networking sites, and others). As Greenstein (2013) notes, "Looking at broadband demand, which does have a price, helped capture the demand for all the gains a user would get from using a faster form of Internet access." They estimate that the new consumer surplus created by households that switched from the earlier technology (dialup) was between 31–47 percent of broadband's incremental revenue over dialup. At the end of their analysis sample in 2006, this consumer surplus totaled $4.8–6.7 billion. In 2015, total US broadband revenues are estimated to be $55 billion (see The Statistics Portal, http://www.statista.com/statistics/280435/fixed-broadband-access-revenues-in-the-united-states).

Supposing broadband's overall ratio of consumer surplus to revenues is the same in 2015 as Greenstein and McDevitt (2009) estimated, this implies that the consumer surplus of broadband was $17–26 billion in 2015. Some of this value is likely priced into GDP indirectly through broadband's use by producers as an intermediate input, and as such should not be considered part of the missing output due to the productivity slowdown. But even absent any such adjustment, this surplus is two orders of magnitude smaller than the $3 trillion of missing output.

Dutz, Orszag, and Willig (2009) apply demand estimation techniques to household data on internet service take-up and prices. They estimate a consumer surplus from broadband (again relative to dialup) on the order of $32 billion per year in 2008. To scale up this value for the growth in broadband since then, I use the fact that their estimates implied the same consumer surplus was $20 billion in 2005. Assuming this robust 60 percent growth over three years (a compounded annual growth rate of 17 percent) held until 2015, consumer surplus in 2015 would be $96 billion. While this is notably larger than the Greenstein and McDevitt (2009) valuation, it is still only 3.2 percent of $3 trillion.

In another attempt to measure broadband's consumer surplus, Rosston, Savage, and Waldman (2010) use a different methodology and dataset. Their estimate is $33.2 billion in 2010. I bring this forward to 2015 using their assessment that this surplus had doubled or perhaps even tripled between 2003 and 2010, which implies a compound annual growth rate between 10.4 and 17.0 percent (which as it happens is on the order of the growth rate in Dutz, Orszag, and Willig 2009). This extrapolation implies consumer surplus was in the range of $54–73 billion in 2015. Once again, this is miniscule compared to the lost output.

Nevo, Turner, and Williams (2015) use household-level data on broadband purchases to estimate a dynamic model of broadband demand. They find an average consumer surplus among households in their data between $85 and $112 per month ($1,020–1,344 per year) in 2012. Applying this to the 80 percent of US households that had broadband access in 2015, this totals at most $132 billion—larger than the estimates above, but again less than 5 percent of the $3 trillion in missing GDP.[7]

Goolsbee and Klenow (2006) take a different approach. They use the time people spend online as an indicator of "full expenditure" on internet-based technologies. In their methodology, consumption of a good generally involves expenditure of both income and time. Therefore, even if financial expenditures on a good are relatively small, the good can deliver substantial welfare if people spend a lot of time consuming it. They argue this is a realistic possibility for the internet, which in their data (for 2005) has a time expenditure share 30 times greater than its income

[7] They also use their estimates to infer the total surplus (revenues plus consumer surplus) of access to 1 Gb/s networks, which is currently unavailable in most locations. This extrapolation implies a total surplus of $3,350 per year. Some of this would surely be captured as revenues of downstream firms and thus measured in GDP. A conservative price for this service would be $900 per year, so consumer surplus per household would be around $2,450. Even if service were obtained by every household in the country that has broadband, this adds up to $241 billion of consumer surplus, which is 8 percent of $3 trillion.

expenditure share. Applying their theoretical framework to data, they find that the consumer surplus of internet access could be as large as 3 percent of full income (the sum of actual income and the value of leisure time). This surplus would be $3,000 annually for the median person in their dataset. Brynjolfsson and Oh (2012) extended this analysis with updated data. They pay particular attention to incremental gains from free internet services, valuing these at over $100 billion (about $320 per capita) annually.

To extend the Goolsbee and Klenow (2006) value-of-time analysis to the question of the mismeasurement hypothesis, I must first compute total income in 2015. Disposable personal income totaled $13.52 trillion, about $42,100 per capita, in 2015. For the value of leisure time, I start with the fact that according to the American Time Use Survey (ATUS), the average person in 2014 spent 10.8 hours a day on non-work-related, non-personal-care activities. (Personal care includes sleep, so sleep is not included in the 10.8 hours.) I make the (very) generous assumptions that all of these 10.8 hours are leisure time and that people value them at the average after-tax wage of $22.08, regardless of employment status and whether the hours are inframarginal or marginal. This value of time is based on the estimate by the Bureau of Labor Statistics that average pre-tax hourly earnings for all nonfarm private business employees were $25.25 over the final quarter of 2015. To impute an after-tax wage, I multiply this value by the ratio of that quarter's disposable personal income ($13.52 trillion) to total pre-tax personal income ($15.46 trillion), reflecting an average tax rate of 12.5 percent. This yields a total annual value of leisure time of about $87,000 per person. Adding this to personal income gives a total income equal to $129,100 per capita.

Applying the Goolsbee and Klenow (2006) top-end estimate that it is 3 percent of total income, I end up with a measure of the consumer surplus from the internet in 2015 of around $3,900 per capita.[8] Assuming this surplus accrues mainly to the 80 percent of people with broadband access in their household, the aggregate benefit is $995 billion. Going through the same set of computations with 2004 data (when broadband penetration was about 12 percent according to OECD data) and subtracting the result so as to estimate incremental gains from broadband-based technologies yields a post-2004 incremental surplus from broadband of $863 billion.[9]

[8] As noted in the text, the 3 percent value is determined in part from Goolsbee and Klenow's (2006) time use data. It is plausible that the ratio of the internet's time expenditure share to its income expenditure share could have risen in the intervening decade, thereby raising this number. However, comparable contemporaneous data necessary to check this is difficult to find. The ATUS does not offer a separate item for online activity save for an email category that accounts for a tiny share of time. Many commercially available data products do not separate online leisure from online work time (the latter being an input into production rather than a final output) and allow multitasking, so a day can be filled with more than 24 hours of activity. In absence of specific guidance, I keep the original 3 percent value here.

[9] The specific figures for 2004 are $9 trillion of nominal disposable income ($30,700 per capita given a population of 293 million), 11 hours of leisure time per day, and $18.19 per hour after-tax nominal hourly earnings (based on Bureau of Labor Statistics earnings data for 2006, the start of the all-worker-compensation series). This implies a total nominal income of $103,800 per capita. Applying the 2004–2015 GDP deflator ratio of 1.21 and multiplying by the Goolsbee–Klenow estimate of 3 percent

The Goolsbee–Klenow time-based estimate is by far the highest valuation of the internet in the literature, essentially an order of magnitude larger than the other estimates. Time-of-use valuation approaches can produce large numbers; there are always 24 hours in a day to allocate and value, and it is hard to estimate the monetary value of a minute. Indeed, one could have used a similar logic to argue that productivity numbers in the 1950s and 1960s—the height of the post-World War II productivity acceleration—were missing the allegedly massive social gains of families' fast-increasing TV viewing. I stick with common practice and apply a (generous) wage-related valuation here, but in principle the wage only applies to the unit of time on the margin of work. Inframarginal leisure time should be valued by the incremental surplus relative to the next-best use of that time: for example, the extra amount someone is willing to pay to be online as opposed to, say, watch television. This increment could be much smaller than the person's wage, and the increment and wage may be uncorrelated across people, making the $863 billion figure a large overstatement. Even given these measurement issues, the implied valuation from the time-of-use approach is still less than one-third the $3 trillion of lost income from the productivity slowdown.

Most of the technologies cited by proponents of the mismeasurement hypothesis require internet access of some sort, so these estimates of the surplus delivered by that gateway should embody the surplus of the technologies that are not priced on the margin. It is possible that some post-2004 technologies that deliver a high ratio of consumer surplus to revenue do not require internet access. The numbers above indicate, however, that to explain the bulk of the productivity slowdown in quantitative terms, these products would need to deliver surplus that is both somehow not priced either directly or through complementary goods and services, and that is as large as or larger than the biggest estimates of the surplus of internet-linked products.

What If the "Missing" Output Were Measured?

Yet another calculation of the quantitative plausibility of the mismeasurement hypothesis relates the $3 trillion of missing GDP to the value-added of the specific products associated with post-2004 technologies. I take an expansive view of which products include such technologies, in an attempt to construct something of an upper bound of the lost output that can be explained by the hypothesis.

The first step in this calculation is to select the set of technologies that would be most implicated in the mismeasurement, if GDP mismeasurement results from the migration of value from output to consumer surplus since 2004. I include the

yields a benefit of $3,800 per capita in 2015 dollars. This is very close to the 2015 figure, so almost all incremental surplus from broadband by this calculation comes from diffusion of broadband to a larger population. This increase in population with broadband is (0.8 x 321 million) – (0.12 x 293 million) = 222 million.

following sectors in this group: computer and electronic products manufacturing (NAICS 334), the entire information sector (NAICS 51), and computer systems design and related services (NAICS 5415). The first and last are self-explanatory. The information sector includes the following four subindustries: publishing (including software), except internet; motion picture and sound recording; broadcasting and telecommunications; and data processing, internet publishing, and other information services. Both internet service providers and mobile telephony carriers are in this sector (in particular, NAICS 517, telecommunications).

These industries comprise the segments of the economy most likely to produce the technologies that are the focus of claims of the mismeasurement hypothesis. They also doubtlessly contain some activity that has *not* seen considerable technological expansion over the past decade (or even the past couple of decades, for that matter). As will be clear, this overexpansive definition of the output tied to the mismeasurement hypothesis is conservative in the sense that it will tend to overestimate the missing output of these industries for which technological developments in these industries might account.

The value added of these industries in 2015 were as follows: computer/electronics manufacturing, $278 billion; information, $840 billion; computer systems design and services, $266 billion. This totals $1,384 billion.

At the precipice of the productivity slowdown in 2004, nominal value added of the sectors was $945 billion ($202 billion in computer/electronics manufacturing, $621 billion in information, and $123 billion in computer systems design and services). Applying the Bureau of Economic Analysis value-added price indices of the three sectors yields 2004 value-added expressed in 2015 dollars: $813 billion.[10]

These industries therefore saw measured real value added growth between 2004 and 2015 of about $571 billion (that is, $1,384 billion – $813 billion). If measurement problems in the products of these industries are to account for the lion's share of $3 trillion in missing GDP, the incremental consumer surplus these industries would have created would need to be over six times their measured incremental value-added. Or to put this another way, if the incremental consumer surplus implied by the mismeasurement hypothesis would in fact have been captured as measured value added (and therefore the productivity slowdown observed in the data never materialized), the real value added of the industries would actually have increased by 440 percent (($1.384 trillion + $3 trillion)/$813 billion), over six times the 70 percent growth ($1.384 trillion/$813 billion) that was actually observed in

[10] This method divides the industries' summed nominal value added in 2004 by a Tornqvist price index I constructed for the combined industries. This index is equal to the average-share-weighted sum of the log changes in each of the three components' price indexes from 2004 to 2015. Note that all three industries saw drops in their value-added price indices over the period, which is why the figure in 2015 dollars is smaller than the 2004 figure. An alternative approach of deflating each industry's 2004 nominal value added by the industry-specific deflator and summing the result implies 2004 real value added in 2015 dollars of $829 billion. The difference in the methods mostly reflects the effect of the 36 percent decline in the computer equipment manufacturing price index during the period. Note that using this latter figure for 2004 value added in the calculations below would make the "missing" output of the mismeasurement hypothesis even larger in terms of the industries' measured incremental value added.

the data. This implies an enormous amount of mismeasurement. Even to account for just one-third of the missing output, by far the largest estimate of surplus from internet-related products discussed in the prior section, the industries' "correct" value added would have had to have grown by 190 percent from 2004–2015, almost triple the measured growth.

Looking at the dual to this calculation—that is, not how much larger the "real" output would need to be, but how much larger the price deflator would need to be—is also instructive. The (Tornqvist) value-added price index for this bundle of industries fell 14 percent over 2004–2015, a compound annual growth rate of −1.4 percent. If real GDP growth has been misstated because deflators have improperly accounted for quality changes in these products, the true deflator would be that which raises measured real value added growth by the extra $3 trillion. This deflator would have a compound annual growth rate of −9.9 percent, sustained over 11 years—seven times the magnitude of the official deflator. Prices would have fallen not by 14 percent since the productivity slowdown began, but by 68 percent instead.

Some of the outputs of these industries are intermediate inputs used to make other products. Therefore, they do not directly deliver surplus to final demanders. It is possible that some of the gains from the new technologies might arise as (again mismeasured) productivity gains in the production of goods for which they are used as inputs. For example, in the 2015 input-output tables for the national income and product accounts, 83 percent of computer equipment manufacturing output was used as an intermediate in the production of another commodity. The corresponding values for information and computer services are 46 and 42 percent, respectively. The total "multiplier" effect of technological progress through input use is captured by the industry's ratio of gross output (revenues) to its value added (Domar 1961; Hulten 1978). Incremental revenues capture the gains associated not just with the industry's products per se but also any embodied productivity gains obtained through their use as inputs. To gauge the potential influence of this usage, I repeat the calculations above using revenues—that is, gross output—in place of value added.

The nominal gross output of the three sectors in 2015 was $2.29 trillion ($387 billion in computer/electronics manufacturing, $1,550 billion in information, and $353 billion in computer systems design and services). The corresponding values in 2004 were $1.67 trillion ($392 billion, $1,080 billion, and $195 billion). Again applying the Bureau of Economic Analysis price deflators (this time for gross output) to express these values in 2015 dollars yields a real gross output of $1.61 trillion.

Incremental real gross output (that is, real revenue) for this set of industries was therefore about $680 billion. A full accounting for the mismeasurement hypothesis would imply an increment to consumer surplus that is five times as large as this. Had such a surplus been captured in revenue figures, the industries' real revenues would have more than tripled over 2004–2015, rather than risen 42 percent as observed in the data. The dual calculation implies a mismeasurement-corrected deflator with a compound average growth rate of −7.3 percent over 2004–2015 instead of the official gross output price index compound average growth rate of −0.3 percent, for a total price decline of 57 percent rather than 3 percent.

These calculations reveal how severely one must believe the measured growth of these industries understates their true growth if measurement problems are to explain the overall productivity slowdown for the entire US economy. What was measured and what would have actually had to happen would be multiples apart.

A final set of calculations reinforces this point. If the data miss industry output growth, they of course also miss productivity growth. In this case, it would need to be a lot of missing productivity. These industries, combined, saw their total employment rise 3.2 percent over 2004–2015 (from 5.58 million to 5.76 million, about 0.3 percent annually). Assuming they actually produced all of the output lost to the productivity slowdown, real value added per worker, properly measured, would have risen by 415 percent over those 11 years, an astounding rate of productivity growth. For example, it is notably larger than the 83 percent productivity growth seen in durable goods manufacturing during the productivity acceleration of 1995 to 2004, when durables had the fastest labor productivity growth of any major sector and they were a primary driver of the acceleration (Oliner, Sichel, and Stiroh 2007).

Perhaps these numbers are not that surprising when one considers that these digital-technology industries accounted for only 7.7 percent of GDP in 2004. A full accounting of the productivity slowdown by the mismeasurement hypothesis requires this modest share of economic activity to account for lost *incremental* output that in 2015 is about 17 percent of GDP—over twice the 2004 size of the entire sector.

One should be mindful that it is possible that unmeasured incremental gains are being made in industries outside these. For example, more intensive use of information technologies has been a recent focus of attention (including public policy efforts) in the sizeable health-care sector. Yet evidence on the productivity benefits of specific technologies in the sector has been mixed (for example, Agha 2014; Bhargava and Mishra 2014). There does not appear to be a clear case for large missing gains in the sector. Moreover, further balancing this out is the fact that, as discussed above, the digital-product-focused industries here are defined expansively. It is unlikely that every segment in this grouping (as one example, radio broadcasting) experienced similarly rapid technological progress.

National Income versus National Product

In national income accounting, it is an identity that gross domestic product (GDP) is equal to gross domestic income (GDI)—the sum of employee compensation, net operating surplus, net taxes on production and imports, and consumption of fixed capital (that is, depreciation). However, GDP and GDI are never equal in practice, because different data are used to construct each—expenditure data on the one hand and income information on the other.

In recent years, the gap between GDI and GDP—the so-called "statistical discrepancy"—has widened, with GDI on average outpacing GDP. Table 2 shows

Table 2
Gross Domestic Income versus Gross Domestic Product

				Percent of GDI going to			
Year	*GDI ($ billions)*	*GDP ($ billions)*	*GDI–GDP gap ($ billions)*	*Labor income*	*Net operating surplus*	*Net taxes*	*Depreciation*
1995	7,573.5	7,664.1	–90.6	55.5	22.7	6.9	14.8
1996	8,043.6	8,100.2	–56.6	55.0	23.6	6.8	14.6
1997	8,596.2	8,608.5	–12.3	54.9	24.0	6.7	14.4
1998	9,149.3	9,089.2	60.1	55.5	23.5	6.6	14.3
1999	9,698.1	9,660.6	37.5	55.9	23.2	6.5	14.4
2000	10,384.3	10,284.8	99.5	56.5	22.6	6.4	14.6
2001	10,736.8	10,621.8	115	56.4	22.4	6.2	14.9
2002	11,050.3	10,977.5	72.8	55.7	22.8	6.5	15.0
2003	11,524.3	11,510.7	13.6	55.3	23.1	6.6	15.0
2004	12,283.5	12,274.9	8.6	54.9	23.5	6.7	14.9
2005	13,129.2	13,093.7	35.5	54.1	24.2	6.7	15.1
2006	14,073.2	13,855.9	217.3	53.4	24.7	6.7	15.2
2007	14,460.1	14,477.6	–17.5	54.7	22.9	6.8	15.7
2008	14,619.2	14,718.6	–99.4	55.3	21.7	6.8	16.2
2009	14,343.4	14,418.7	–75.3	54.4	22.4	6.7	16.5
2010	14,915.2	14,964.4	–49.2	53.4	23.9	6.7	16.0
2011	15,556.3	15,517.9	38.4	53.2	24.3	6.7	15.8
2012	16,358.5	16,155.3	203.2	52.7	25.3	6.6	15.5
2013	16,829.5	16,691.5	138	52.6	25.2	6.6	15.6
2014	17,651.1	17,393.1	258	52.5	25.4	6.5	15.6
2015	18,290.3	18,036.6	253.7	53.1	25.0	6.5	15.5

Note: Data are from the US Bureau of Economic Analysis, national income accounts Table 1.10.

GDI, GDP, and the gap between them in annual data for 1995–2015.[11] Over 2005–2015, a cumulative gap of $903 billion (nominal) grew between GDI and GDP. This is an average gap of about 0.5 percent of GDP per year, though not every single year saw domestic income exceed domestic product. One could argue that this gap reflects workers being paid to make products (whose labor earnings are included in GDI) that are being given away for free or at highly discounted prices relative to their value (reducing measured expenditures on these products and therefore GDP in turn). This would be an indicator of the forces surmised by the mismeasurement hypothesis.

A closer examination of the data, however, strongly suggests that the GDI–GDP gap is not a sign of the mismeasurement hypothesis.

First, the gap started opening before the productivity slowdown. GDI was larger than GDP in each of the seven years running from 1998 to 2004, all of which were a time of fast productivity growth. The average annual gap was 0.6 percent of GDP, even larger than in the slowdown period.

Second, a closer look at the composition of national income reveals patterns inconsistent with the "workers paid for making free (or nearly free) products" story.

[11] The US Bureau of Economic Analysis defines the statistical discrepancy as GDP minus GDI, so a negative reported value implies that GDI is larger than GDP. I am focusing on the extent to which GDI is greater than GDP, so I am discussing the behavior of the negative of the statistical discrepancy.

The four right-most columns in Table 2 follow the evolution of the shares of GDI paid to each of the four major income categories that comprise it. Between 2004 and 2015, employee compensation's share of GDI *fell* by 1.8 percentage points, while net operating surplus grew by 1.5 percentage points. The net taxes share fell by 0.2 percentage points and depreciation rose by 0.6 percentage points. Thus, the GDI gains over the period were tied to payments to capital that came at the expense of labor income.[12] Nor is this link between GDI and capital income only manifested in long differences; the correlation in annual data from 1995 to 2015 between the GDI–GDP percentage gap and labor's share is −0.35, while it is 0.58 for net operating surplus.

Growth in domestic income measures relative to measured domestic product therefore seems to reflect increases in capital income rather than labor income. "Abnormally" high measured income relative to measured expenditures is positively related to growth in businesses' profitability and negatively related to payments to employees. This is inconsistent with—and indeed implies the opposite of—the "pay people to build free goods" story.

Conclusion

What I have termed the "mismeasurement hypothesis" argues that true productivity growth has not slowed (or has slowed considerably less than measured) since 2004, but recent gains have not been reflected in productivity statistics, either because new goods' total surplus has shifted from (measured) revenues to (unmeasured) consumer surplus, or because price indices are overstated. My evaluation focuses on four pieces of evidence that pose challenges for mismeasurement-based explanations for the productivity slowdown that the US economy has been experiencing since 2004. Two patterns—the size of the slowdown across countries is uncorrelated with the information and communications technology intensities of those countries' economies, and the GDI–GDP gap began opening before the slowdown and in any case reflects capital income growth—are flatly inconsistent with the implications of the mismeasurement hypothesis. Two others—the modest size of the existing literature's estimates of surplus from internet-linked products and the large implied missing growth rates of digital technology industries that the mismeasurement hypothesis would entail—show the quantitative hurdles the hypothesis

[12] These income share changes are a reflection of the trends that other researchers have been exploring in other contexts (for example, Elsby, Hobijn, and Şahin 2013; Karabarbounis and Neiman 2014). An alternative decomposition of income yields the same implications as those described here. This alternative divides national income (gross domestic income adjusted for international transfers minus depreciation) into employee compensation, proprietor's income, capital income (the sum of rental income, corporate profits, and net interest), and a residual category that is the sum of net taxes on production and imports plus business transfer payments plus the surplus of government enterprises. As with the results above, labor's share fell as capital's share rose over 2004–2015. Employee compensation's share of national income fell by 2.1 percentage points while capital income grew by 2.5 percentage points. (Proprietors' income share fell by 0.3 percentage points and the share of taxes fell by 0.1 percentage point over the period.)

must clear to account for a substantial share of what is an enormous amount of measured output lost to the slowdown (around $9,300 per person per year).

These results do not definitively rule out the possibility that productivity measurement problems may have developed over the past decade for specific products or product classes. However, the combined weight of the patterns presented here makes clear that the intuitive and plausible empirical case for the mismeasurement hypothesis faces a higher bar in the data, at least in terms of its ability to account for a substantial portion of the measured output lost to the productivity slowdown.

In addition to the quantitative analyses above, several qualitative points further bolster the case for skepticism about the mismeasurement hypothesis.

As briefly mentioned above, concerns about GDP mismeasurement preceded the recent slowdown, particularly regarding GDP's disconnect with social welfare. Perhaps, the argument goes, even if true productivity growth has slowed, it need not be the case that welfare growth has. I agree that GDP does not measure social welfare; it was not designed to do so. But the GDP-welfare disconnect is not a recent phenomenon. The mere fact that GDP is an imperfect measure of welfare is insufficient as evidence for the measurement hypothesis; instead, to support the hypothesis one must argue that a *break* in the GDP-welfare disconnect somehow developed around 2004. None of the evidence presented above indicates this. In fact, the estimates of the benefits of internet-linked technologies are measures of consumer surplus, which by definition are not in GDP. In other words, even if all of that surplus (recall the largest estimate is $863 billion) were somehow captured in GDP—which is not typically the case—it would still fall considerably short of making up for the GDP lost because of the productivity slowdown.

A second point is that my four analyses took as given the possibility that, as the mismeasurement hypothesis asserts, many new goods post-slowdown are missed in GDP because of low or zero prices. However, it is not clear at all that this baseline assertion is correct. To enjoy all these free goods—Facebook, the camera on your phone, Google searches, and so on—one must purchase complementary goods: a smart phone, an iPad, broadband access, mobile telephony, and so on. If companies that sell those complements know what they are doing, they ought to be pricing the value of those "free goods" into the price of the complementary products. Their value ought to be captured in the product accounts through the prices of the complementary products that are required to consume them. As an example, at least one of these complementary goods sellers, Apple, has been famously profitable during the slowdown.

Finally, in parallel with this study, other researchers have been conducting independent work that also looked at the mismeasurement hypothesis. Their approaches used different methods and data than mine, yet they came to the same conclusion. I mentioned earlier the work by Cardarelli and Lusinyan (2015), which shows that the differing rates of productivity slowdown across US states are not related to variations in the intensity of information and communications technology production across states. Nakamura and Soloveichik (2015) estimate the value of advertising-supported internet consumer entertainment and information. They apply the existing procedures for valuing advertising-supported media content in GDP, and find that accounting for free-to-consumers content on the internet raises GDP growth by less than 0.02 percent per year. Byrne, Fernald, and Reinsdorf (2016) offer two main arguments. First, they readily admit that information technology hardware is mismeasured since 2004, but they argue that the mismeasurement was even larger in the 1995–2004 period. Moreover, more of the information technology hardware was produced in the United States in the 1995–2004 period. Taken together, these adjustments imply that the slowdown in labor productivity since 2005 looks worse, not better. The second main point is that consumers are using many information and communications technologies to produce service for their nonmarket time, which means that consumers benefit, but gains in nonmarket production (which in any event are small) do not suggest that market sector productivity is understated.

If the theory that new products caused the productivity slowdown is to be resurrected, it may well need to take on a different form. For example, one very speculative mechanism that would tie a *true* productivity slowdown to people spending a large share of their time on zero-to-low-marginal-price activities would be if workers substituted work effort for technology consumption—for example, spending time while they are at work on social networking sites. This pattern would heighten consumer surplus in a way largely unmeasured by standard statistics while at the same time reducing output per hour—that is, measured labor productivity. Of course, to explain a slowdown in annual labor productivity growth, this substitution would need to be occurring in ever-greater magnitudes over time.

The empirical burdens facing the mismeasurement hypothesis are heavy, and more likely than not, much if not most of the productivity slowdown since 2005 is real. Whether that slowdown will end anytime soon remains an open question.

■ *I thank Erik Brynjolfsson, Dave Byrne, Austan Goolsbee, Bob Gordon, Jan Hatzius, Pete Klenow, Rachel Soloveichik, Hal Varian, and the JEP editors for comments. I have no financial interests relevant to this study.*

References

Agha, Leila. 2014. "The Effects of Health Information Technology on the Costs and Quality of Medical Care." *Journal of Health Economics* 34: 19–30.

Alloway, Tracy. 2015. "Goldman: How 'Grand Theft Auto' Explains One of the Biggest Mysteries of the U.S. Economy." Bloomberg, May 26. http://www.bloomberg.com/news/articles/2015-05-26/goldman-how-grand-theft-auto-explains-one-of-the-biggest-mysteries-of-the-u-s-economy.

Baily, Martin Neil, James Manyika, and Shalabh Gupta. 2013. "U.S. Productivity Growth: An Optimistic Perspective." *International Productivity Monitor* 25(Spring): 3–12.

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

Bhargava, Hemant K., and Abhay Nath Mishra. 2014. "Electronic Medical Records and Physician Productivity: Evidence from Panel Data Analysis." *Management Science* 60(10): 2543–62.

Brynjolfsson, Erik, and Andrew McAfee. 2011. *Race Against the Machine: How the Digital Revolution Is Accelerating Innovation, Driving Productivity, and Irreversibly Transforming Employment and the Economy.* Lexington, MA: Digital Frontier Press.

Brynjolfsson, Erik, and Andrew McAfee. 2014. *The Second Machine Age: Work, Progress, and Prosperity in a Time of Brilliant Technologies.* New York: W.W. Norton & Company.

Brynjolfsson, Erik, and JooHee Oh. 2012. "The Attention Economy: Measuring the Value of Free Digital Services on the Internet." *Proceedings of the International Conference on Information Systems*, Orlando. http://aisel.aisnet.org/cgi/viewcontent.cgi?article=1045&context=icis2012.

Byrne, David M., John G. Fernald, and Marshall B. Reinsdorf. 2016. "Does the United States Have a Productivity Slowdown or a Measurement Problem?" *Brookings Papers on Economic Activity*, Spring, 109–182.

Byrne, David M., Stephen D. Oliner, and Daniel E. Sichel. 2013. "Is the Information Technology Revolution Over?" *International Productivity Monitor* 25(Spring): 20–36.

Byrne, David M., Stephen D. Oliner, and Daniel E. Sichel. 2015. "How Fast Are Semiconductor Prices Falling?" NBER Working Paper 21074.

Cardarelli, Roberto, and Lusine Lusinyan. 2015. "U.S. Total Factor Productivity Slowdown: Evidence from the U.S. States." IMF Working Paper 15/116.

Cette, Gilbert, John Fernald, and Benoît Mojon. 2015. "The Pre-Global-Financial-Crisis Slowdown in Productivity." Unpublished Paper.

Connolly, Ellis, and Linus Gustafsson. 2013. "Australian Productivity Growth: Trends and Determinants." *Australian Economic Review* 46(4): 473–82.

Cowen, Tyler. 2011. *The Great Stagnation: How America Ate All the Low-Hanging Fruit of Modern History, Got Sick, and Will (Eventually) Feel Better.* New York: Dutton.

Davis, Steven J., and John Haltiwanger. 2014. "Labor Market Fluidity and Economic Performance." NBER Working Paper 20479.

Decker, Ryan, John Haltiwanger, Ron Jarmin, and Javier Miranda. 2014. "The Role of Entrepreneurship in US Job Creation and Economic Dynamism." *Journal of Economic Perspectives* 28(3): 3–24.

Diewert, W. Erwin, and Kevin J. Fox. 1999. "Can Measurement Error Explain the Productivity Paradox?" *Canadian Journal of Economics* 32(2): 251–80.

Domar, Evsey D. 1961. "On the Measurement of Technological Change." *Economic Journal* 71(284): 709–29.

Doyle, Dara. 2016. "Ireland's Economists Left Speechless by 26% Growth Figure." *Bloomberg*, July 12, 2016.

Dutz, Mark, Jonathan Orszag, and Robert Willig. 2009. "The Substantial Consumer Benefits of Broadband Connectivity for U.S. Households." Commissioned by the Internet Innovation Alliance. https://internetinnovation.org/files/special-reports/CONSUMER_BENEFITS_OF_BROADBAND.pdf.

Elsby, Michael W. L., Bart Hobijn, and Ayşegül Şahin. 2013. "The Decline of the U.S. Labor Share." *Brookings Papers on Economic Activity*, Fall, pp. 1–63.

Feldstein, Martin. 2015. "The U.S. Underestimates Growth." *Wall Street Journal*, May 18.

Fernald, John G. 2014a. "Productivity and Potential Output Before, During, and After the Great Recession." NBER Working Paper 20248.

Fernald, John G. 2014b. "A Quarterly, Utilization-Adjusted Series on Total Factor Productivity." Working Paper 2012-19, Federal Reserve Bank of San Francisco (updated March 2014).

Goodridge, Peter, Jonathan Haskel, and Gavin Wallis. 2015. "Accounting for the UK Productivity Puzzle: A Decomposition and Predictions." Imperial College London Discussion Paper 2015/02.

Goolsbee, Austan, and Peter J. Klenow. 2006. "Valuing Consumer Products by the Time Spent Using Them: An Application to the Internet." *American Economic Review* 96(2): 108–13.

Gordon, Robert J. 2016. *The Rise and Fall of American Growth: The U.S. Standard of Living since* *the Civil War*. Princeton University Press.

Greenstein, Shane. 2013. "Measuring Consumer Surplus Online." *Economist*, March 11.

Greenstein, Shane, and Ryan C. McDevitt. 2009. "The Broadband Bonus: Accounting for Broadband Internet's Impact on U.S. GDP." NBER Working Paper 14758.

Hatzius, Jan, and Kris Dawsey. 2015. "Doing the Sums on Productivity Paradox v2.0." *Goldman Sachs U.S. Economics Analyst*, No. 15/30.

Hulten, Charles R. 1978. "Growth Accounting with Intermediate Inputs." *Review of Economic Studies* 45(3): 511–18.

Karabarbounis, Loukas, and Brent Neiman. 2014. "The Global Decline of the Labor Share." *Quarterly Journal of Economics* 129(1): 61–103.

Klenow, Peter J. 2003. "Measuring Consumption Growth: The Impact of New and Better Products." *Federal Reserve Bank of Minneapolis Quarterly Review* 27(2): 10–23.

Mas, Matilde, and Robert Stehrer, eds. 2012. *Industrial Productivity in Europe: Growth and Crisis.* Northhampton, MA: Edward Elgar.

Mokyr, Joel. 2014. "Secular Stagnation? Not in Your Life." Chap. 6 in *Secular Stagnation: Facts, Causes and Cures,* edited by Coen Teulings and Richard Baldwin. London: CEPR Press.

Nakamura, Leonard, and Rachel Soloveichik. 2015. "Valuing 'Free' Media across Countries in GDP." Federal Reserve Bank of Philadelphia Working Paper 15-25.

Nevo, Aviv, John L. Turner, and Jonathan W. Williams. 2015. "Usage-Based Pricing and Demand for Residential Broadband." Available at SSRN: http://ssrn.com/abstract=2330426.

Oliner, Stephen D., Daniel E. Sichel, and Kevin J. Stiroh. 2007. "Explaining a Productive Decade." *Brookings Papers on Economic Activity* no. 1, pp. 81–152.

Pessoa, João Paulo, and John Van Reenen. 2014. "The UK Productivity and Jobs Puzzle: Does the Answer Lie in Wage Flexibility?" *Economic Journal* 124(576): 433–52.

Rosston, Gregory L., Scott J. Savage, and Donald M. Waldman. 2010. "Household Demand for Broadband Internet in 2010." *B.E. Journal of Economic Analysis & Policy: Advances* 10(1): 1–45.

Smith, Noah. 2015. "The Internet's Hidden Wealth." *Bloomberg View,* June 10, 2015. http://www.bloombergview.com/articles/2015-06-10/wealth-created-by-the-internet-may-not-appear-in-gdp.

Syverson, Chad. 2013. "Will History Repeat Itself? Comments on 'Is the Information Technology Revolution Over?'" *International Productivity Monitor* 25: 37–40.

Tarullo, Daniel K. 2014. "Longer-Term Challenges for the American Economy." Speech given to 23rd Annual Hyman P. Minsky Conference: Stabilizing Financial Systems for Growth and Full Employment, Washington, D.C., April 9, 2014.

US Congress. 1996. "Toward a More Accurate Measure of the Cost of Living." Final Report to the Senate Finance Committee from the Advisory Commission to Study the Consumer Price Index. (The Boskin Commission Report.) 104th Congress, 2nd sess., S.Prt. 104-072. http://www.ssa.gov/history/reports/boskinrpt.html.

Journal of Economic Perspectives—Volume 31, Number 2—Spring 2017—Pages 187–210

How Government Statistics Adjust for Potential Biases from Quality Change and New Goods in an Age of Digital Technologies: A View from the Trenches

Erica L. Groshen, Brian C. Moyer, Ana M. Aizcorbe, Ralph Bradley, and David M. Friedman

> *"[W] hen you can measure what you are speaking about, and express it in numbers, you know something about it; but when you cannot measure it, when you cannot express it in numbers, your knowledge is of a meagre and unsatisfactory kind."*
>
> William Thomson, Lord Kelvin, *Electrical Units of Measurement* (1883)

The US Congress created and funds the US Bureau of Labor Statistics (BLS) and the US Bureau of Economic Analysis (BEA) to produce essential information on economic conditions to inform public and private decisions. A key economic indicator—what our economy produces during a time period—is called "real output" and measured as the nominal dollar value of gross domestic product (GDP) deflated by price indexes to remove the influence of inflation. To get this right, we need to measure accurately both the value of nominal GDP (done by BEA) and key price indexes (done mostly by BLS). Otherwise, real output and related measures like productivity growth statistics will be biased.

■ *Erica L. Groshen was Commissioner, US Bureau of Labor Statistics, Washington, DC, from January 2013 through January 2017. Brian C. Moyer is Director, US Bureau of Economic Analysis, Washington, DC. Ana M. Aizcorbe is Senior Research Economist, US Bureau of Economic Analysis, Washington, DC. Ralph Bradley is Chief of the Division of Price and Index Number Research, US Bureau of Labor Statistics, Washington, DC. David Friedman is Associate Commissioner for Prices and Living Conditions, US Bureau of Labor Statistics, Washington, DC. Their email addresses are egroshen@yahoo.com, brian.moyer@bea.gov, ana.aizcorbe@bea.gov, Bradley.Ralph@bls.gov, and Friedman.David@bls.gov.*

† For supplementary materials such as appendices, datasets, and author disclosure statements, see the article page at
https://doi.org/10.1257/jep.31.2.187

doi=10.1257/jep.31.2.187

All of us have worked on these measurements while at the Bureau of Labor Statistics and the Bureau of Economic Analysis. In this article, we explore some of the thorny statistical and conceptual issues related to measuring a dynamic economy. An often-stated concern in recent years is that the national economic accounts miss some of the value of some goods and services arising from the growing digital economy. We agree that measurement problems related to quality changes and new goods have likely caused growth of real output and productivity to be understated. Nevertheless, these measurement issues are far from new and, based on the magnitude and timing of recent changes, we conclude that it is unlikely that they can account for the pattern of slower growth in recent years.

We begin by discussing how the Bureau of Labor Statistics currently adjusts price indexes to reduce the bias from quality changes and the introduction of new goods, along with some alternative methods that have been proposed. We then present estimates of the extent of remaining bias in real GDP growth that stem from potential biases in growth of consumption and investment. Based on our analysis of bias estimates performed by experts external to the Bureau of Labor Statistics and the Bureau of Economic Analysis, we find that these influences on existing price indexes may overstate inflation in the categories of "personal consumption expenditures" and "private fixed investment," leading to a corresponding understatement of real economic growth of less than one-half percentage point per year. Furthermore, we find this to be fairly stable over time. We then also take a look at potential biases that could result from challenges in measuring nominal GDP, including assessing the significance of the argument that the digital economy has created some valuable goods and services (such as smartphone apps and Internet searches) that are not bought or sold, and thus are not counted in GDP. Finally, we review some of the ongoing work at BLS and BEA to reduce potential biases and further improve measurement.

Challenges in Measuring Price Indexes

The Bureau of Labor Statistics publishes monthly indexes of consumer prices, producer prices, and import and export prices. The Consumer Price Index (CPI) measures average changes in the prices paid by urban consumers for a representative set of goods and services. The CPI is often used to make cost-of-living adjustments; indeed, the BLS uses the concept of a Cost of Living Index (as defined by Konüs 1939) as a unifying framework and is the standard by which BLS defines any bias. The Producer Price Index (PPI) measures average change in selling prices received by domestic producers for their output. Nominal value of production can be deflated by the PPI to get an output measure. Finally, Import and Export Price Indexes (MXP) measure average changes in the prices of nonmilitary goods and select services traded between the US economy and the rest of the world. The MXP allows one to estimate real aggregate trade volumes from nominal trade amounts.[1]

[1] For more information on the PPI, see Chapter 14 of the BLS *Handbook of Methods* (2006). For more detail on the CPI, see Abraham (2003) as well as Chapter 17 of the BLS *Handbook of Methods*. For more information about the MXP, see Chapter 15 of the BLS *Handbook of Methods*.

The Bureau of Labor Statistics seeks to produce the best possible monthly price indexes, subject to some very practical and binding constraints. On average, no more than 20 days elapse between the collection of a price and the final publication of the index. This must be done within a rigid budget constraint. Furthermore, the confidentiality of all collected data must be strictly protected at every stage of index construction (more at the BLS website at https://www.bls.gov/bls/confidentiality.htm). Respondent participation is entirely voluntary, so the Bureau of Labor Statistics also aims to minimize respondent burden because its field representatives must be able to persuade respondents to participate.[2] These considerations mean that for a methodological improvement to be implemented, it must meet the following criteria: it 1) is feasible within the BLS budget constraint; 2) is computable and reviewable within 20 days; 3) is compatible with the skill set of BLS staff; 4) requires no increase in samples or new surveys (unless there is budget approval for a new survey); 5) does not unduly burden respondents; and 6) is proven to reduce bias in a statistically significant manner.

So how does the Bureau of Labor Statistics treat new and evolving goods and services? To begin with, this problem is hardly a recent development. For example, consider the 1920s. That decade saw a rapid introduction of new goods such as indoor plumbing, electricity, and radios, as well as dramatic quality improvements of existing products such as automobiles and airplanes. Over the past century, technical innovation has continued to improve existing goods and has led to the introduction of myriad new products.

From a price index perspective, the biases caused by technological innovation are distinct from the sometimes-conflated issue of substitution bias. Substitution bias may occur when either a new outlet enters the market and offers existing products at a lower price, or when a foreign country starts producing a lower-price product that was already produced domestically. For example, not accounting for the substitution from US-produced manufacturing inputs to lower-priced foreign inputs, as studied by Houseman, Kurz, Lengermann, and Mandel (2011), does not cause a new goods problem but rather a substitution bias problem. In particular, this input substitution exerts an *upward* bias on US productivity statistics, while not adjusting for quality improvements or new products leads to a *downward* bias on the same statistics.

The decades-long search of academics and statistical agency staff for better ways to adjust price indexes for innovations has generated a vast literature on this topic.[3] These methods fall into six groups: 1) quality adjustment from producers, 2) outside surveys to measure quality changes, 3) hedonic approaches, 4) discrete

[2] Countries vary on whether participation in price surveys is voluntary. For example, participation in Norway's and Canada's price index programs is mandatory (for details, on Statistics Norway, see Johannessen 2016, https://ec.europa.eu/eurostat/cros/system/files/randi_johansenn_the_use_of_scanner_data_in_the_norwegian_cpi.pdf, and for Statistics Canada, see http://www.statcan.gc.ca/eng/survey/business/2301).

[3] A partial list of references for studies done by BLS staff appears in http://www.bls.gov/cpi/cpihqa-blsbib.pdf. Those not listed in this link include Armknecht, Lane, and Stewart (1996) and Erickson and Pakes (2011).

Table 1
Summary of Methods to Account for New and Improved Goods and Services

Method	*Requires demand estimation*	*Based on characteristics, product, or other*	*Example of studies*	*In production*	*Reason not in production*
Quality adjustment from producer	No	Characteristics	Moulton, LaFleur, and Moses (1998)	Yes; PPI, MXP, CPI[a]	
Input from other surveys	No	Characteristics	Murphy et. al (2008)	Yes; primarily PPI	
Explicit hedonic quality adjustment	No	Characteristics	Fixler, Fortuna, Greenlees, and Lane (1999)	Yes; CPI[b], PPI[c], MXP[c]	
Time dummy hedonic index	No	Characteristics	Byrne, Oliner, and Sichel (2015); Berndt, Griliches, and Rappaport (1995); Griliches (1961)	No	Restrictive assumptions
Imputed hedonic index	No	Characteristics	Erickson and Pakes (2011)	No	Requires larger sample sizes
Discrete choice	Yes	Characteristics	Berry, Levinsohn, and Pakes (1995); Nevo (2001); Petrin (2002)	No	High computational intensity and cost; poor timeliness
Consumer surplus	Yes	Product	Lee and Pitt (1986); Hausman (1997); Broda and Weinstein (2010)	No	Endogeneity problems (under investigation); high cost
Disease-based price indexes	No	Treated disease	Aizcorbe and Nestoriak (2011); Bradley (2013)	Partial; BEA and BLS experimental indexes	Do not yet adjust for differences in outcomes

[a] For example this is done for new vehicles in the CPI and PPI.
[b] See http://www.bls.gov/cpi/cpihqablsbib.pdf for CPI items that are quality adjusted.
[c] PPI and MXP do explicit hedonic quality adjustment for computers.

choice models, 5) explicit measurement of increased consumer surplus from new goods, and 6) the special case of disease-based price indexes. Table 1 lists the various methods and shows which are used in the current production of price indexes by the Bureau of Labor Statistics. The last column explains briefly why some methods are not being used currently. While any method that government statistical agencies use for computing price indexes needs to have a sound theoretical basis, the table illustrates that the operational requirements (like allowing for index computation on a timely basis in a transparent fashion and within budgetary constraints) are also often binding.

Current Quality and New-Product Adjustment Methods

The *matched model* is the cornerstone of constructing price indexes at the Bureau of Labor Statistics. When products match over time, the characteristics of each product are held constant. Thus, any price change can only be attributed to inflation, and not to changes in characteristics. For example, from December 2013 through November 2014, matches were found for items in the Consumer Price Index 73 percent of the time. Of the remaining 27 percent of items that were not matched, 22 percent reflected temporarily missing items, such as a bathing suit in Milwaukee in December. The other 5 percent represented a permanent disappearance. (Note that similar figures are not available for the Producer Price Index or the Import and Export Price Indexes.)

When a match permanently ends in the Consumer Price Index and the same good cannot be tracked from one period to the next, then (except for housing) the Bureau of Labor Statistics initiates a *quality adjustment procedure* after a replacement good has been established. When the replacement has characteristics very similar to the exiting product, the price of the replacement product is used in place of the exiting product. For example, of the 5 percent of the CPI that represented permanently disappearing items during the period noted above, three-fifths of those items were replaced by a similar good. For the remaining two-fifths, where the characteristics were judged to be insufficiently close, BLS staff made a quality adjustment to the replacement product's price.

In some cases, an item's producer can provide a value for the change in its characteristics. The Bureau of Labor Statistics uses this value to adjust the transaction price before it is entered into the index. This method is referred to as *explicit quality adjustment* and is most prevalent in the Producer Price Index or the Import and Export Price Indexes. It is especially important for automobiles, machinery, and other types of goods that undergo periodic model changes. For example, the price increment due to a new protective coating or electronic feature on a machine part can often be provided by its manufacturer. In general, explicit quality adjustment is easier to do for goods industries than for services industries, because producers are more likely to be willing to estimate a monetary value for the quality of a good, rather than a service. If the item's producer cannot provide a value for a characteristic change, then the *overlap* method is used. This entails having historical prices on both the exiting item and the replacement item at the same time. When computing a price index, the price of the new replacement is reduced by dividing by the ratio of the price of the new good to the exiting good (for a more detailed explanation by the BLS see "Quality Adjustment in the Producer Price Index" http://www.bls.gov/ppi/qualityadjustment.pdf).[4]

Quality adjustments *from producers* are generally cost-based. Some observers argue that utility- or welfare-based quality adjustments would be an improvement. Triplett (1982) attempts to find a resolution for this disagreement. He concludes, "In output price indexes (the fixed-weight forms as well as the theoretical ones based on production possibility curves), the quality adjustment required is equal

[4] This method is used by the Billion Prices Project, discussed in this journal by Cavallo and Rigobon (2016). They argue that with the overlap method, "As we increase the number of models included in the index, we more closely approximate the results of the hedonic price index constructed by the BLS." Statistics Canada also uses this approach (see "The Canadian Consumer Price Index Reference Paper," http://www.statcan.gc.ca/pub/62-553-x/62-553-x2015001-eng.pdf).

to the resource usage of the characteristics that changed. Only with a resource-cost adjustment does the index price a set of outputs that can be produced with the resources available in the reference period." Because the measurement objective of real GDP is economic output, rather than welfare, the Bureau of Labor Statistics believes that cost information is appropriate for adjusting prices in an index whose purpose is to deflate nominal industry revenues to measure real output. Other practitioners agree. For example, the International Monetary Fund (2004, chapter 7) views this approach as a best practice for producer price indexes.

Finally, other surveys can be used to adjust for quality. For example, the US Department of Health and Human Services has created a Hospital Compare and a Nursing Home Compare database, which looks at inputs that experts believe can serve as proxies for quality of health care. The Bureau of Labor Statistics uses these data to adjust the hospital and nursing home components of the PPI (Bureau of Labor Statistics 2008 and "Quality Adjustment in the Producer Price Index" at http://www.bls.gov/ppi/qualityadjustment.pdf). In addition, the Insurance Services Office (a private firm) creates a database on the risk characteristics of cars, which BLS uses for quality adjustments in auto insurance prices (as reported in Bathgate 2011).

Hedonic Adjustments

Fifty years ago, Lancaster (1966) suggested, "It is the properties or characteristics of the goods from which utility is derived." For example, we do not consume a car; we consume its horsepower, transmission, size, audio system, and other amenities. This insight helped lead to the hedonic, or the "demand for characteristics," model, which estimates how each characteristic contributes to the value of a good. The term "hedonics" derives from the Greek root for satisfaction. Court (1939) is often viewed as the first hedonic study that estimated a statistical relationship between prices and characteristics, but Lancaster (1966) and Rosen (1974) provide the major microeconomic insights of the demand for characteristics.

In the Consumer Price Index, about 33 percent of the total expenditures in the underlying basket of goods are eligible for quality adjustment with hedonics when price-determining characteristics change. Housing-related expenditures account for most of this share (as described by BLS at http://www.bls.gov/cpi/cpihqaitem.htm). Liegey (1993, 1994, 2001a, b, 2003) explains the investigations made by Bureau of Labor Statistics staff on the use of hedonics for various items in the CPI. In addition, the Producer Price Index or the Import and Export Price Indexes use hedonic adjustment procedures for computers.

Hedonic indexes start with a simple regression of the price (or log price) of a good on its observable characteristics (for more detail, see Aizcorbe 2014; Silver and Heravi 2007). The explicit hedonic quality adjustment makes the prices of two differing products comparable by adjusting the price of one good using the coefficients of the regression so that it accounts for the differences in characteristics. While the hedonic index approach needs only to be a measure of the compensating variation required to keep utility constant, the explicit quality adjustment approach requires that the coefficients on the hedonic regressions be consistent estimators of the characteristic's shadow price.

There are two variations on the hedonic approach that are not currently used by the Bureau of Labor Statistics. One approach uses a time dummy coefficient in the regression to compute the price index. In a related approach, a hedonic imputation index can be constructed by estimating a regression for each time period. The levels of characteristics are then held constant over time and a quality-adjusted price is imputed for each period. However, the assumptions behind the basic time dummy hedonic index are very restrictive (Erickson and Pakes 2011). The alternative hedonic imputation index can in theory give a more accurate measure of the compensating variation than the matched model (Pakes 2003; Erickson and Pakes 2011). Unfortunately, attempts by BLS to implement imputed hedonic price indexes revealed that our sample sizes are too small for this approach. Thus, BLS has so far continued to use explicit hedonic quality adjustment, because it has more general applicability than the time dummy approach and does not require the large sample sizes needed for the hedonic imputation indexes.

Hedonic methods are feasible when adequate sample sizes and information on relevant characteristics are available. Unlike the discrete choice and consumer surplus methods discussed in the next section, hedonic methods do not require estimating demand for a good and can be implemented with data on only prices and characteristics. These data are already collected by the price index programs. The Bureau of Labor Statistics plans to expand use of hedonic methods. However, hedonics must be implemented carefully, case by case, to ensure that key conditions are met: 1) product characteristics must be observable, ruling out features such as enhancing the user's social status; 2) the set of relevant characteristics cannot change, ruling out this approach for goods where stark new attributes are introduced frequently, such as the smartphone; and 3) the market for the product must be competitive, with markups playing only a very limited role (which ensures that a characteristic's coefficient is an unbiased estimate of its shadow price, as discussed in Rosen 1974).

When all relevant characteristics for hedonics are not available, Statistics New Zealand and Statistics Netherlands have collaborated to develop a "fixed effects window-splice" price index for scanner data (Krsinich 2016; de Haan and Krsinich 2014). If the characteristics by bar code remain fixed, then this index is equivalent to a hedonic time dummy index. This method shows much promise and could be more efficient than using the traditional time dummy approach as it would require fewer parameters. Also, this approach incorporates the effects of the entire set of characteristics whether or not they are observable. The Bureau of Labor Statistics plans to review this method.

Models Not in Production: Discrete Choice Model and Consumer Surplus

Academic researchers have proposed alternative and more sophisticated procedures that are intriguing. But, as we discuss below, thus far their implementation requires restrictive assumptions and are not feasible given the production constraints at the Bureau of Labor Statistics.

The discrete choice model is a demand-for-characteristics model, like that used to motivate hedonics. However, unlike hedonic methods, discrete choice methods do not assume that the quantity of each good consumed is a continuous variable. McFadden (1978) introduces the discrete choice model, which has several theoretical advantages over hedonics. First, fewer assumptions are required for the coefficients on the characteristics to be an unbiased estimate of the characteristics' shadow prices. Second, the supply side can be explicitly modeled and markups can be allowed. Third, unlike hedonics, this method yields estimates of both consumer demand and aggregate utility based on the prices and characteristics of each product for each time period. Fourth, estimation of these models is able to take advantage of information from volume or sales data, while hedonic methods do not. In the discrete choice model, the quality-adjusted price index is usually computed as the percent change in income that consumers need to keep their expected utility constant over time (for discussion, see Nevo 2003 on how to compute quality-adjusted price indexes from discrete choice models).

However, *in practice* the discrete model approach runs into difficulties. It originally came under criticism for its highly restrictive assumptions, but additional modifications introduced later relaxed these assumptions (as discussed in Berry 1994; Berry, Levinsohn, and Pakes 1995; Nevo 2000). Yet these modifications did not solve key practical difficulties including large computational and personnel cost increases, to the extent that computationally intensive problems may take years to complete (Berry 1994; Berry, Levinsohn, and Pakes 1995; Petrin 2002; Nevo 2001). In addition, the Bureau of Labor Statistics does not have access to the necessary volume sales or share data for estimating the parameters of these models. Even if these volume data were available, to implement these models in the Consumer Price Index would pose serious logistical challenges. The CPI is constructed from over 6,000 item-area subprice indexes. For each price index item that needs quality adjustment, BLS would need to solve a separate discrete choice model for each of that item's areas and/or expenditure categories during the 20 days that BLS has available between price collection and publishing the monthly price index. This could easily require solving over a thousand models each month, which is not currently feasible.

The consumer surplus method is based on the demand for products and not characteristics. Using this method to adjust for changes in quality or new goods is problematic for other reasons. In this method, when a new product is introduced, its demand is estimated econometrically as a function of the new price and prices of incumbent products in a system of structural demand equations that include the other competing products (Hausman 1997; Redding and Weinstein 2016). The parameter estimates of the demand system are then used to solve for the *virtual price* where no one purchases this new product. The price index is computed by plugging in this virtual price for the period when the product had not been introduced. This idea that before a new good is introduced, it should be considered available but at a price at which the quantity demanded is zero, was introduced by Hicks (1940). Generally, a higher estimated virtual price will generate a higher consumer surplus coming from the new good, leading in turn to a lower price index.

Implementing this approach has proven most controversial. For example, the Hausman (1997) study of the cereal market is based on the consumer surplus model. A separate demand equation is estimated for each brand of cereal, such as Special K or Cheerios. This method is more computationally intensive than the hedonic method, where only one reduced form regression of log price on characteristics is estimated. This approach has the potential for bias from specification error, using the wrong instruments, and making incorrect assumptions on how demand shocks affect prices.[5] These sources of econometric bias could produce a biased estimate of the virtual price. The Committee on National Statistics advises against using the Hausman (1997) method, citing the risks of introducing new errors (National Research Council 2002). Moreover, a potentially troubling aspect of this approach is that the bias of the virtual price may be extremely large. For example, Hausman (1997) concludes that the CPI for cereal "may be too high by about 25%." His average virtual price of the new cereal product is $7.14 while the observed average price is $3.78. It is difficult to imagine that introducing a new flavor of Cheerios could have such a large impact on a true cereal price index.

Newer versions of the consumer surplus approach are still being evaluated for production feasibility. As one example, Redding and Weinstein (2016) and Broda and Weinstein (2010) use a constant elasticity of substitution form of preferences. They assume that the elasticity of substitution is bounded below by one. With these restrictions, they only need to estimate the elasticity of a substitution parameter and not additional preference parameters. This method is far more computationally efficient than Hausman's (1997) approach, and it sidesteps the issue that a "virtual price" where demand is literally zero can be very high. Typically, this method is used with scanner data, and it is possible that when prices bounce over a fixed time period and then return to their original level, the price index does not equal one. This problem is referred to as *chain drift*. Any index that does not satisfy the transitivity and time reversal axioms can be subject to chain drift. Ivancic, Diewert, and Fox (2009) develop a multilateral price index method that corrects for this problem, and Krsinich (2015) shows how this method has been adopted for the New Zealand electronics equipment price index.[6]

[5] Hausman (2003) argues that his consumer surplus method is the solution to the new goods problem in the Consumer Price Index. When Hausman (1997) accounted for the introduction of Apple Cinnamon Cheerios, he had to estimate the demand of all other brands of cereal. Bresnahan (1997) and Bradley (2005) critique the approach used in Hausman (1997). Among other concerns, Hausman's use of prices in other markets as instruments and his use of unit values could have distorted his estimation of the virtual price. Abraham (2003) provides a thorough explanation of the problems of implementing the Hausman (1997) method.

[6] This index is called a GEKS index, after Gini (1931), Eltetö and Köves (1964), and Szulc (1964). While this approach eliminates chain drift, month to month price indexes are functions of prices that are outside the base and comparison months. This is called "loss of characteristicity" (de Haan and van der Grient 2011). Even if the individual bilateral indexes that are plugged into this GEKS index are superlative indexes, there is no guarantee that the GEKS index itself is a superlative index.

Disease-Based Price Indexes

In our view, a substantial current challenge for adjusting for new products lies in the medical sector because of its size, rapid innovation, and unique market features. While many observers focus on technological innovations in the digital economy like the smartphone or other aspects of information technology, the medical sector is still critical because healthcare spending is large: 17.5 percent of GDP in 2014. In the same year, all spending on phones was $16.6 billion—about 0.1 percent of GDP. Several studies conclude that medical price indexes are upwardly biased because many new treatments for particular diseases replace more expensive approaches (for examples, for heart disease, see Cutler, McClellan, Newhouse, and Remler 1998; for depression, see Berndt et al. 2002; more broadly, see Shapiro, Shapiro, and Wilcox 2001).

Presently, there are two sources of medical price data: providers' billing offices and health insurance claims. Neither source provides characteristics data, such as the remission of disease, length of time for healing, and other indicators for patient wellness. This is a major constraint for quality-adjusting medical prices.

As a partial solution, the Bureau of Labor Statistics and the Bureau of Economic Analysis have created experimental disease-based price indexes that correct for a portion of the new goods bias in medical care, specifically the part that arises when less-expensive goods and services substitute for more-expensive treatments (for discussion, see Aizcorbe and Nestoriak 2011; Bradley 2013).[7] These substitutions often occur as the result of an innovation, such as a new drug that lowers the need to use expensive therapies to treat a disease. For example, the introduction of selected serotonin re-uptake inhibitors (SSRIs) represented a new generation of antidepressants that allowed fewer more-expensive therapy visits. Disease-based price indexes report medical inflation by the treatment of disease, rather than by the good or service that treats this disease. However, they do not at this point account for improved outcomes, such as increases in life expectancy coming from an innovation such as coronary bypass surgery. Disease-based price indexes are still a work in progress, and are not yet ready to be officially published medical price indexes, which supplement the medical practice indexes that are reported on a goods and services basis.

Estimated Quality-Adjustment and New Goods Biases and Measured Real GDP Growth

The Bureau of Economic Analysis uses price indexes to decompose changes in nominal GDP growth into a portion that reflects inflation and a portion that reflects growth in real output. This use of price indexes to deflate nominal spending implies

[7] See https://www.bls.gov/pir/diseasehome.htm at the BLS website for more information on BLS experimental disease-based price indexes. See https://www.bea.gov/national/health_care_satellite_account.htm at the BEA website for more information on BEA's treatment of disease-based measures as reflected in Health Satellite Accounts.

that any bias that overstates inflation will result in a downward bias in the attendant measure of real GDP. Bias can also result from challenges in measurement of nominal GDP, which is discussed in the next section.

Here, in order to get some sense of the degree to which real GDP growth may be understated, we apply results from studies that use available empirical evidence to form judgmental assessments of biases in the price indexes. Such studies exist for indexes underlying two of the major expenditure categories of GDP that, together, make up about 85% of the GDP: personal consumption expenditures and private fixed investment. For example, the Boskin Commission concluded that, in 1996, the Consumer Price Index was likely biased by +1.1 percentage points per year, with about half of the bias attributable to problems with accounting for quality change and new goods (Boskin et al. 1997, table 3). Several years later, Lebow and Rudd (2003) estimated that in 2001, CPI growth was biased by +0.87 percentage points per year, with +0.37 percentage points of that stemming from quality change and new goods bias (Lebow and Rudd 2003, table 1). Among the reasons they cited for a lower estimate than had been reported by the Boskin Commission were the use of new empirical studies of biases in the underlying CPI components and improvements in methods implemented by the Bureau of Labor Statistics after the Boskin Commission convened. Beyond the CPI, Byrne, Fernald, and Reinsdorf (2016) used price indexes developed by Byrne and Corrado (2017) to provide a similar analysis for the effect of bias in price indexes for information and communications technology equipment on real growth of private fixed investment.

These careful studies inherently involve an element of judgment, and the investigators clearly emphasize the uncertainty surrounding their estimates. For example, Lebow and Rudd (2003) assigned a subjective 90 percent confidence interval to their point estimate of total bias in the Consumer Price Index, yielding an estimated bias that ranged from 0.3 to 1.4 percentage points. Likewise, Byrne, Fernald, and Reinsdorf (2016) provide two sets of estimates, which they label "conservative" and "liberal." Despite the uncertainty, we believe these estimates are of great value to help direct improvement efforts, inform users of data limitations, and (in the case of Byrne et al.) rule out certain hypotheses, such as a recent large increase in bias.

To consider the effect of imperfect adjustments for quality change and new goods purchased by households, we use the Lebow and Rudd (2003) assessments of biases in the individual components of the Consumer Price Index, representing the most recent comprehensive set of estimates available, many of which are used to deflate the components of personal consumption expenditures. We make adjustments to their estimates for medical care spending and internet services to reflect results available from more recent studies. For medical care, Cutler, Rosen, and Vijan (2006) studied the impact of innovations in medical care on life expectancy and found that improved mortality over the 1960–2000 period was concentrated in five conditions (see their table 2): cardiovascular disease (4.9 years), death in infancy (1.4 years), death from external causes (4 months), pneumonia and influenza (3 months), and malignant cancers (2.5 months). According to the Agency for Healthcare Research and Quality (2017), spending on these conditions totaled $363

Table 2

Impact of Estimated Biases to Personal Consumption Expenditures (PCE) Deflators on Measured Real GDP Growth, 2000–2015

Expenditure Category	*Share of GDP*				*Lebow–Rudd (2003) bias*
	2000	*2005*	*2010*	*2015*	*2003*
Medical care:					
Prescription drugs	1.3%	1.6%	1.9%	2.3%	1.20%
Nonprescription drugs	0.2%	0.2%	0.3%	0.3%	0.50%
Medical care services[a]	9.8%	10.9%	12.2%	12.5%	0.76%
PC services (including internet)[b]	0.2%	0.2%	0.4%	0.6%	6.50%
	Contributions to real GDP growth (percentage points per year)				
Medical care:					
Prescription drugs	–0.02	–0.02	–0.02	–0.03	
Nonprescription drugs	0.00	0.00	0.00	0.00	
Medical care services	–0.07	–0.08	–0.09	–0.09	
PC services (including internet)	–0.01	–0.01	–0.03	–0.04	
All other PCE categories	–0.10	–0.10	–0.10	–0.09	
All personal consumption expenditive categories	**–0.20**	**–0.22**	**–0.24**	**–0.26**	

Note: Table 2 presents our translation of the impact of bias in each segment of personal consumption expenditures on measured real GDP growth, calculated by multiplying each Lebow and Rudd (2003) bias estimate—the percentage point overstatement of price change in the individual price index per year—by the segment's share of GDP (with the adjustments noted in the text). Total for All personal consumption expenditure categories may not add up exactly to the subcomponents shown in the columns due to rounding.

[a]Bias estimate for medical care services has been adjusted based on data from AHRQ (2017).

[b]Bias estimate for PC services (including internet) is based on Greenstein and McDevitt (2011).

billion, or about 20 percent of all spending on medical care. To fold this analysis into an estimated bias, we follow Lebow and Rudd in reducing the 4.5 percentage point bias reported in Cutler et al. (1998) by 0.7 percentage point, to account for improvements to the methods used by the Bureau of Labor Statistics. However, given the new information in Cutler (2006), we apply the resulting bias (4.5 – 0.7 = 3.8) to a smaller share of spending (1/5 versus the 2/3 used in Lebow–Rudd). This translates into a +0.76 bias for medical care services overall.

Similarly, Greenstein and McDevitt (2011) estimate that price indexes for broadband access overstate price growth by 3 to 10 percentage points, and we apply the midpoint of this range to the personal consumption services (including internet) category in overall personal consumption expenditures. For the remaining categories of personal consumption expenditures, we use the Lebow and Rudd (2003) estimates, suitably reweighted.

Table 2 presents our translation of the impact of bias in each segment of personal consumption expenditures on measured real GDP growth, calculated by multiplying each Lebow and Rudd (2003) bias estimate—the percentage point overstatement of price change in the individual price index per year—by the segment's

Table 3

Impact of Estimated Biases to Private Fixed Investment (PFI) Deflators on Measured Real GDP Growth, 2000–2015

Equipment type	*Share of GDP*				*Byrne, Fernald, and Reinsdorf (2016) estimated bias*	
	2000	*2005*	*2010*	*2015*	*1995–2004*	*2004–2014*
Communication equipment	1.2%	0.7%	0.6%	0.6%	5.8%	7.6%
Computers and peripherals	1.0%	0.6%	0.5%	0.4%	8.0%	12.0%
Other information systems equipment	0.7%	0.7%	0.7%	0.8%	8.3%	5.4%
Software	1.8%	1.7%	1.7%	1.8%	1.4%	0.9%
	Contributions to real GDP growth (percentage points per year)					
Communication equipment	–0.07	–0.04	–0.03	–0.03		
Computers and peripherals	–0.08	–0.05	–0.04	–0.03		
Other information systems equipment	–0.05	–0.06	–0.06	–0.06		
Software	–0.03	–0.02	–0.02	–0.03		
All private fixed investment categories	**–0.23**	**–0.17**	**–0.16**	**–0.15**		

Note and Source: To assess the importance of quality change and new goods bias in investment goods, we use the Byrne, Fernald, and Reinsdorf (2016, table 1) estimates for biases in the deflators that the Bureau of Economic Analysis uses for information and communications technology products. The contributions to GDP growth for 2000 and 2005 are calculated using the bias estimates for 1995–2004; the contributions for 2010 and 2015 use the bias estimates for 2004–2014. Total for All private fixed investment categories may not add up exactly to the subcomponents shown in the columns due to rounding.

share of GDP (with the adjustments noted above). Note that a *positive* bias in the Consumer Price Index leads to a *negative* bias in GDP growth. To assess whether growth of problematic sectors has increased the potential bias to GDP growth, we report the resulting contributions using GDP shares for 2000, 2005, 2010, and 2015.

The estimated overall impact of the biases of consumption deflators on real GDP appears in the bottom line: measured growth in real GDP was reduced by –0.20 percentage point in 2000, with this downward bias growing only modestly over time to –0.26 percentage point in 2015. Of course, because the bias estimates that we apply are point-in-time estimates, we cannot assess here how the biases to the individual components may have changed over time, including the impact of continued improvements in measurement of the Consumer Price Index; hence, the increase in overall bias reported here only reflects the effects of changes over time in GDP shares.

To assess the importance of quality change and new goods bias in investment goods, we use the Byrne, Fernald, and Reinsdorf (2016, table 1) estimates for biases in the deflators that the Bureau of Economic Analysis uses for information and communications technology products. As shown in Table 3, they report separate estimates for biases in the pre- and post-slowdown periods, which show increases in the biases for communications equipment and computers and peripherals but declines in the biases for other information systems equipment. All told, the implied impact to measured real GDP growth from biases in these information and communications technology components of private fixed investment is small: less than 1/4 percentage point in all four years, with a decline to –0.15 percentage point in 2015.[8]

Taken together, the reduction in measured real GDP growth from biases in both personal consumption expenditures and private fixed investment would be about –0.4 percentage point in all four periods.

Challenges in Measuring Nominal GDP in the Digital Age

In addition to the concern that price indexes do not adequately capture quality changes and new goods, there is a separate concern that estimates of GDP by the Bureau of Economic Analysis ignore valuable new goods and services that aren't sold, such as internet searches or encyclopedia services that are provided essentially free to the user. In addition, changes in the way that households and firms obtain goods and services, such as an evolving ability of firms to outsource goods and services previously provided in-house, have also raised questions about whether and how these phenomena are measured in GDP. Overall, these concerns seem overstated. For most cases mentioned, either the official government statistics have always excluded the value of products similar to these because they are outside the scope of what GDP aims to measure, or the official economic statistics do actually include their value, although it is embedded in other measured market activity.

As economists tell students in every introductory economics class, GDP measures the *market* value of the goods, services, and structures produced by the nation's economy in a particular period. GDP is not designed to measure well-being or welfare: for example, it does not account for rates of poverty, crime, or literacy. Nor does GDP attempt to comprehensively measure nonmarket activity, such as household production or "free" services, digital or otherwise. Whether GDP should be measured more expansively has long been debated by economists. Indeed, Kuznets (1934), an early architect of the national economic accounts, noted the shortcomings of focusing exclusively on market activities and of excluding nonmarket activities and assets that have productive value.

Since then, a tremendous amount of research has sought to develop methods to better address the need to include nonmarket or near-market activities and to better measure economic welfare. Given the conceptual and practical difficulties of such efforts, however, several National Academy of Sciences studies on the environment (Nordhaus and Kokkelenberg 1999) and nonmarket production (Abraham and Mackie 2005), as well as the *System of National Accounts 2008* (European Commission

[8] Hatzius (2015) finds that the estimated biases in price indexes related to information and communications technology are sufficiently high and grow enough in 2005 to explain a significant part of the slowdown in measured productivity that happened at about that time. However, as pointed out in Syverson (2016) and Gordon (2015a), the assumed biases that require this result seem implausibly high.

et al. 2009) guidelines for measuring GDP, all recommend that the core GDP account concepts remain as is, while encouraging the creation of supplemental (or satellite) accounts to address other issues. For example, BEA began publishing a satellite account for household production in 2012 that provides estimates for GDP that incorporate the value of home production by households, the largest component of which is production of nonmarket services like cooking, gardening, or housework (Bridgman et al. 2012). In a similar spirit, the Stiglitz–Sen–Fitoussi commission (2009) report on expanded welfare measures suggests ways that "'classical GDP issues' can be addressed within existing GDP accounts or through an extension and improvement of measures included in existing accounts" (Jorgenson, Landefeld, and Nordhaus 2006; Landefeld, Moulton, Platt, and Villones 2010). We agree with Ahmad and Schreyer (2016) that while building alternative welfare measures can certainly be useful and informative, such measures have a different purpose than GDP statistics.

To evaluate how emerging economic phenomena are treated in the national accounts, we look separately and systematically at how the Bureau of Economic Analysis assembles three of the four expenditure components of GDP. (We omit government for brevity, since it has not been the focus of concern about productivity mismeasurement.) Specifically, we focus on some central examples of goods and services related to the digital economy, intellectual property, and globalization, and discuss how they are included, or not, in these components of GDP.

To begin, we note that consumption spending in GDP statistics consists of *purchases* of goods and services by households (families and unrelated individuals) and nonprofit institutions serving households (such as Goodwill Industries International). The Bureau of Economic Analysis derives these estimates from statistical reports and surveys, primarily from the Census Bureau but also from other government agencies, administrative and regulatory agency reports, and reports from private organizations, such as trade associations (BEA 2016a). Most new goods and services that involve market transactions are likely properly recorded in GDP. For example, purchases made on the Internet have been reported to the Census Bureau's retail trade surveys since 1999. However, there may be lags between the time when the new good is first introduced and when it is represented in the source data.

Some observers are concerned that consumption as measured by the GDP misses "free" digital services used by households. For example, Internet services such as Google search or Facebook are provided without any direct charge to users, which might suggest that they aren't included in GDP. However, most providers of these "free" Internet services charge advertisers. In GDP, ad-supported content—like television—is treated as an intermediate input, which means that its value is reflected in the value of the goods or services that are sold through advertising. Nakamura and Soloveichik (2015) find that an alternative treatment that regards ad-supported content as part of consumption in GDP would likely have only a small impact on GDP: US spending on advertising has been about 1.3 percent of nominal GDP for decades. However, content and services provided online without advertising such as Wikipedia, blogs, photo archives, and so on, will not be directly recorded in GDP, because there is no associated market transaction.

Others have raised questions about what happens when economic activity shifts between market and household production. For example, the Internet and its powerful search engines have lowered the cost of finding information. This has moved some activities into the home, like the services that used to be provided by paid travel agencies, for example. While estimates of these kinds of changes would be useful, perhaps in the household satellite account, such activities are appropriately excluded from GDP when they become a nonmarket activity; after all, if an activity does not provide income to some party, it is not part of the market activity that makes up GDP.

Next, consider business sector purchases. The business sector comprises all for-profit corporate and noncorporate private entities, a broad category that includes mutual financial institutions, private noninsured pension funds, cooperatives, nonprofit organizations that primarily serve businesses, Federal Reserve Banks, federally sponsored credit agencies, and government enterprises. From the standpoint of calculating GDP, this sector is responsible for gross private domestic investment and consists of purchases of fixed assets (structures, equipment, and intellectual property) by private businesses that contribute to production and have a useful life of more than one year, purchases of homes by households, and investment in inventories. Of course, purchases of other goods and services used by businesses for production of final goods are counted in GDP as intermediate (not final) goods.

The digital economy poses several challenges for measuring investment, particularly investment in intellectual property. That category of investment currently includes software, scientific research and development, and artistic originals. Because the intangible nature of these assets makes this type of investment especially hard to measure (Corrado, Hulten, and Sichel 2009), some notable examples of intellectual property that are currently not counted as investment in national accounts include research and development spending on social sciences and humanities and certain economic competencies (for example, as embodied in advertising, marketing, or organizational structure).

Some observers wonder what happens to GDP when services previously conducted internally by firms are moved to the Internet cloud. For example, whereas firms previously invested in their own servers and software, now they increasingly pay for cloud services from a central firm. To the extent that investments in servers and other equipment are shifted from firms to cloud service providers, this development would not cause mismeasurement of GDP. Servers are still purchased and recorded as investment, and purchased cloud services are recorded as intermediate goods.

While we think that goods and services that involve market transactions are likely properly included in total GDP spending, there are cases where the transactions can be misallocated across GDP categories. Consider, for example, the introduction of ride-sharing applications like Uber and Lyft and alternates to hotel services like Airbnb. To the extent that individuals earn income from selling their services through these platforms, income and output measures will accurately reflect the contribution of these new products to consumer spending in GDP final expenditures. But such changes still raise measurement issues. One is that to the extent that (unincorporated) individuals are using assets to provide transportation services, for example, then spending on such assets should be recorded as business investment, not consumption. That said, any distortions from this are likely small. In the case of Uber, for example, Bean (2016) estimates that this kind of misclassification for vehicles could perhaps amount to no more than 1.5 percent of business investment.

Since many US-made products are not consumed here and others used here are made abroad, GDP must include exports and subtract imports. Measurement issues also extend to this adjustment. In a globalized economy, many goods are US-designed (an investment) but manufactured abroad. This practice creates particular problems in accounting for the domestic value of intellectual property. For example, consider a smartphone that is designed in the United States, produced in an Asian country, and then purchased and imported by the US firm for final sale. The Bureau of Economic Analysis counts the wholesale value of the phone, which may include the value of the US firm's intellectual property, as an import and in final sales. Ideally, BEA would also capture the export of the intellectual property to the foreign producer on its surveys of international trade in services. However, under certain contract manufacturing arrangements, there may be no separate transaction for exports of design/software to the Asian manufacturer, thus understating exports in the national accounts. Ongoing work at the BEA and elsewhere continue to explore the potential magnitude of potential omissions like this (for example, Houseman and Mandel 2015).

All told, we believe that concerns about a downward bias on output are overstated because for most cases mentioned, either the value of these products is outside the scope of GDP or is embedded in other measured market activity.

Improving Measurement of Prices and Output

In this section, we review some of the projects underway at the Bureau of Labor Statistics and the Bureau of Economic Analysis that will improve the ability to measure changes in real output.

Initiatives at the Bureau of Labor Statistics

The Bureau of Labor Statistics continually looks at possible new data sources and for ways to expand the use of quality adjustment methods like hedonic analysis. The most frequent issue is that these alternative sources of data on prices usually lack data on characteristics of goods or on the arrival of new goods. These deficiencies can make it challenging to address the issues of adjusting for quality change and new goods that are the focus of this article.

For instance, the Billion Prices Project (discussed in this journal in Cavallo and Rigobon 2016) scrapes prices from the internet while the Adobe Digital Economy

Project (ADEP) uses price and quantity information stored by its cloud service customers. Currently, if respondents do not send electronic records and do not permit use of their application programming interface, then web-scraping is an option for which the Bureau of Labor Statistics is developing expertise. Note that many transactions prices are not available online, so it is impossible to generate an "all items" index with scraped data or with ADEP's cloud. For example, transactions-level medical, college tuition, new vehicle, and utility prices cannot be scraped, nor are they stored on a marketing cloud. Furthermore, many price-determining characteristics are missing, and highly sophisticated programming is required to identify new goods. (However, the "fixed effects window-splice" approach discussed in the price index section could provide the same results as a traditional time dummy approach.) For prices where web-scraping is viable, it can be contracted out. Indeed, Statistics New Zealand has hired PriceStats, the commercial arm of the Billion Prices Project, to do web-scraping. The other route is to develop skills in-house for a method that will address the confidentiality and informed consent responsibilities of the BLS.

A few respondents for the Consumer Price Index have offered their electronic transactions data to the Bureau of Labor Statistics. To be useful as building blocks for price indexes that adjust for quality changes, these records must contain both transaction prices and price-determining characteristics. To date, BLS receives electronically transferred data from two major national retailers and one market research firm. The data from one of the retailers is now in production. BLS continues to work with the other two respondents to obtain an adequate set of price-determining characteristics.

The Bureau of Labor Statistics began investigating use of scanner data over 20 years ago and has used it over the years to validate and diagnose current Consumer Price Index samples, but not directly for monthly production. Purchasing real-time data is expensive. Recently, BLS obtained an estimate for purchasing real-time scanner data for grocery stores from a third party, and the estimated cost would exceed CPI's current grocery store data collection costs by as much as 72 percent. It also can be challenging to adjust for new goods in this data, although computationally efficient new goods adjustment methods using scanner data are also being investigated (for example, Broda and Weinstein 2010; Redding and Weinstein 2016). Testing of these kinds of models is underway to see if the methods can be implemented under the BLS time and budget constraints. BLS has also observed other countries' uses of scanner data. Statistics New Zealand, Statistics Netherlands, and Statistics Norway have already incorporated scanner data in their indexes (Krsinich 2015; van der Grient and de Haan 2010, 2011; Johannessen 2016). However, there is a large difference between the way these countries and the United States can get their scanner data. Both New Zealand and Norway receive such data directly from the outlets, while the BLS would have to purchase scanner data from a private vendor. Hence the higher cost to BLS mentioned above.

When it comes to quality changes and new goods, the medical care sector poses particular challenges. Thus, the Bureau of Economic Analysis and the Bureau of

Labor Statistics are working on quality-adjusting the medical price indexes on a disease-by-disease basis discussed earlier. For each disease, BLS plans to use Medicare claims data to follow both the illnesses and the treatments of the same patients over time. The output measure is the number of days that a patient survives treatment without being readmitted to an inpatient hospital or contracting additional illnesses (like infections). This method closely follows the Romley, Goldman, and Sood (2015) measure of the output of inpatient hospital services, although BLS will instead focus on the output of treatment bundles, similar to the approach used in Berndt et al. (2002).

The Bureau of Labor Statistics continues to expand use of hedonics. In the Producer Price Index, for example, hedonic methods are now used to quality-adjust broadband services. Further hedonic adjustments are under investigation for women's dresses and network switches. While the PPI currently uses hedonic adjustment for computers, improvements are being investigated for microprocessors where new cross-validation methods will allow a more parsimonious use of parameters. The added method of cross-validation is important because the sample sizes used to compute the hedonic regression are usually small. In addition, possible approaches to implementing imputed hedonic price index methods are under consideration (for example, Pakes 2003; Erickson and Pakes 2011). BLS will investigate whether new cross-validation and machine-learning techniques will allow estimation of hedonic price indexes with current sample sizes.

Initiatives at the Bureau of Economic Analysis

The Bureau of Economic Analysis has a number of initiatives underway to address challenges related to measuring GDP in an economy characterized by digital technologies and commerce, as well as by global value chains.

Regarding consumer spending, the Bureau of Economic Analysis will continue its research on ways to improve the treatment of advertising-supported media, acknowledging that the Internet has fundamentally changed the way households consume entertainment services. BEA will seek to improve consumer spending estimates to reflect the growing importance of e-commerce. As online purchases represent an increasing share of consumer spending, BEA will update its data sources and methods to capture more information about these purchases.

With respect to digital technologies and commerce, the Bureau of Economic Analysis plans to publish a roadmap that outlines the research needed to more accurately measure the contribution of information technology to economic growth and to improve the measurement of digitally enabled commerce. In addition, BEA has begun publishing an annual report that examines trends in services for which digital technologies likely play an important role in facilitating trade, such as telecommunications, insurance, and financial services (Grimm 2016; Bureau of Economic Analysis 2016b).

Concerning issues raised by globalization and trade, the Bureau of Economic Analysis will be expanding coverage to include intellectual property transactions, which will provide a more complete picture of foreign trade in computer, audio-visual media, and research and development services. Also, BEA plans to take several steps to better measure the value that is created at various steps in such production chains: for example, it will to continue to update and refine its supply-use tables, including the production of extended supply-use tables that introduced firm-level characteristics such as ownership type and multinational status, allowing for a more refined analysis of global value chains.

Finally, the Bureau of Economic Analysis has embarked on several initiatives with statistical agency partners to leverage alternative data sources to improve the measurement of high-tech goods and services prices. For example, with regard to software, which accounts for half of investment in information and communications technology goods, BEA has purchased three types of data for potential construction of price indexes: scanner data for consumer titles, administrative data for commercial applications, and data to shed light on custom software development. Regarding cell phones and wireless plans, which pose thorny measurement challenges, BEA has obtained survey data that can potentially be used to construct price indexes for both phone and wireless services; the data contain information on households' annual payments, whether the payment includes the phone, specific features of the cell phones, or features of the plans. In addition, BEA is also supporting academic researchers who are exploring how best to measure prices for cloud computing services.

The task of calculating price indexes and output in the 21st century, and doing so in a way that provides timely monthly data within budget constraints, is not for the rigid or the fainthearted. The Bureau of Labor Statistics and Bureau of Economic Analyisis agree that price index mismeasurement continues to lead to understated growth in real output over time, perhaps especially in healthcare but also possibly in areas related to information and communications technology. Although rapid innovation and globalization present numerous measurement challenges, we are optimistic that they can be addressed. Government statistical agencies and academic researchers have successfully, if often incrementally, overcome many previous problems that once seemed intractable, whether in nominal GDP, price indexes, or elsewhere. Statistics are always estimates; they will never be perfect. Yet official economic statistics particularly possess a unique combination of accuracy, objectivity, relevance, timeliness, and accessibility that serve as infrastructure in support of efficient markets. They are essential to help policymakers and citizens form opinions and make decisions.

■ *The authors thank Alyssa Holdren for excellent research assistance; Dee Bathgate, Rob Cage, Jeff Hill, Crystal Konny, John Layng, Bonnie Murphy, and Bill Thompson for information regarding the handling of quality adjustment and the introduction of new goods in BLS price indexes; and Abe Dunn, Lucy Eldridge, David Lebow, Jeremy Rudd, Jon Samuels, and Jay Stewart for very useful comments.*

References

Abraham, Katharine G. 2003. "Toward a Cost-of-Living Index: Progress and Prospects." *Journal of Economic Perspectives* 17(1): 45–58.

Abraham, Katharine G., and Christopher Mackie, eds. 2005. *Beyond the Market: Designing Nonmarket Accounts for the United States.* Washington, DC: National Academies Press.

Agency for Healthcare Research and Quality. (AHRQ) 2017. Total Expenses and Percent Distribution for Selected Conditions by Type of Service: United States, 2014. Medical Expenditure Panel Survey Household Component Data. Generated interactively on March 21, 2017.

Ahmad, Nadim, and Paul Schreyer. 2016. "Measuring GDP in a Digitalised Economy." OECD Statistics Working Papers 2016/07.

Aizcorbe, Ana. 2014. *A Practical Guide to Price Index and Hedonic Techniques.* Oxford University Press.

Aizcorbe, Ana, and Nicole Nestoriak. 2011. "Changing Mix of Medical Care Services: Stylized Facts and Implications for Price Indexes." *Journal of Health Economics* 30(3): 568–74.

Armknecht, Paul A., Walter Lane, and Ken Stewart. 1996. "New Products and the U.S. Consumer Price Index." Chap. 9 in *The Economics of New Goods*, 373–96. National Bureau of Economic Research.

Bathgate, Deanna. 2011. "Mini-Presentation by David Friedman: The 26th Voorburg Group Meeting on Service Statistics, Newport, U.K., 19–23 September 2011." http://www4.statcan.ca/english/voorburg/Documents/2011%20 Newport/Papers/2011%20-%2036.pdf.

Bean, Charles. 2016. *Independent Review of UK Economic Statistics.* Cabinet Office, United Kingdom.

Berndt, Ernst R., Anupa Bir, Susan H. Busch, Richard G. Frank, and Sharon-Lise Normand. 2002. "The Medical Treatment of Depression, 1991–1996: Productive Inefficiency, Expected Outcome Variations, and Price Indexes." *Journal of Health Economics* 21(3): 373–96.

Berndt, Ernst R., Zvi Griliches, and Neil J. Rappaport. 1995. "Econometric Estimates of Price Indexes for Personal Computers in the 1990's." *Journal of Econometrics* 68(1): 243–68.

Berry, Steven T. 1994. "Estimating Discrete-Choice Models of Product Differentiation." *RAND Journal of Economics* 25(2): 242–62.

Berry, Steven, James Levinsohn, and Ariel Pakes. 1995. "Automobile Prices in Market Equilibrium." *Econometrica* 63(4): 841–90.

Boskin, Michael, Ellen R. Dulberger, Robert J. Gordon, Zvi Griliches, and Dale W. Jorgenson. 1997. "The CPI Commission: Findings and Recommendations." *American Economic Review* 87(2): 78–83.

Bradley, Ralph. 2005. "Pitfalls of Using Unit Values as a Price Measure or Price Index." *Journal of Economic and Social Measurement* 30(1): 39–61.

Bradley, Ralph. 2013. "Feasible Methods to Estimate Disease Based Price Indexes." *Journal of Health Economics* 32(3): 504–514.

Bresnahan, Timothy. 1997. Comment on "Valuation of New Goods under Perfect and Imperfect Competition," by Jerry A. Hausman. In chap. 5 of *The Economics of New Goods*, edited by Timothy F. Bresnahan and Robert J. Gordon. University of Chicago Press.

Bridgmen, Benjamin, Andrew Dugan, Mikhael Lal, Matthew Osborne, and Shaunda Villones. 2012. "Accounting for Household Production in the National Accounts, 1965–2010." *Survey of Current Business*, May.

Broda, Christian, and David E. Weinstein. 2010. "Product Creation and Destruction: Evidence and Price Implications." *American Economic Review* 100(3): 691–723.

Bureau of Economic Analysis. 2015. "Measuring the Economy: A Primer on GDP and the National Income and Product Accounts." December. Available at: https://bea.gov/methodologies/index.htm.

Bureau of Economic Analysis. 2016a. "Concepts and Methods of the U.S. National Income and Product Accounts." https://www.bea.gov/national/pdf/allchapters.pdf.

Bureau of Economic Analysis. 2016b. "U.S. International Services: Trade in Services in 2015 and Services Supplied Through Affiliates in 2014." *Survey of Current Business*, December.

Bureau of Labor Statistics (BLS). 2008. "Proposal for Adjusting the General Hospital Producer Price Index for Quality Change." February 15. http://conference.nber.org/confer/2008/si2008/PRCR/murphy2.pdf.

Bureau of Labor Statistics (BLS). 2016. *Handbook of Methods.* Division of BLS Publishing, Office of Publications and Special Studies.

Byrne, David M., and Carol Corrado. 2017. "ICT Prices and ICT Services: What Do They Tell Us about Productivity and Technology?" Finance and Economics Discussion Paper 2017-015, Federal Reserve Board, May.

Byrne, David M., John G. Fernald, and Marshall B. Reinsdorf. 2016. "Does the United States have a Productivity Slowdown or a Measurement

Problem?" *Brookings Paper on Economic Activity*, Spring. https://www.brookings.edu/bpea-articles/does-the-united-states-have-a-productivity-slowdown-or-a-measurement-problem/.

Byrne, David M., Stephen D. Oliner, and Daniel E. Sichel. 2015. "How Fast are Semiconductor Prices Falling?" NBER Working Paper 21074.

Cavallo, Alberto, and Roberto Rigobon. 2016. "The Billion Prices Project: Using Online Prices for Measurement and Research." *Journal of Economic Perspectives* 30(2): 151–78.

Corrado, Carol, Charles Hulten, and Daniel Sichel. 2009. "Intangible Capital and U.S. Economic Growth." *Review of Income and Wealth* 55(3): 661–85.

Court, Andrew T. 1939. "Hedonic Price Indexes with Automotive Examples. In *The Dynamics of Automotive Demand*, edited by Charles F. Roos, 99–117. New York: General Motors Corporation.

Cutler, David M., Mark McClellan, Joseph P. Newhouse, and Dahlia Remler. 1998. "Are Medical Prices Declining? Evidence from Heart Attack Treatments." *Quarterly Journal of Economics* 113(4): 991–1024.

Cutler, David M., Allison B. Rosen, and Sandeep Vijan. 2006. "The Value of Medical Spending in the United States, 1960–2000." *New England Journal of Medicine* 355(9): 920–27.

de Haan, Jan, and Haymerik A. van der Grient. 2011. "Eliminating Chain Drift in Price Indexes Based on Scanner Data." *Journal of Econometrics* 161(1): 36–46.

de Haan, Jan, and Frances Krsinich. 2014. "Scanner Data and the Treatment of Quality Change in Non-Revisable Price Indexes." *Journal of Business & Economic Statistics* 32(3): 341–58.

Eltetö, O., and P. Köves. 1964. "On a Problem of Index Number Computation Relating to International Comparisons." *Statisztikai Szemle* 42: 507–518 (in Hungarian).

Erickson, Timothy, and Ariel Pakes. 2011. "An Experimental Component Index for the CPI: From Annual Computer Data to Monthly Data on Other Goods." *American Economic Review* 101(5): 1707–38.

European Commission, International Monetary Fund, the Organiation for Economic Co-operation and Development, the United Nations, and the World Bank. 2009. *System of National Accounts, 2008.* http://unstats.un.org/unsd/nationalaccount/docs/SNA2008.pdf.

Fixler, Dennis, Charles Fortuna, John Greenlees, and Walter Lane. 1999. "The Use of Hedonic Regressions to Handle Quality Change: The Experience in the U.S. CPI." US Bureau of Labor Statistics, Presented at the Fifth Meeting of the International Working Group on Price Indices, Reykjavik, Iceland.

Gini, C. 1931. "On the Circular Test of Index Numbers." *Metron* 9(2): 3–24.

Gordon, Robert J. 2015a. "Productivity, Prices, and Measurement." Presentation at BoskinFest, Stanford University.

Gordon, Robert J. 2015b. "The Sources of Slowing Growth in Productivity Growth and Potential Output." Presentation at the Philadelphia Fed Policy Forum, December 4.

Greenstein, Shane, and Ryan McDevitt. 2011. "Evidence of a Modest Price Decline in US Broadband Services." *Information Economics and Policy* 23(2): 200–211.

Griliches, Zvi. 1961. "Hedonic Prices for Automobiles: An Econometric Analysis of Quality Change." *The Price Statistics of the Federal Government, General Series No. 73*, pp, 137–96. Columbia University Press for the National Bureau of Economic Research, New York.

Grimm, Alexis N. 2016. "Trends in U.S. Trade in Information and Communications Technology (ICT) Services and in ICT-Enabled Services." *Survey of Current Business*, May, 96(5).

Hatzius, Jan. 2015. "Productivity Paradox 2.0." *Top of Mind*, Issue 39, p. 6–7. Goldman Sachs. http://www.goldmansachs.com/our-thinking/pages/macroeconomic-insights-folder/the-productivity-paradox/report.pdf.

Hausman, Jerry. 1997. "Valuation of New Goods under Perfect and Imperfect Competition." Chapter 4 in *The Economics of New Goods*, edited by Timothy F. Bresnahan and Robert J. Gordon.

Hausman, Jerry. 2003. "Sources of Bias and Solutions to Bias in the Consumer Price Index." *Journal of Economic Perspectives* 17(1): 23–44.

Hicks, John R. 1940. "The Valuation of the Social Income." *Economica* 7(26): 105–124.

Houseman, Susan, Christopher Kurz, Paul Lengermann, and Benjamin Mandel. 2011. "Offshoring Bias in U.S. Manufacturing." *Journal of Economic Perspectives* 25(2): 111–32.

Houseman, Susan N., and Michael Mandel, eds. 2015. *Measuring Globalization: Better Trade Statistics for Better Policy*, vols. 1 and 2. Kalamazoo, MI: W.E. Upjohn Institute for Employment Research.

International Monetary Fund. 2004. *Producer Price Index Manual: Theory and Practice.*

Ivancic, Lorraine, Walter Erwin Diewert, and Kevin J. Fox. 2009. "Scanner Data, Time Aggregation and the Construction of Price Indexes." Discussion Paper 09-09, Department of Economics, University of British Columbia, Vancouver, Canada.

Johannessen, Randi. 2016. "Scanner Data in CPI/HICP." Presentation at ESS Modernization Workshop in Bucharest, Romania, March 16–17, 2016. https://ec.europa.eu/eurostat/

cros/system/files/randi_johansenn_the_use_of_scanner_data_in_the_norwegian_cpi.pdf.

Jorgenson, Dale W., Steven Landefeld, and William Nordhaus. 2006. *A New Architecture for the U.S. National Accounts.* University of Chicago Press.

Konüs, Alexander. 1939. "The Problem of the True Index of the Cost of Living." *Econometrica* 7(1): 10–29.

Krsinich, Frances. 2015. "Implementation of Consumer Electronics Scanner Data in the New Zealand CPI." Paper presented at the Ottawa Group on Price Indices, May 20–22, Tokyo, Japan.

Krsinich, Frances. 2016. "The FEWS Index: Fixed Effects with a Window Splice." *Journal of Official Statistics* 32(2): 375–404.

Kuznets, Simon. 1934. "National Income 1929–1932." Senate Document No. 124, 73rd Congress, 2nd Session. Washington, DC: US Government Printing Office.

Lancaster, Kelvin J. 1966. "A New Approach to Consumer Theory." *Journal of Political Economy* 74(2): 132–57.

Landefeld, J. Steven, Brent R. Moulton, Joel D. Platt, and Shaunda M. Villones. 2010. "GDP and Beyond: Measuring Economic Progress and Sustainability." *Survey of Current Business*, April.

Landefeld, J. Steven, Eugene P. Seskin, and Barbara M. Fraumeni. 2008. "Taking the Pulse of the Economy: Measuring GDP." *Journal of Economic Perspectives* 22(2): 193–216.

Lebow, David, and Jeremy B. Rudd. 2003. "Measurement Error in the Consumer Price Index: Where Do We Stand?" *Journal of Economic Literature* 41(1): 159–201.

Lee, Lung-Fei, and Mark M. Pitt. 1986. "Microeconometric Demand System with Binding Nonnegativity Constraints: The Dual Approach." *Econometrica* 54(5): 1237–42.

Liegey, Paul. R., Jr. 2001a. "Developing a Hedonic Regression Model for DVD Players in the U.S. CPI." http://www.bls.gov/cpi/cpidvd.htm.

Liegey, Paul. R., Jr. 2001b. "Hedonic Quality Adjustment Methods for Microwave Ovens in the U.S. CPI." http://www.bls.gov/cpi/cpimwo.htm.

Liegey, Paul. R., Jr. 2003. "Hedonic Quality Adjustment Methods for Clothes Dryers in the U.S. CPI." http://www.bls.gov/cpi/cpidryer.htm.

Liegey, Paul. R., Jr. 1993. "Adjusting Apparel Indexes in the Consumer Price Index for Quality Differences." Chap. 6 in *Price Measurements and Their Uses,* edited by Murray F. Foss, Marilyn Manser, and Allan H. Young. University of Chicago Press.

Liegey, Paul. R., Jr. 1994. "Apparel Price Indexes: Effects of Hedonic Adjustment." *Monthly Labor Review* 117(5): 38–45. http://www.bls.gov/opub/mlr/1994/05/art6full.pdf.

McFadden, Daniel. 1978. "Modelling the Choice of Residential Location." In *Spatial Interaction Theory and Planning Models*, edited by A. Karlqvist et al., pp. 75–96. Amsterdam: North Holland.

Moulton, Brent R., Timothy J. LaFleur, and Karin E. Moses. 1998. "Research on Improved Quality Adjustment in The CPI: The Case of Televisions." Bureau of Labor Statistics Working Paper. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.506.3183&rep=rep1&type=pdf.

Murphy Bonnie H., Michael Holdway, John L. Lucier, Jason Carnival, Elizabeth Garabis, and Elaine Cardenas. 2008. "Proposal for Adjusting the General Hospital Producer Price Index for Quality Change." http://conference.nber.org/confer/2008/si2008/PRCR/murphy2.pdf.

Nakamura, Leonard I., and Rachel H. Soloveichik. 2015. "Valuing 'Free' Media across Countries in GDP." Available at SSRN: http://papers.ssrn.com/abstract=2631621.

National Research Council. 2002. *At What Price? Conceptualizing and Measuring Cost-of-Living and Price Indexes.* Panel on Conceptual, Measurement, and Other Statistical Issues in Developing Cost-of-Living Indexes, Charles L. Schultze and Christopher Mackie, Editors. Committee on National Statistics, Division of Behavioral and Social Sciences and Education. Washington, DC: National Academy Press.

Nevo, Aviv. 2000. "A Practitioner's Guide to Estimation of Random-Coefficients Logit Models of Demand." *Journal of Economics & Management Strategy* 9(4): 513–548.

Nevo, Aviv. 2001. "Measuring Market Power in the Ready-to-Eat Cereal Industry." *Econometrica* 69(2): 307–42.

Nevo, Aviv. 2003. "New Products, Quality Changes, and Welfare Measures Computed from Estimated Demand Systems." *Review of Economics and Statistics* 85(2): 266–75.

Nordhaus, William D., and Edward C. Kokkelenberg, eds. 1999. *Nature's Numbers: Expanding the National Economic Accounts to Include the Environment.* National Academies Press.

Pakes, Ariel. 2003. "A Reconsideration of Hedonic Price Indexes with an Application to PC's." *American Economic Review* 93(5): 1578–96.

Petrin, Amil. 2002. "Quantifying the Benefits of New Products: The Case of the Minivan." *Journal of Political Economy* 110(4): 705–29.

Redding, Stephen J., and David E. Weinstein. 2016. "A Unified Approach to Estimating Demand and Welfare." NBER Working Paper 22479.

Romley, John A., Dana P. Goldman, and Neeraj Sood. 2015. "US Hospitals Experienced Substantial Productivity Growth during 2002–11." *Health Affairs* 34(3): 511–18.

Rosen, Sherwin. 1974. “Hedonic Prices and Implicit Markets: Product Differentiation in Pure Competition.” *Journal of Political Economy* 82(1): 34–55.

Shapiro, Irving, Matthew D. Shapiro, and David W. Wilcox. 2001. “Measuring the Value of Cataract Surgery.” In *Medical Care Output and Productivity*, edited by David M. Cutler and Ernst R. Berndt, 411–437. *Studies in Income and Wealth*, vol. 62. University of Chicago Press.

Silver, Mick, and Heravi, Saeed. 2007. “The Difference between Hedonic Imputation Indexes and Time Dummy Hedonic Indexes.” *Journal of Business and Economic Statistics* 25(2): 239–46.

Stiglitz, Joseph E., Amartya Sen, and Jean-Paul Fitoussi. 2009. *Report by the Commission on the Measurement of Economic Performance and Social Progress.* United Nations Press.

Syverson, Chad. 2016. “Challenges to Mismeasurement Explanations for the U.S. Productivity Slowdown.” NBER Working Paper 21974.

Szulc, B. 1964. “Indices for Multiregional Comparisons.” *Przeglad Statystycny* 3: 239–254 (in Polish).

Triplett, Jack E. 1982. “Concepts of Quality in Input and Output Price Measures: A Resolution of the User-Value Resource-Cost Debate.” In *The U.S. National Income and Product Accounts: Selected Topics*, edited by Murray F. Foss, 269–312. University of Chicago Press.

van der Grient, Heymerik A., and Jan de Haan. 2010. “The Use of Supermarket Scanner Data in the Dutch CPI.” Paper presented at the Joint ECE/ILO Workshop on Scanner Data, May 10, 2010. Geneva.

van der Grient, Heymerik A., and Jan de Haan. 2011. “Scanner Data Price Indexes: The ‘Dutch Method’ versus RYGEKS.” Paper presented at the Twelfth Meeting of the International Working Group on Price Indices (Ottawa Group).

Journal of Economic Perspectives—Volume 31, Number 2—Spring 2017—Pages 211–236

Social Media and Fake News in the 2016 Election

Hunt Allcott and Matthew Gentzkow

American democracy has been repeatedly buffeted by changes in media technology. In the 19th century, cheap newsprint and improved presses allowed partisan newspapers to expand their reach dramatically. Many have argued that the effectiveness of the press as a check on power was significantly compromised as a result (for example, Kaplan 2002). In the 20th century, as radio and then television became dominant, observers worried that these new platforms would reduce substantive policy debates to sound bites, privilege charismatic or "telegenic" candidates over those who might have more ability to lead but are less polished, and concentrate power in the hands of a few large corporations (Lang and Lang 2002; Bagdikian 1983). In the early 2000s, the growth of online news prompted a new set of concerns, among them that excess diversity of viewpoints would make it easier for like-minded citizens to form "echo chambers" or "filter bubbles" where they would be insulated from contrary perspectives (Sunstein 2001a, b, 2007; Pariser 2011). Most recently, the focus of concern has shifted to social media. Social media platforms such as Facebook have a dramatically different structure than previous media technologies. Content can be relayed among users with no significant third party filtering, fact-checking, or editorial judgment. An individual user with no track record or reputation can in some cases reach as many readers as Fox News, CNN, or the *New York Times*.

■ *Hunt Allcott is Associate Professor of Economics, New York University, New York City, New York. Matthew Gentzkow is Professor of Economics, Stanford University, Stanford, California. Both authors are Research Associates, National Bureau of Economic Research, Cambridge, Massachusetts.*

† For supplementary materials such as appendices, datasets, and author disclosure statements, see the article page at
https://doi.org/10.1257/jep.31.2.211 doi=10.1257/jep.31.2.211

Following the 2016 election, a specific concern has been the effect of false stories—"fake news," as it has been dubbed—circulated on social media. Recent evidence shows that: 1) 62 percent of US adults get news on social media (Gottfried and Shearer 2016); 2) the most popular fake news stories were more widely shared on Facebook than the most popular mainstream news stories (Silverman 2016); 3) many people who see fake news stories report that they believe them (Silverman and Singer-Vine 2016); and 4) the most discussed fake news stories tended to favor Donald Trump over Hillary Clinton (Silverman 2016). Putting these facts together, a number of commentators have suggested that Donald Trump would not have been elected president were it not for the influence of fake news (for examples, see Parkinson 2016; Read 2016; Dewey 2016).

Our goal in this paper is to offer theoretical and empirical background to frame this debate. We begin by discussing the economics of fake news. We sketch a model of media markets in which firms gather and sell signals of a true state of the world to consumers who benefit from inferring that state. We conceptualize fake news as distorted signals uncorrelated with the truth. Fake news arises in equilibrium because it is cheaper to provide than precise signals, because consumers cannot costlessly infer accuracy, and because consumers may enjoy partisan news. Fake news may generate utility for some consumers, but it also imposes private and social costs by making it more difficult for consumers to infer the true state of the world—for example, by making it more difficult for voters to infer which electoral candidate they prefer.

We then present new data on the consumption of fake news prior to the election. We draw on web browsing data, a new 1,200-person post-election online survey, and a database of 156 election-related news stories that were categorized as false by leading fact-checking websites in the three months before the election.

First, we discuss the importance of social media relative to sources of political news and information. Referrals from social media accounted for a small share of traffic on mainstream news sites, but a much larger share for fake news sites. Trust in information accessed through social media is lower than trust in traditional outlets. In our survey, only 14 percent of American adults viewed social media as their "most important" source of election news.

Second, we confirm that fake news was both widely shared and heavily tilted in favor of Donald Trump. Our database contains 115 pro-Trump fake stories that were shared on Facebook a total of 30 million times, and 41 pro-Clinton fake stories shared a total of 7.6 million times.

Third, we provide several benchmarks of the rate at which voters were exposed to fake news. The upper end of previously reported statistics for the ratio of page visits to shares of stories on social media would suggest that the 38 million shares of fake news in our database translates into 760 million instances of a user clicking through and reading a fake news story, or about three stories read per American adult. A list of fake news websites, on which just over half of articles appear to be false, received 159 million visits during the month of the election, or 0.64 per US adult. In our post-election survey, about 15 percent of respondents recalled seeing each of 14 major pre-election fake news headlines, but about 14 percent also recalled seeing a set of placebo fake news headlines—untrue headlines that we invented and that never actually circulated. Using the difference between fake news headlines and placebo headlines as a measure of true recall and projecting this to the universe of fake news articles in our database, we estimate that the average adult saw and remembered 1.14 fake stories. Taken together, these estimates suggest that the average US adult might have seen perhaps one or several news stories in the months before the election.

Fourth, we study inference about true versus false news headlines in our survey data. Education, age, and total media consumption are strongly associated with more accurate beliefs about whether headlines are true or false. Democrats and Republicans are both about 15 percent more likely to believe ideologically aligned headlines, and this ideologically aligned inference is substantially stronger for people with ideologically segregated social media networks.

We conclude by discussing the possible impacts of fake news on voting patterns in the 2016 election and potential steps that could be taken to reduce any negative impacts of fake news. Although the term "fake news" has been popularized only recently, this and other related topics have been extensively covered by academic literatures in economics, psychology, political science, and computer science. See Flynn, Nyhan, and Reifler (2017) for a recent overview of political misperceptions. In addition to the articles we cite below, there are large literatures on how new information affects political beliefs (for example, Berinsky 2017; DiFonzo and Bordia 2007; Taber and Lodge 2006; Nyhan, Reifler, and Ubel 2013; Nyhan, Reifler, Richey, and Freed 2014), how rumors propagate (for example, Friggeri, Adamic, Eckles, and Cheng 2014), effects of media exposure (for example, Bartels 1993, DellaVigna and Kaplan 2007, Enikolopov, Petrova, and Zhuravskaya 2011, Gerber and Green 2000, Gerber, Gimpel, Green, and Shaw 2011, Huber and Arceneaux 2007, Martin and Yurukoglu 2014, and Spenkuch and Toniatti 2016; and for overviews, DellaVigna and Gentzkow 2010, and Napoli 2014), and ideological segregation in news consumption (for example, Bakshy, Messing, and Adamic 2015; Gentzkow and Shapiro 2011; Flaxman, Goel, and Rao 2016).

Background: The Market for Fake News

Definition and History

We define "fake news" to be news articles that are intentionally and verifiably false, and could mislead readers. We focus on fake news articles that have political implications, with special attention to the 2016 US presidential elections. Our definition includes intentionally fabricated news articles, such as a widely shared article from the now-defunct website denverguardian.com with the headline, "FBI agent suspected in Hillary email leaks found dead in apparent murder-suicide." It also includes many articles that originate on satirical websites but could be misunderstood as factual, especially when viewed in isolation on Twitter or Facebook feeds. For example, in July 2016, the now-defunct website wtoe5news.com reported that

Pope Francis had endorsed Donald Trump's presidential candidacy. The WTOE 5 News "About" page disclosed that it is "a fantasy news website. Most articles on wtoe-5news.com are satire or pure fantasy," but this disclaimer was not included in the article. The story was shared more than one million times on Facebook, and some people in our survey described below reported believing the headline.

Our definition rules out several close cousins of fake news: 1) unintentional reporting mistakes, such as a recent incorrect report that Donald Trump had removed a bust of Martin Luther King Jr. from the Oval Office in the White House; 2) rumors that do not originate from a particular news article;[1] 3) conspiracy theories (these are, by definition, difficult to verify as true or false, and they are typically originated by people who believe them to be true);[2] 4) satire that is unlikely to be misconstrued as factual; 5) false statements by politicians; and 6) reports that are slanted or misleading but not outright false (in the language of Gentzkow, Shapiro, and Stone 2016, fake news is "distortion," not "filtering").

Fake news and its cousins are not new. One historical example is the "Great Moon Hoax" of 1835, in which the *New York Sun* published a series of articles about the discovery of life on the moon. A more recent example is the 2006 "Flemish Secession Hoax," in which a Belgian public television station reported that the Flemish parliament had declared independence from Belgium, a report that a large number of viewers misunderstood as true. Supermarket tabloids such as the *National Enquirer* and the *Weekly World News* have long trafficked in a mix of partially true and outright false stories.

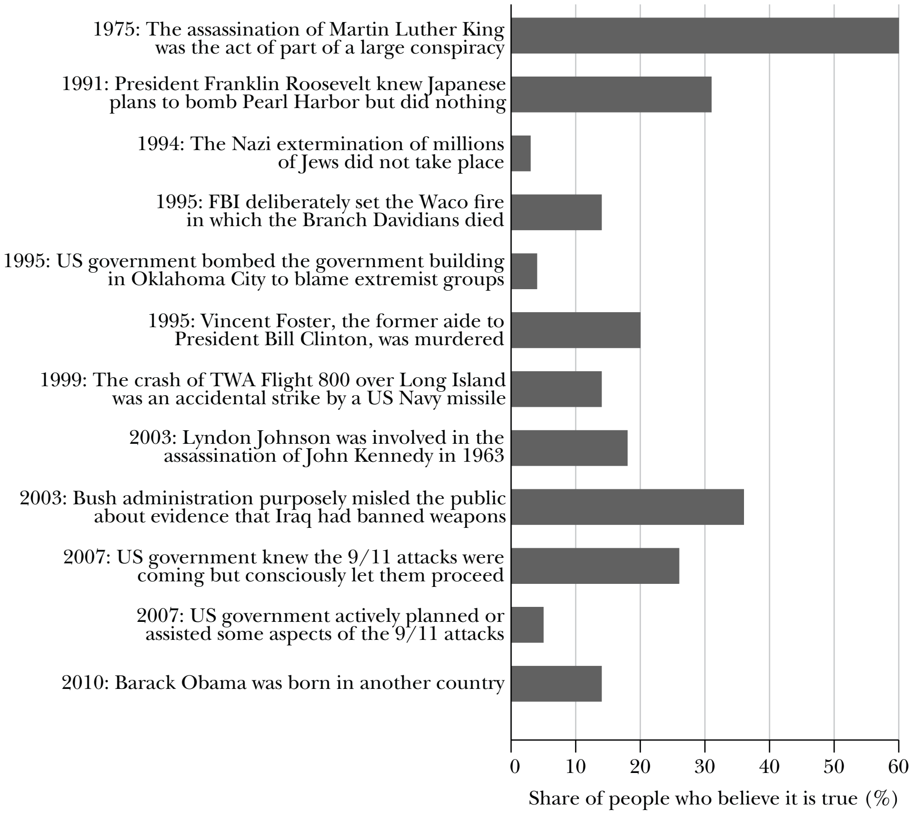
Figure 1 lists 12 conspiracy theories with political implications that have circulated over the past half-century. Using polling data compiled by the American Enterprise Institute (2013), this figure plots the share of people who believed each statement is true, from polls conducted in the listed year. For example, substantial minorities of Americans believed at various times that Franklin Roosevelt had prior knowledge of the Pearl Harbor bombing, that Lyndon Johnson was involved in the Kennedy assassination, that the US government actively participated in the 9/11 bombings, and that Barack Obama was born in another country.

The long history of fake news notwithstanding, there are several reasons to think that fake news is of growing importance. First, barriers to entry in the media industry have dropped precipitously, both because it is now easy to set up websites and because it is easy to monetize web content through advertising platforms. Because reputational concerns discourage mass media outlets from knowingly reporting false stories, higher entry barriers limit false reporting. Second, as we discuss below, social media are well-suited for fake news dissemination, and social

[1] Sunstein (2007) defines rumors as "claims of fact—about people, groups, events, and institutions—that have not been shown to be true, but that move from one person to another, and hence have credibility not because direct evidence is available to support them, but because other people seem to believe them."

[2] Keeley (1999) defines a conspiracy theory as "a proposed explanation of some historical event (or events) in terms of the significant causal agency of a relatively small group of persons—the conspirators—acting in secret."

Figure 1
Share of Americans Believing Historical Partisan Conspiracy Theories

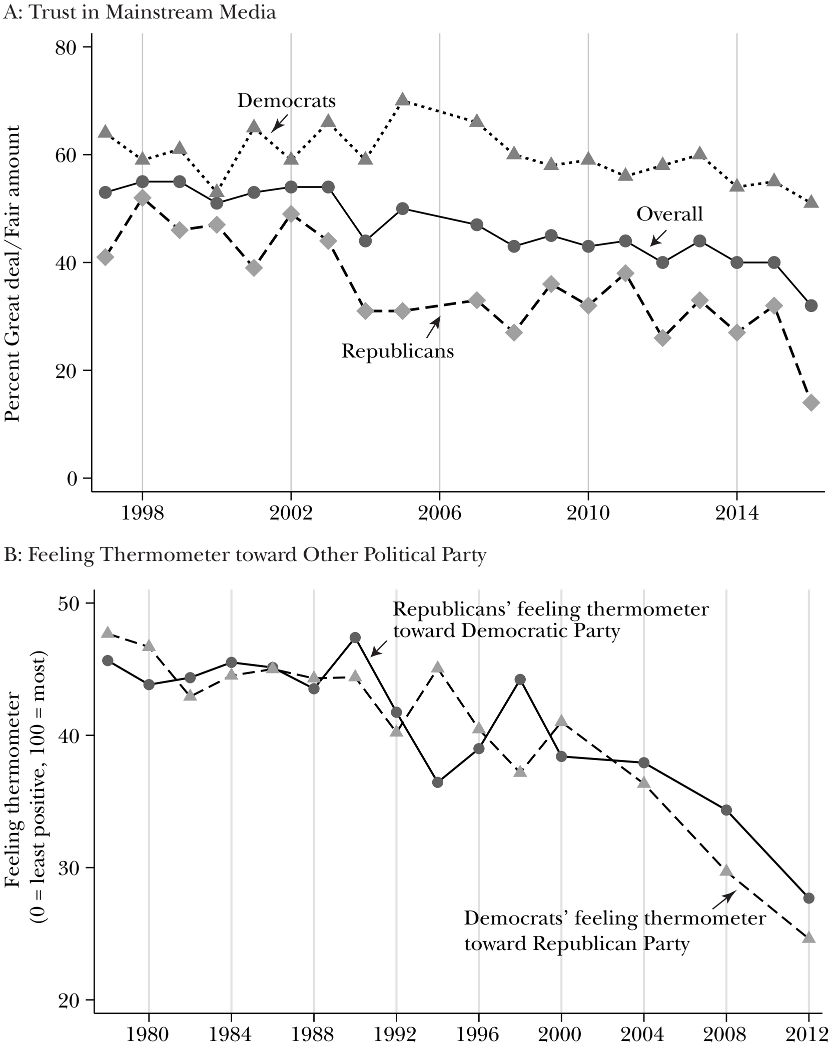


Note: From polling data compiled by the American Enterprise Institute (2013), we selected all conspiracy theories with political implications. This figure plots the share of people who report believing the statement listed, using opinion polls from the date listed.

media use has risen sharply: in 2016, active Facebook users per month reached 1.8 billion and Twitter's approached 400 million. Third, as shown in Figure 2A, Gallup polls reveal a continuing decline of "trust and confidence" in the mass media "when it comes to reporting the news fully, accurately, and fairly." This decline is more marked among Republicans than Democrats, and there is a particularly sharp drop among Republicans in 2016. The declining trust in mainstream media could be both a cause and a consequence of fake news gaining more traction. Fourth, Figure 2B shows one measure of the rise of political polarization: the increasingly negative feelings each side of the political spectrum holds toward the other.[3] As we

[3] The extent to which polarization of voters has increased, along with the extent to which it has been driven by shifts in attitudes on the right or the left or both, are widely debated topics. See Abramowitz and Saunders (2008), Fiorina and Abrams (2008), Prior (2013), and Lelkes (2016) for reviews.

Figure 2
Trends Related to Fake News



A: Trust in Mainstream Media

B: Feeling Thermometer toward Other Political Party

Note: Panel A shows the percent of Americans who say that they have "a great deal" or "a fair amount" of "trust and confidence" in the mass media "when it comes to reporting the news fully, accurately, and fairly," using Gallup poll data reported in Swift (2016). Panel B shows the average "feeling thermometer" (with 100 meaning "very warm or favorable feeling" and 0 meaning "very cold or unfavorable feeling") of Republicans toward the Democratic Party and of Democrats toward the Republican Party, using data from the American National Election Studies (2012).

discuss below, this could affect how likely each side is to believe negative fake news stories about the other.

Who Produces Fake News?

Fake news articles originate on several types of websites. For example, some sites are established entirely to print intentionally fabricated and misleading articles, such as the above example of denverguardian.com. The names of these sites are often chosen to resemble those of legitimate news organizations. Other satirical sites contain articles that might be interpreted as factual when seen out of context, such as the above example of wtoe5news.com. Still other sites, such as endingthefed.com, print a mix between factual articles, often with a partisan slant, along with some false articles. Websites supplying fake news tend to be short-lived, and many that were important in the run-up to the 2016 election no longer exist.

Anecdotal reports that have emerged following the 2016 election provide a partial picture of the providers behind these sites. Separate investigations by BuzzFeed and the *Guardian* revealed that more than 100 sites posting fake news were run by teenagers in the small town of Veles, Macedonia (Subramanian 2017). Endingthefed.com, a site that was responsible for four of the ten most popular fake news stories on Facebook, was run by a 24-year-old Romanian man (Townsend 2016). A US company called Disinfomedia owns many fake news sites, including NationalReport.net, USAToday.com.co, and WashingtonPost.com.co, and its owner claims to employ between 20 and 25 writers (Sydell 2016). Another US-based producer, Paul Horner, ran a successful fake news site called National Report for years prior to the election (Dewey 2014). Among his most-circulated stories was a 2013 report that President Obama used his own money to keep open a Muslim museum during the federal government shutdown. During the election, Horner produced a large number of mainly pro-Trump stories (Dewey 2016).

There appear to be two main motivations for providing fake news. The first is pecuniary: news articles that go viral on social media can draw significant advertising revenue when users click to the original site. This appears to have been the main motivation for most of the producers whose identities have been revealed. The teenagers in Veles, for example, produced stories favoring both Trump and Clinton that earned them tens of thousands of dollars (Subramanian 2017). Paul Horner produced pro-Trump stories for profit, despite claiming to be personally opposed to Trump (Dewey 2016). The second motivation is ideological. Some fake news providers seek to advance candidates they favor. The Romanian man who ran endingthefed.com, for example, claims that he started the site mainly to help Donald Trump's campaign (Townsend 2016). Other providers of right-wing fake news actually say they identify as left-wing and wanted to embarrass those on the right by showing that they would credulously circulate false stories (Dewey 2016; Sydell 2016).

A Model of Fake News

How is fake news different from biased or slanted media more broadly? Is it an innocuous form of entertainment, like fictional films or novels? Or does it have larger social costs? To answer these questions, we sketch a model of supply and demand for news loosely based on a model developed formally in Gentzkow, Shapiro, and Stone (2016).

There are two possible unobserved states of the world, which could represent whether a left- or right-leaning candidate will perform better in office. Media firms receive signals that are informative about the true state, and they may differ in the precision of these signals. We can also imagine that firms can make costly investments to increase the accuracy of these signals. Each firm has a reporting strategy that maps from the signals it receives to the news reports that it publishes. Firms can either decide to report signals truthfully, or alternatively to add bias to reports.

Consumers are endowed with heterogeneous priors about the state of the world. Liberal consumers' priors hold that the left-leaning candidate will perform better in office, while conservative consumers' priors hold that the right-leaning candidate will perform better. Consumers receive utility through two channels. First, they want to know the truth. In our model, consumers must choose an action, which could represent advocating or voting for a candidate, and they receive private benefits if they choose the candidate they would prefer if they were fully informed. Second, consumers may derive psychological utility from seeing reports that are consistent with their priors. Consumers choose the firms from which they will consume news in order to maximize their own expected utility. They then use the content of the news reports they have consumed to form a posterior about the state of the world. Thus, consumers face a tradeoff: they have a private incentive to consume precise and unbiased news, but they also receive psychological utility from confirmatory news.

After consumers choose their actions, they may receive additional feedback about the true state of the world—for example, as a candidate's performance is observed while in office. Consumers then update their beliefs about the quality of media firms and choose which to consume in future periods. The profits of media firms increase in their number of consumers due to advertising revenue, and media firms have an incentive to build a reputation for delivering high levels of utility to consumers. There are also positive social externalities if consumers choose the higher-quality candidate.

In this model, two distinct incentives may lead firms to distort their reports in the direction of consumers' priors. First, when feedback about the true state is limited, rational consumers will judge a firm to be higher quality when its reports are closer to the consumers' priors (Gentzkow and Shapiro 2006). Second, consumers may prefer reports that confirm their priors due to psychological utility (Mullainathan and Shleifer 2005). Gentzkow, Shapiro, and Stone (2016) show how these incentives can lead to biased reporting in equilibrium, and apply variants of this model to understand outcomes in traditional "mainstream" media.

How would we understand fake news in the context of such a model? Producers of fake news are firms with two distinguishing characteristics. First, they make no investment in accurate reporting, so their underlying signals are uncorrelated with the true state. Second, they do not attempt to build a long-term reputation for quality, but rather maximize the short-run profits from attracting clicks in an initial period. Capturing precisely how this competition plays out on social media would require extending the model to include multiple steps where consumers see "headlines" and then decide whether to "click" to learn more detail. But loosely speaking, we can imagine that such firms attract demand because consumers cannot distinguish them from higher-quality outlets, and also because their reports are tailored to deliver psychological utility to consumers on either the left or right of the political spectrum.

Adding fake news producers to a market has several potential social costs. First, consumers who mistake a fake outlet for a legitimate one have less-accurate beliefs and are worse off for that reason. Second, these less-accurate beliefs may reduce positive social externalities, undermining the ability of the democratic process to select high-quality candidates. Third, consumers may also become more skeptical of legitimate news producers, to the extent that they become hard to distinguish from fake news producers. Fourth, these effects may be reinforced in equilibrium by supply-side responses: a reduced demand for high-precision, low-bias reporting will reduce the incentives to invest in accurate reporting and truthfully report signals. These negative effects trade off against any welfare gain that arises from consumers who enjoy reading fake news reports that are consistent with their priors.

Real Data on Fake News

Fake News Database

We gathered a database of fake news articles that circulated in the three months before the 2016 election, using lists from three independent third parties. First, we scraped all stories from the Donald Trump and Hillary Clinton tags on Snopes (snopes.com), which calls itself "the definitive Internet reference source for urban legends, folklore, myths, rumors, and misinformation." Second, we scraped all stories from the 2016 presidential election tag from PolitiFact (politifact.com), another major fact-checking site. Third, we use a list of 21 fake news articles that had received significant engagement on Facebook, as compiled by the news outlet BuzzFeed (Silverman 2016).[4] Combining these three lists, we have a database of 156 fake news articles. We then gathered the total number of times each article was shared on Facebook as of early December 2016, using an online content database called BuzzSumo (buzzsumo.com). We code each article's content as either pro-Clinton (including anti-Trump) or pro-Trump (including anti-Clinton).

This list is a reasonable but probably not comprehensive sample of the major fake news stories that circulated before the election. One measure of comprehensiveness is to look at the overlap between the lists of stories from Snopes, PolitiFact, and BuzzFeed. Snopes is our largest list, including 138 of our total of 156 articles. As

[4] Of these 21 articles, 12 were fact-checked on Snopes. Nine were rated as "false," and the other three were rated "mixture," "unproven," and "mostly false."

a benchmark, 12 of the 21 articles in the BuzzFeed list appear in Snopes, and 4 of the 13 articles in the PolitiFact appear in Snopes. The lack of perfect overlap shows that none of these lists is complete and suggests that there may be other fake news articles that are omitted from our database.

Post-Election Survey

During the week of November 28, 2016, we conducted an online survey of 1208 US adults aged 18 and over using the SurveyMonkey platform. The sample was drawn from SurveyMonkey's Audience Panel, an opt-in panel recruited from the more than 30 million people who complete SurveyMonkey surveys every month (as described in more detail at https://www.surveymonkey.com/mp/audience/).

The survey consisted of four sections. First, we acquired consent to participate and a commitment to provide thoughtful answers, which we hoped would improve data quality. Those who did not agree were disqualified from the survey. Second, we asked a series of demographic questions, including political affiliation before the 2016 campaign, vote in the 2016 presidential election, education, and race/ethnicity. Third, we asked about 2016 election news consumption, including time spent on reading, watching, or listening to election news in general and on social media in particular, and the most important source of news and information about the 2016 election. Fourth, we showed each respondent 15 news headlines about the 2016 election. For each headline, we asked, "Do you recall seeing this reported or discussed prior to the election?" and "At the time of the election, would your best guess have been that this statement was true?" We also received age and income categories, gender, and census division from profiling questions that respondents had completed when they first started taking surveys on the Audience panel. The survey instrument can be accessed at https://www.surveymonkey.com/r/RSYD75P.

Each respondent's 15 news headlines were randomly selected from a list of 30 news headlines, six from each of five categories. Within each category, our list contains an equal split of pro-Clinton and pro-Trump headlines, so 15 of the 30 articles favored Clinton, and the other 15 favored Trump. The first category contains six fake news stories mentioned in three mainstream media articles (one in the *New York Times*, one in the *Wall Street Journal*, and one in BuzzFeed) discussing fake news during the week of November 14, 2016. The second category contains the four most recent pre-election headlines from each of Snopes and PolitiFact deemed to be unambiguously false. We refer to these two categories individually as "Big Fake" and "Small Fake," respectively, or collectively as "Fake." The third category contains the most recent six major election stories from the *Guardian's* election timeline. We refer to these as "Big True" stories. The fourth category contains the two most recent pre-election headlines from each of Snopes and PolitiFact deemed to be unambiguously true. We refer to these as "Small True" stories. Our headlines in these four categories appeared on or before November 7.

The fifth and final category contains invented "Placebo" fake news headlines, which parallel placebo conspiracy theories employed in surveys by Oliver and Wood (2014) and Chapman University (2016). As we explain below, we include these

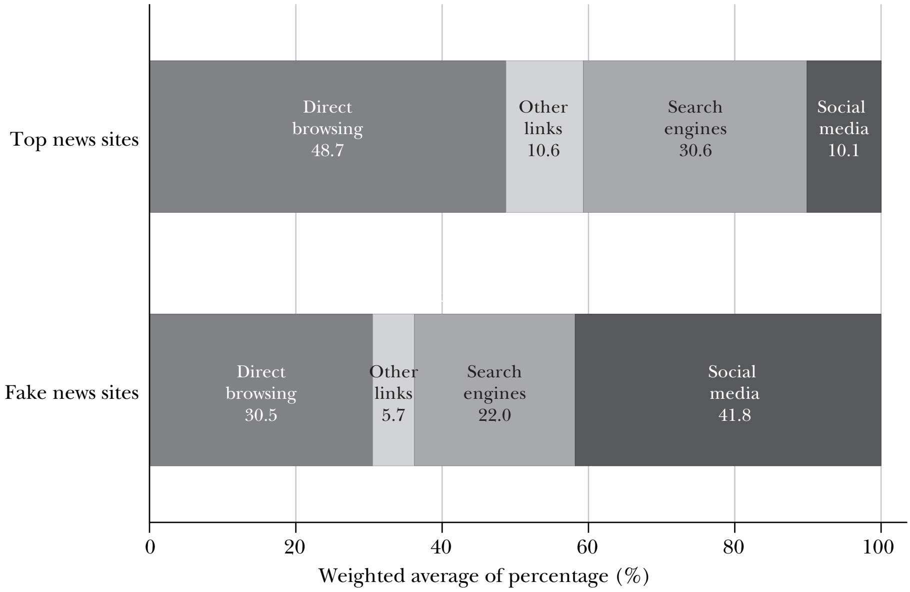
Placebo headlines to help control for false recall in survey responses. We invented three damaging fake headlines that could apply to either Clinton or Trump, then randomized whether a survey respondent saw the pro-Clinton or pro-Trump version. We experimented with several alternative placebo headlines during a pilot survey, and we chose these three because the data showed them to be approximately equally believable as the "Small Fake" stories. (We confirmed using Google searches that none of the Placebo stories had appeared in actual fake news articles.) Online Appendix Table 1, available with this article at this journal's website (http://e-jep.org), lists the exact text of the headlines presented in the survey. The online Appendix also presents a model of survey responses that makes precise the conditions under which differencing with respect to the placebo articles leads to valid inference.

Yeager et al. (2011) and others have shown that opt-in internet panels such as ours typically do not provide nationally representative results, even after reweighting. Notwithstanding, reweighting on observable variables such as education and internet usage can help to address the sample selection biases inherent in an opt-in internet-based sampling frame. For all results reported below, we reweight the online sample to match the nationwide adult population on ten characteristics that we hypothesized might be correlated with survey responses, including income, education, gender, age, ethnicity, political party affiliation, and how often the respondent reported consuming news from the web and from social media. The online Appendix includes summary statistics for these variables; our unweighted sample is disproportionately well-educated, female, and Caucasian, and those who rely relatively heavily on the web and social media for news. The Appendix also includes additional information on data construction.

Social Media as a Source of Political Information

The theoretical framework we sketched above suggests several reasons why social media platforms may be especially conducive to fake news. First, on social media, the fixed costs of entering the market and producing content are vanishingly small. This increases the relative profitability of the small-scale, short-term strategies often adopted by fake news producers, and reduces the relative importance of building a long-term reputation for quality. Second, the format of social media—thin slices of information viewed on phones or news feed windows—can make it difficult to judge an article's veracity. Third, Bakshy, Messing, and Adamic (2015) show that Facebook friend networks are ideologically segregated—among friendships between people who report ideological affiliations in their profiles, the median share of friends with the opposite ideology is only 20 percent for liberals and 18 percent for conservatives—and people are considerably more likely to read and share news articles that are aligned with their ideological positions. This suggests that people who get news from Facebook (or other social media) are less likely to receive evidence about the true state of the world that would counter an ideologically aligned but false story.

Figure 3
Share of Visits to US News Websites by Source



Note: This figure presents the share of traffic from different sources for the top 690 US news websites and for 65 fake news websites. "Other links" means impressions that were referred from sources other than search engines and social media. "Direct browsing" means impressions that did not have a referral source. Sites are weighted by number of monthly visits. Data are from Alexa.

One way to gauge the importance of social media for fake news suppliers is to measure the source of their web traffic. Each time a user visits a webpage, that user has either navigated directly (for example, by typing www.wsj.com into a browser) or has been referred from some other site. Major referral sources include social media (for example, clicking on a link in the Facebook news feed) and search engines (for example, searching for "Pope endorsed Trump?" on Google and clicking on a search result). Figure 3 presents web traffic sources for the month around the 2016 US presidential election (late October through late November) from Alexa (alexa.com), which gathers data from browser extensions installed on people's computers as well as from measurement services offered to websites. These data exclude mobile browsing and do not capture news viewed directly on social media sites, for example, when people read headlines within Facebook or Twitter news feeds.

The upper part of the graph presents referral sources for the top 690 US news sites, as ranked by Alexa. The lower part of the graph presents web traffic sources for a list of 65 major fake news sites, which we gathered from lists compiled by Zimdars (2016) and Brayton (2016). For the top news sites, social media referrals represent only about 10 percent of total traffic. By contrast, fake news websites rely on social media for a much higher share of their traffic. This demonstrates the importance of social media for fake news providers. While there is no definitive list of fake news sites, and one might disagree with the inclusion or exclusion of particular sites in this list of 65, this core point about the importance of social media for fake news providers is likely to be robust.

A recent Pew survey (Gottfried and Shearer 2016) finds that 62 percent of US adults get news from social media. To the extent that fake news is socially costly and fake news is prevalent on social media, this statistic could appear to be cause for concern. Of this 62 percent, however, only 18 percent report that they get news from social media "often," 26 percent do so "sometimes," and 18 percent do so "hardly ever." By comparison, the shares who "often" get news from local television, national broadcast television, and cable television are 46 percent, 30 percent, and 31 percent respectively. Moreover, only 34 percent of web-using adults trust the information they get from social media "some" or "a lot." By contrast, this share is 76 percent for national news organizations and 82 percent for local news organizations.

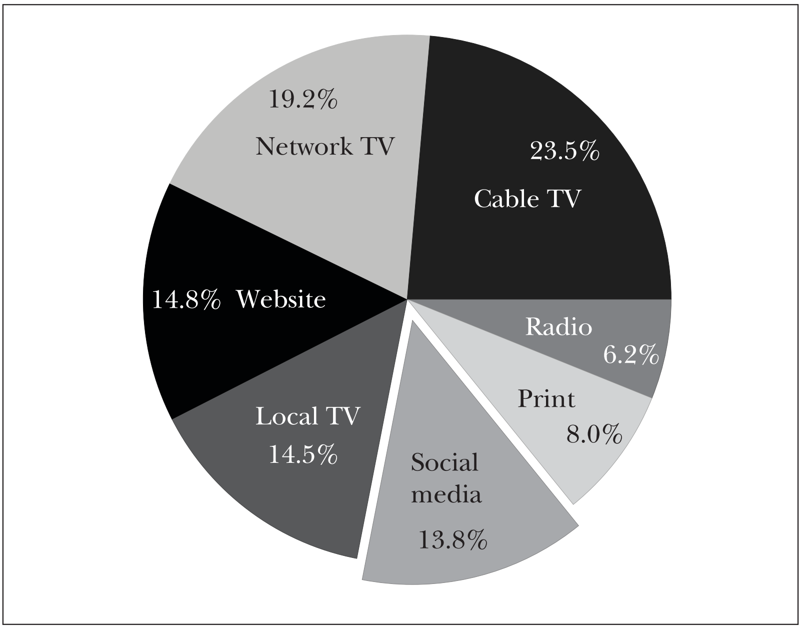
The results of our post-election survey are broadly consistent with this picture. For the month before the 2016 election, our respondents report spending 66 minutes per day reading, watching, or listening to election news. (Again, these and all other survey results are weighted for national representativeness.) Of this, 25 minutes (38 percent) was on social media. Our survey then asked, "Which of these sources was your most important source of news and information about the 2016 election?" The word "important" was designed to elicit a combination of consumption frequency and trust in information. Figure 4 presents responses. In order, the four most common responses are cable TV, network TV, websites, and local TV. Social media is the fifth most common response, with 14 percent of US adults listing social media as their most "important" news source.

Taken together, these results suggest that social media have become an important but not dominant source of political news and information. Television remains more important by a large margin.

Partisanship of Fake News

In our fake news database, we record 41 pro-Clinton (or anti-Trump) and 115 pro-Trump (or anti-Clinton) articles, which were shared on Facebook a total of 7.6 million and 30.3 million times, respectively. Thus, there are about three times more fake pro-Trump articles than pro-Clinton articles, and the average pro-Trump article was shared more on Facebook than the average pro-Clinton article. To be clear, these statistics show that more of the fake news articles *on these three fact-checking sites* are right-leaning. This could be because more of the actual fake news is right-leaning, or because more right-leaning assertions are forwarded to and/or reported by fact-checking sites, or because the conclusions that fact-checking sites draw have a left-leaning bias, or some combination. Some anecdotal reports support the idea that the majority of election-related fake news was pro-Trump: some fake

Figure 4
Most Important Source of 2016 Election News



Notes: Our post-election survey asked, "Which of these sources was your most important source of news and information about the 2016 election?" This figure plots responses. Observations are weighted for national representativeness.

news providers reportedly found higher demand for pro-Trump (or anti-Clinton) fake news, and responded by providing more of it (Sydell 2016).

There could be several possible explanations for a preponderance of pro-Trump fake news. The more marked decline of trust in the mainstream media among Republicans shown in Figure 2 could have increased their relative demand for news from nontraditional sources, as could a perception that the mainstream media tended to favor Clinton. Pro-Trump (and anti-Clinton) storylines may have simply been more compelling than pro-Clinton (and anti-Trump) storylines due to particulars of these candidates, perhaps related to the high levels of media attention that Trump received throughout the campaign. Or, it could theoretically be that Republicans are for some reason more likely to enjoy or believe fake news.

Some prior evidence argues against the last hypothesis. McClosky and Chong (1985) and Uscinski, Klofstad, and Atkinson (2016) find that people on the left and right are equally disposed to conspiratorial thinking. Furthermore, Bakshy, Messing, and Adamic (2015) find that conservatives are actually exposed to more cross-cutting news content than liberals, which could help conservatives to be better at detecting partisan fake news. Below, we present further evidence on this hypothesis from our survey.

Exposure to Fake News

How much fake news did the typical voter see in the run-up to the 2016 election? While there is a long literature measuring media exposure (for example, Price and Zaller 1993), fake news presents a particular challenge: much of its circulation is on Facebook (and other social media) news feeds, and these data are not public. We provide three benchmarks for election-period fake news exposure, which we report as average exposure for each of the 248 million American adults.

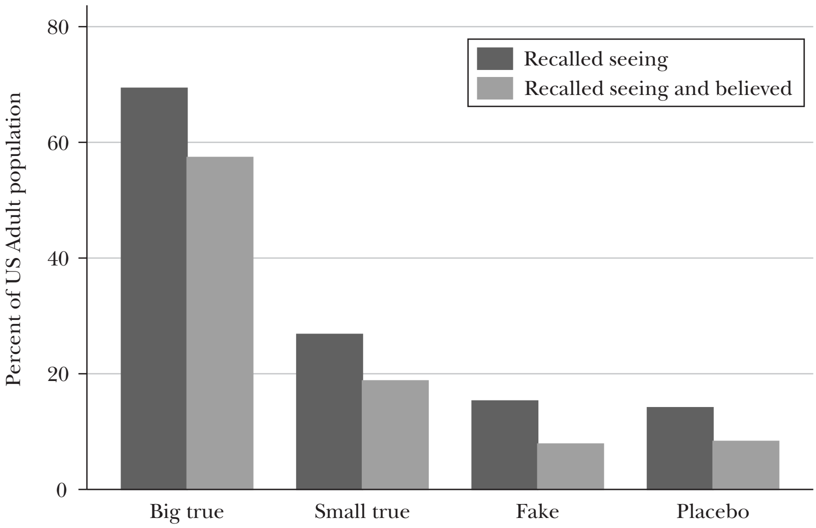
First, we can use prior evidence to predict the number of times the articles in our database were read based on the number of times they were shared. The corporate website of Eventbrite (2012) reports that links to its events on Facebook generate 14 page visits per share. A blog post by Jessica Novak (undated) reports that for a set of "top performing" stories on Facebook the ratio of visits to shares was also 14. Zhao, Wang, Tao, Ma, and Guan (2013) report that the ratio of views to shares for videos on the Chinese social networking site Renren ranges from 3 to 8. Based on these very rough reference points, we consider a ratio of 20 page visits per share as an upper bound on the plausible range. This implies that the 38 million shares of fake news in our database translate into 760 million page visits, or about three visits per US adult.

Second, we can use web browsing data to measure impressions on fake news websites. For the month around the 2016 election, there were 159 million impressions on the 65 websites in the bottom part of Figure 3, or 0.64 impressions per adult. This is dwarfed by the 3 billion impressions on the 665 top news websites over the same period. Furthermore, not all content on these 65 sites is false: in a random sample of articles from these sites, we categorized just under 55 percent as false, either because the claim was refuted by a mainstream news site or fact-checking organization, or because the claim was not covered on any other sites despite being important enough that it would have been covered on other sites if it were true. When comparing these first two approaches to estimating election-period fake news exposure, remember that the first approach uses cumulative Facebook shares as of early December 2016 for fake news articles that were fact-checked in the three months before the election, while the second approach uses web traffic from a one month period between late October to late November 2016.

Third, we can use our post-election survey to estimate the number of articles respondents saw and remembered. The survey gave respondents 15 news headlines—three headlines randomly selected from each of the five categories detailed earlier—and asked if they recalled seeing the headline ("Do you recall seeing this reported or discussed prior to the election?") and if they believed it ("At the time of the election, would your best guess have been that this statement was true?").

Figure 5 presents the share of respondents that recalled seeing (left bar) and seeing and believing (right bar) headlines, averaging responses across all the headlines within each of our main categories. Rates of both seeing and believing are much higher for true than fake stories, and they are substantially higher for the "Big True" headlines (the major headlines leading up to the election) than for the

Figure 5
Percent of US Adult Population that Recall Seeing or that Believed Election News



Notes: In our post-election survey, we presented 15 headlines. For each headline, the survey asked whether respondents recall seeing the headline ("Do you recall seeing this reported or discussed before the election?") and whether they believed it ("At the time of the election, would your best guess have been that this statement was true?"). The left bars present the share of respondents who recall seeing the headlines in each category, and the right bars present the share of respondents who recall seeing and believed the headlines. "Big True" headlines are major headlines leading up to the election; "Small True" headlines are the minor fact-checked headlines that we gathered from Snopes and PolitiFact. The Placebo fake news headlines were made-up for the research and never actually circulated. Observations are weighted for national representativeness.

"Small True" headlines (the minor fact-checked headlines that we gathered from Snopes and PolitiFact). The Placebo fake news articles, which never actually circulated, are approximately equally likely to be recalled and believed as the Fake news articles which did actually circulate. This implies that there is a meaningful rate of false recall of articles that people never actually saw, which could cause the survey measure to significantly overstate true exposure. On the other hand, people likely forgot some of the Fake articles that they were actually exposed to, which causes the survey responses to understate true exposure.

In summary, one can think of recalled exposure as determined both by actual exposure and by the headline's perceived plausibility—people might think that if a headline is plausible, they probably saw it reported somewhere. Then, we show that if the Placebo headlines are equally plausible as the Fake headlines, the difference between recall of Fake and Placebo headlines represents the rate of true exposure that was remembered. The Appendix available online with this paper at http://e-jep.org presents additional theoretical and empirical discussion of false recall in our data.

After weighting for national representativeness, 15 percent of survey respondents recalled seeing the Fake stories, and 8 percent both recalled seeing the story and said they believed it.[5] By comparison, about 14 percent of people report seeing the placebo stories, and about 8 percent report seeing and believing them. We estimate that the average Fake headline was 1.2 percentage points more likely to be seen and recalled than the average Placebo headline, and the 95 percent confidence interval allows us to exclude differences greater than 2.9 percent.

We can use these results to provide a separate estimate of fake news exposure. The average Fake article that we asked about in the post-election survey was shared 0.386 million times on Facebook. If the average article was seen and recalled by 1.2 percent of American adults, this gives (0.012 recalled exposure)/(0.386 million shares) ≈ 0.03 chance of a recalled exposure per million Facebook shares. Given that the Fake articles in our database had 38 million Facebook shares, this implies that the average adult saw and remembered 0.03/million × 38 million ≈ 1.14 fake news articles from our fake news database.

All three approaches suggest that election-period fake news exposure was on the order of one or perhaps several articles read per adult. We emphasize several important caveats. First, each of these measures excludes some forms of exposure that could have been influential. All of them exclude stories or sites omitted from our database. Estimated page visits or impressions exclude cases in which users saw a story within their Facebook news feed but did not click through to read it. Our survey-based recall measure excludes stories that users saw but did not remember, and may be subject to other biases associated with survey-based estimates of media exposure (Bartels 1993; Prior 2009; Guess 2015).

Who Believes Fake News?

It is both privately and socially valuable when people can infer the true state of the world. What factors predict the ability to distinguish between real and fake news? This analysis parallels a literature in political science measuring and interpreting correlates of misinformation, including Lewandowsky, Oberauer, and Gignac (2013), Malka, Krosnick, and Langer (2009), and Oliver and Wood (2014).

We construct a variable C_{ia}, that takes value 1 if survey respondent i correctly identifies whether article a is true or false, 0.5 if respondent i is "not sure," and value 0 otherwise. For example, if headline a is true, then C_{ia} takes value 1 if person i responded "Yes" to "would your best guess have been that this statement was true?"; 0.5 if person i responded "Not sure"; and 0 if person i responded "No." We use C_{ia}

[5] These shares are broadly consistent with the results of a separate survey conducted by Silverman and Singer-Vine (2016): for a set of five fake news stories, they find that the share of respondents who have heard them ranges from 10 to 22 percent and the share who rate them as "very accurate" ranges from 28 to 49 percent.

as the dependent variable and a vector $\mathbf{X}_i$ of individual characteristics in a linear regression:

$$C_{ia} = \boldsymbol{\alpha}_1\mathbf{X}_i + \alpha_0 + \varepsilon_{ia}.$$

Table 1 reports results. Column 1 includes only false articles (both Fake and Placebo), and focuses only on party affiliation; the omitted category is Independents. In these data, it is indeed true that Republicans were statistically less likely than Democrats to report that they (correctly) did not believe a false article. Column 2 includes only true articles (both Big True and Small True categories). This suggests that Republicans are also more likely than Democrats to correctly believe articles that were true ($p = 0.124$). These results suggest that in our data, Republicans were not generally worse at inference: instead, they tended to be more credulous of both true and false articles. Of course, it is possible that this is simply an artifact of how different respondents interpreted the survey design. For example, it could be that Republicans tended to expect a higher share of true headlines in our survey, and thus were less discerning.

Another possible explanation is that the differences between parties hide other factors associated with party affiliation. Columns 3 and 4 test this possibility, including a vector of additional covariates. The differences between the Democrat and Republican indicator variables are relatively robust. Column 5 includes all articles, which weights true and false articles by the proportions in our survey sample. Given that our survey included a large proportion of fake articles that Republicans were less likely to recognize as false, Democrats are overall more likely to correctly identify true versus false articles. Three correlations tend to be statistically significant: people who spend more time consuming media, people with higher education, and older people have more accurate beliefs about news. As with Republicans relative to Democrats, people who report that social media were their most important sources of election news were more likely both to correctly believe true headlines and to incorrectly believe false headlines.

The association of education with correct beliefs should be highlighted. Flynn, Nyhan, and Reifler (2017) argue that education could have opposing effects on political misperceptions. On the one hand, education should increase people's ability to discern fact from fiction. On the other hand, in the presence of motivated reasoning, education gives people better tools to counterargue against incongruent information. To the extent that the association in our data is causal, it would reinforce many previous arguments that the social return to education includes cognitive abilities that better equip citizens to make informed voting decisions. For example, Adam Smith (1776) wrote, "The more [people] are instructed, the less liable they are to the delusions of enthusiasm and superstition, which, among ignorant nations, frequently occasion the most dreadful disorders."

A common finding in the survey literature on rumors, conspiracy theories, and factual beliefs is that partisan attachment is an important predictor of beliefs (for example, Oliver and Wood 2014; Uscinski, Klofstad, and Atkinson 2016).

Table 1
What Predicts Correct Beliefs about News Headlines?

	(1)	*(2)*	*(3)*	*(4)*	*(5)*
Democrat	0.029	–0.004	0.028	–0.010	0.015
	(0.020)	(0.023)	(0.019)	(0.021)	(0.013)
Republican	–0.024	0.040	–0.037*	0.021	–0.018
	(0.024)	(0.027)	(0.020)	(0.023)	(0.014)
ln(Daily media time)			–0.002	0.042***	0.013***
			(0.007)	(0.008)	(0.004)
Social media most important			–0.066***	0.065***	–0.023
			(0.025)	(0.024)	(0.016)
Use social media			0.014	–0.023	0.002
			(0.030)	(0.038)	(0.019)
Social media ideological segregation			–0.027	0.028	–0.008
			(0.036)	(0.046)	(0.024)
Education			0.014***	0.004	0.011***
			(0.004)	(0.004)	(0.003)
Undecided			–0.011	0.006	–0.005
			(0.017)	(0.022)	(0.013)
Age			0.002***	0.000	0.002***
			(0.000)	(0.001)	(0.000)
N	12,080	6,040	12,080	6,040	18,120
p-value (Democrat = Republican)	0.029	0.124	0.004	0.207	0.035
Articles in sample	False	True	False	True	All

Note: This table presents estimates of a regression of a dependent variable measuring correct beliefs about headlines on individual characteristics. Columns 1 and 3 include only false headlines, columns 2 and 4 contain only true headlines, and column 5 contains all headlines. All columns include additional demographic controls: income, race, and gender. "Social media most important" means social media were the respondent's most important sources of election news. "Social media ideological segregation" is the self-reported share (from 0 to 1) of social media friends that preferred the same presidential candidate. "Undecided" is an indicator variable for whether the respondent decided which candidate to vote for less than three months before the election. Observations are weighted for national representativeness. Standard errors are robust and clustered by survey respondent.
*, **, *** indicate statistically significantly different from zero with 90, 95, and 99 percent confidence, respectively.

For example, Republicans are more likely than Democrats to believe that President Obama was born outside the United States, and Democrats are more likely than Republicans to believe that President Bush was complicit in the 9/11 attacks (Cassino and Jenkins 2013). Such polarized beliefs are consistent with a Bayesian framework, where posteriors depend partially on priors, as well as with models of motivated reasoning (for example, Taber and Lodge 2006, or see the symposium in the Summer 2016 issue of this journal). Either way, the ability to update one's priors in response to factual information is privately and socially valuable in our model, and polarized views on factual issues can damage society's ability to come to agreement on what social problems are important and how to address them (Sunstein 2001a, b, 2007).

Given this discussion, do we also see polarized beliefs with respect to fake news? And if so, what factors moderate ideologically aligned inference—that is, what factors predict a lower probability that a Republican is more likely to believe pro-Trump news than pro-Clinton news, or that a Democrat is more likely to believe pro-Clinton than pro-Trump news? To gain insight into this question, we define B_{ia} as a measure of whether individual i believed article a, taking value 1 if "Yes," 0.5 if "Not sure," and 0 if "No." We also define D_i and R_i as Democrat and Republican indicators, and C_a and T_a as indicators for whether headline a is pro-Clinton or pro-Trump. We then run the following regression in the sample of Democrats and Republicans, excluding Independents:

$$B_{ia} = \beta_D D_i C_a + \beta_R R_i T_a + \gamma_D D_i + \gamma_R R_i + \varepsilon_{ia}.$$

The first two independent variables are interaction terms; their coefficients β_D and β_R measure whether a Democrat is more likely to believe a pro-Clinton headline and whether a Republican is more likely to believe a pro-Trump headline. The second two independent variables control for how likely Democrats or Republicans are as a group are to believe all stories. Since headlines are randomly assigned to respondents, with equal balance of true versus false and pro-Trump versus pro-Clinton, the estimated β parameters will measure ideologically aligned inference

Table 2 presents the results. Column 1 presents estimates of β_D and β_R. Democrats and Republicans, respectively, are 17.2 and 14.7 percentage points more likely to believe ideologically aligned articles than they are to believe nonaligned articles. Column 2 takes an intermediate step, constraining the β coefficients to be the same. Column 3 then allows β to vary by the same vector of $\mathbf{X}_i$ variables as reported in Table 1, except excluding D_i to avoid collinearity. In both columns 1 and 3, any differences between Democrats and Republicans in the magnitude of ideologically aligned inference are not statistically significant.

Three variables are strongly correlated with ideologically aligned inference. First, heavy media consumers are more likely to believe ideologically aligned articles. Second, those with segregated social networks are significantly more likely to believe ideologically aligned articles, perhaps because they are less likely to receive disconfirmatory information from their friends. The point estimate implies that a 0.1 (10 percentage point) increase in the share of social media friends that preferred the same presidential candidate is associated with a 0.0147 (1.47 percentage point) increase in belief of ideologically aligned headlines relative to ideologically cross-cutting headlines. Third, "undecided" adults (those who did not make up their minds about whom to vote for until less than three months before the election) are less likely to believe ideologically aligned articles than more decisive voters. This is consistent with undecided voters having less-strong ideologies in the first place. Interestingly, social media use and education are not statistically significantly associated with more or less ideologically aligned inference.

Table 2

Ideological Alignment and Belief of News Headlines

	(1)	*(2)*	*(3)*
Democrat × Pro-Clinton	0.172*** (0.021)		
Republican × Pro-Trump	0.147*** (0.023)		
Aligned		0.161*** (0.016)	0.096 (0.140)
Aligned × Republican			0.000 (0.027)
Aligned × ln(Daily media time)			0.024*** (0.009)
Aligned × Social media most important			–0.031 (0.037)
Aligned × Use social media			–0.068 (0.050)
Aligned × Social media ideological segregation			0.147*** (0.046)
Aligned × Education			–0.004 (0.007)
Aligned × Undecided			–0.099*** (0.030)
Aligned × Age			0.001 (0.001)
N	10,785	10,785	10,785

Notes: This table presents estimates of a regression of a variable measuring belief of news headlines on the interaction of political party affiliation indicators and pro-Clinton or pro-Trump headline indicators. The sample includes all news headlines (both true and false) but excludes survey respondents who are Independents. "Social media most important" means social media were the respondent's most important sources of election news. "Social media ideological segregation" is the self-reported share (from 0 to 1) of social media friends that preferred the same presidential candidate. "Undecided" is an indicator variable for whether the respondent decided which candidate to vote for less than three months before the election. Observations are weighted for national representativeness. Standard errors are robust and clustered by survey respondent. *, **, ***: statistically significantly different from zero with 90, 95, and 99 percent confidence, respectively.

One caveat to these results is that ideologically aligned inference may be exaggerated by respondents' tendency to answer expressively or to want to "cheerlead" for their party (Bullock, Gerber, Hill, and Huber 2015; Gerber and Huber 2009; Prior, Sood, and Khanna 2015). Partisan gaps could be smaller in a survey with strong incentives for correct answers.

Conclusion

In the aftermath of the 2016 US presidential election, it was alleged that fake news might have been pivotal in the election of President Trump. We do not provide an assessment of this claim one way or another.

That said, the new evidence we present clarifies the level of overall exposure to fake news, and it can give some sense of how persuasive fake news would need to have been to have been pivotal. We estimate that the average US adult read and remembered on the order of one or perhaps several fake news articles during the election period, with higher exposure to pro-Trump articles than pro-Clinton articles. How much this affected the election results depends on the effectiveness of fake news exposure in changing the way people vote. As one benchmark, Spenkuch and Toniatti (2016) show that exposing voters to one additional television campaign ad changes vote shares by approximately 0.02 percentage points. This suggests that if one fake news article were about as persuasive as one TV campaign ad, the fake news in our database would have changed vote shares by an amount on the order of hundredths of a percentage point. This is much smaller than Trump's margin of victory in the pivotal states on which the outcome depended.

Of course there are many reasons why a single fake news story could have been more effective than a television commercial. If it were true that the Pope endorsed Donald Trump, this fact would be significantly more surprising—and probably move a rational voter's beliefs by more as a result—than the information contained in a typical campaign ad. Moreover, as we emphasize above, there are many ways in which our estimates could understate true exposure. We only measure the number of stories read and remembered, and the excluded stories seen on news feeds but not read, or read but not remembered, could have had a large impact. Our fake news database is incomplete, and the effect of the stories it omits could also be significant.

We also note that there are several ways in which this back-of-the-envelope calculation is conservative, in the sense that it could overstate the importance of fake news. We consider the number of stories voters read regardless of whether they believed them. We do not account for diminishing returns, which could reduce fake news' effect to the extent that a small number of voters see a large number of stories. Also, this rough calculation does not explicitly take into account the fact that a large share of pro-Trump fake news is seen by voters who are already predisposed to vote for Trump—the larger this selective exposure, the smaller the impact we would expect of fake news on vote shares.

To the extent that fake news imposes social costs, what can and should be done? In theory, a social planner should want to address the market failures that lead to distortions, which would take the form of increasing information about the state of the world and increasing incentives for news consumers to infer the true state of the world. In practice, social media platforms and advertising networks have faced some pressure from consumers and civil society to reduce the prevalence of fake news on their systems. For example, both Facebook and Google are removing fake news sites from their advertising platforms on the grounds that they violate policies against misleading content (Wingfield, Isaac, and Benner 2016). Furthermore, Facebook has taken steps to identify fake news articles, flag false articles as “disputed by 3rd party fact-checkers,” show fewer potentially false articles in users’ news feeds, and help users avoid accidentally sharing false articles by notifying them that a story is “disputed by 3rd parties” before they share it (Mosseri 2016). In our theoretical framework, these actions may increase social welfare, but identifying fake news sites and articles also raises important questions about who becomes the arbiter of truth.

■ *We thank David Deming, Brendan Nyhan, Craig Silverman, Aaron Smith, Joe Uscinski, David Vannette, and many other colleagues for helpful conversations and feedback. We are grateful to Chuan Yu and Nano Barahona for research assistance, and we thank Stanford University for financial support. Our survey was determined to be exempt from human subjects review by the NYU and Stanford Institutional Review Boards.*

References

Abramowitz, Alan I., and Kyle L. Saunders. 2008. “Is Polarization a Myth?” *Journal of Politics* 70(2): 542–55.

American Enterprise Institute. 2013. “Public Opinion on Conspiracy Theories.” AEI Public Opinion Study. Compiled by Karlyn Bowman and Andrew Rugg. November, https://www.aei.org/wp-content/uploads/2013/11/-public-opinion-on-conspiracy-theories_181649218739.pdf.

American National Election Studies. 2010. Times Series Cumulative Data File [dataset]. Produced and distributed by Stanford University and the University of Michigan. http://www.electionstudies.org/studypages/anes_timeseries_cdf/anes_timeseries_cdf.htm.

Bagdikian, Ben H. 1983. *The Media Monopoly.* Beacon Press.

Bakshy, Eytan, Solomon Messing, and Lada A. Adamic. 2015. “Exposure to Ideologically Diverse News and Opinion on Facebook.” *Science* 348(6239): 1130–32.

Bartels, Larry M. 1993. “Messages Received: The Political Impact of Media Exposure.” *American Political Science Review* 87(2): 267–85.

Berinsky, Adam J. 2017. “Rumors and Health Care Reform: Experiments in Political Misinformation.” *British Journal of Political Science* 47(2): 241–62.

Brayton, Ed. 2016. “Please Stop Sharing Links to These Sites.” *Patheos*, September 18. http://www.patheos.com/blogs/dispatches/2016/09/18/please-stop-sharing-links-to-these-sites/.

Bullock, John G., Alan S. Gerber, Seth J. Hill, and Gregory A. Huber. 2015. “Partisan Bias in Factual Beliefs about Politics.” *Quarterly Journal of Political Science* 10(4): 519–78.

BuzzFeed News. No date. “Election Content Engagement.” [A spreadsheet] https://docs.google.com/spreadsheets/d/1ysnzawW6pDGBEqbXqeYuzWa7Rx2mQUip6CXUUUk4jIk/edit#gid=1756764129.

Cassino, Dan, and Krista Jenkins. 2013. “Conspiracy Theories Prosper: 25% of Americans Are ‘Truthers.’” Fairleigh Dickinson University’s Public Mind Poll. January 17. http://publicmind.fdu.edu/2013/outthere.

Chapman University. 2016. “What Aren’t They Telling Us?” Chapman University Survey of American Fears. October 11. https://blogs.chapman.edu/wilkinson/2016/10/11/what-arent-they-telling-us/.

DellaVigna, Stefano, and Matthew Gentzkow. 2010. “Persuasion: Empirical Evidence.” *Annual Review of Economics* 2: 643–69.

DellaVigna, Stefano, and Ethan Kaplan. 2007. “The Fox News Effect: Media Bias and Voting.” *Quarterly Journal of Economics* 122(3): 1187–1234.

Dewey, Caitlin. 2014. “This Is Not an Interview with Banksy.” *Washington Post*, October 22. https://www.washingtonpost.com/news/the-intersect/wp/2014/10/21/this-is-not-an-interview-with-banksy/?tid=a_inl&utm_term=.8a9 5d83438e9.

Dewey, Caitlin. 2016. “Facebook Fake-News Writer: ‘I Think Donald Trump is in the White House because of Me.’” *Washington Post,* November, 17. https://www.washingtonpost.com/news/the-intersect/wp/2016/11/17/facebook-fake-news-writer-i-think-donald-trump-is-in-the-white-house-because-of-me/.

DiFonzo, Nicholas, and Prashant Bordia. 2007. *Rumor Psychology: Social and Organizational Approaches.* American Psychological Association.

Enikolopov, Ruben, Maria Petrova, and Ekaterina Zhuravskaya. 2011. “Media and Political Persuasion: Evidence from Russia.” *American Economic Review* 101(7): 3253–85.

Eventbrite. 2012. “Social Commerce: A Global Look at the Numbers.” October 23. https://www.eventbrite.com/blog/ds00-social-commerce-a-global-look-at-the-numbers/.

Fiorina, Morris P., and Samuel J. Abrams. 2008. “Political Polarization in the American Public.” *Annual Review of Political Science* 11: 563–88.

Flaxman, Seth, Sharad Goel, and Justin M. Rao. 2016. “Filter Bubbles, Echo Chambers, and Online News Consumption.” *Public Opinion Quarterly* 80(1): 298–320.

Flynn, D. J., Brendan Nyhan, and Jason Reifler. 2017. “The Nature and Origins of Misperceptions: Understanding False and Unsupported Beliefs about Politics.” *Advances in Political Psychology* 38(S1): 127–50.

Friggeri, Adrien, Lada Adamic, Dean Eckles, and Justin Cheng. 2014. “Rumor Cascades.” Eighth International AAAI Conference on Weblogs and Social Media.

Gentzkow, Matthew, and Jesse M. Shapiro. 2006. “Media Bias and Reputation.” *Journal of Political Economy* 114(2): 280–316.

Gentzkow, Matthew, and Jesse M. Shapiro. 2011. “Ideological Segregation Online and Offline.” *Quarterly Journal of Economics* 126(4): 1799–1839.

Gentzkow, Matthew, Jesse M. Shapiro, and Daniel F. Stone. 2016. “Media Bias in the Marketplace: Theory.” Chap. 14 in *Handbook of Media Economics,* vol. 1B, edited by Simon Anderson, Joel Waldofgel, and David Stromberg.

Gerber, Alan S., James G. Gimpel, Donald P. Green, and Daron R. Shaw. 2011. “How Large and Long-lasting are the Persuasive Effects of Televised Campaign Ads? Results from a Randomized Field Experiment.” *American Political Science Review* 105(1): 135–150.

Gerber, Alan S., and Donald P. Green. 2000. “The Effects of Canvassing, Telephone Calls, and Direct Mail on Voter Turnout: A Field Experiment.” *American Political Science Review* 94(3): 653–63.

Gerber, Alan S., and Gregory A. Huber. 2009. “Partisanship and Economic Behavior: Do Partisan Differences in Economic Forecasts Predict Real Economic Behavior?” *American Political Science Review* 103(3): 407–26.

Gottfried, Jeffrey, and Elisa Shearer. 2016. “News Use across Social Media Platforms 2016.” Pew Research Center, May 26. http://www.journalism.org/2016/05/26/news-use-across-social-media-platforms-2016.

Guess, Andrew M. 2015. “Measure for Measure: An Experimental Test of Online Political Media Exposure.” *Political Analysis* 23(1): 59–75.

Huber, Gregory A., and Kevin Arceneaux. 2007. “Identifying the Persuasive Effects of Presidential Advertising.” *American Journal of Political Science* 51(4): 957–77.

Kaplan, Richard L. 2002. *Politics and the American Press: The Rise of Objectivity, 1865–1920.* Cambridge University Press.

Keeley, Brian L. 1999. “Of Conspiracy Theories.” *Journal of Philosophy* 96(3): 109–26.

Lang, Kurt, and Gladys Engel Lang. 2002. *Television and Politics.* Transaction Publishers.

Lelkes, Yphtach. 2016. “Mass Polarization: Manifestations and Measurements.” *Public Opinion Quarterly* 80(S1): 392–410.

Lewandowsky, Stephan, Gilles E. Gignac, and Klaus Oberauer. 2013. “The Role of Conspiracist Ideation and Worldviews in Predicting Rejection of Science.” *PloS One* 8(10): e75637.

Malka, Ariel, Jon A. Krosnick, and Gary Langer. 2009. “The Association of Knowledge with Concern about Global Warming: Trusted Information Sources Shape Public Thinking.” *Risk Analysis* 29(5): 633–47.

Martin, Gregory J., and Ali Yurukoglu. 2014. “Bias in Cable News: Persuasion and Polarization.” NBER Working Paper 20798.

McClosky, Herbert, and Dennis Chong. 1985. “Similarities and Differences between Left-Wing and Right-Wing Radicals.” *British Journal of Political Science* 15(3): 329–63.

Mosseri Adam. 2016. “News Feed FYI: Addressing Hoaxes and Fake News.” Newsroom, Facebook, December 15. http://newsroom.fb.com/news/2016/12/news-feed-fyi-addressing-hoaxes-and-fake-news/.

Mullainathan, Sendhil, and Andrei Shleifer. 2005. “The Market for News.” *American Economic Review* 95(4): 1031–53.

Napoli, Philip M. 2014. “Measuring Media Impact: An Overview of the Field.” Norman Lear Center Media Impact Project. https://learcenter.org/pdf/measuringmedia.pdf.

Novak, Jessica. No date. “Quantifying Virality: The Visits to Share Ratio.” http://intelligence.r29.com/post/105605860880/quantifying-virality-the-visits-to-shares-ratio.

Nyhan, Brendan, Jason Reifler, Sean Richey, and Gary L. Freed. 2014. "Effective Messages in Vaccine Promotion: A Randomized Trial." *Pediatrics* 133(4): 835–42.

Nyhan, Brendan, Jason Reifler, and Peter A. Ubel. 2013. "The Hazards of Correcting Myths about Health Care Reform." *Medical Care* 51(2): 127–32.

Oliver, J. Eric, and Thomas J. Wood. 2014. "Conspiracy Theories and the Paranoid Style(s) of Mass Opinion." *American Journal of Political Science* 58(4): 952–66.

Pariser, Eli. 2011. *The Filter Bubble: What the Internet Is Hiding from You.* Penguin UK.

Parkinson, Hannah Jane. 2016. "Click and Elect: How Fake News Helped Donald Trump Win a Real Election." *Guardian,* November 14.

PolitiFact. No date. http://www.politifact.com/truth-o-meter/elections/2016/president-united-states/.

Price, Vincent, and John Zaller. 1993. "Who Gets the News? Alternative Measures of News Reception and Their Implications for Research." *Public Opinion Quarterly* 57(2): 133–64.

Prior, Markus. 2009. "The Immensely Inflated News Audience: Assessing Bias in Self-Reported News Exposure." *Public Opinion Quarterly* 73(1): 130–43.

Prior, Markus. 2013. "Media and Political Polarization." *Annual Review of Political Science* 16: 101–27.

Prior, Markus, Gaurav Sood, and Kabir Khanna. 2015. "You Cannot Be Serious: The Impact of Accuracy Incentives on Partisan Bias in Reports of Economic Perceptions." *Quarterly Journal of Political Science* 10(4): 489–518.

Read, Max. 2016. "Donald Trump Won because of Facebook." *New York Magazine,* November 9.

Silverman, Craig. 2016. "This Analysis Shows how Fake Election News Stories Outperformed Real News on Facebook." *BuzzFeed News,* November 16.

Silverman, Craig and Jeremy Singer-Vine. 2016. "Most Americans Who See Fake News Believe It, New Survey Says." *BuzzFeed News,* December 6.

Smith, Adam. 1776. *The Wealth of Nations. London: W. Strahan.*

Spenkuch, Jörg L., and David Toniatti. 2016. "Political Advertising and Election Outcomes." CESifo Working Paper Series 5780.

Subramanian, Samanth. 2017. "Inside the Macedonian Fake-News Complex, *Wired*, February 15.

Sunstein, Cass R. 2001a. *Echo Chambers: Bush v. Gore, Impeachment, and Beyond.* Princeton University Press.

Sunstein, Cass R. 2001b. *Republic.com.* Princeton University Press.

Sunstein, Cass R. 2007. *Republic.com 2.0.* Princeton University Press.

Swift, Art. 2016. "Americans' Trust in Mass Media Sinks to New Low." Gallup, September 14. http://www.gallup.com/poll/195542/americans-trust-mass-media-sinks-new-low.aspx.

Sydell, Laura. 2016. "We Tracked Down a Fake-News Creator in the Suburbs. Here's What We Learned." National Public Radio, November 23. http://www.npr.org/sections/alltechconsidered/2016/11/23/503146770/npr-finds-the-head-of-a-covert-fake-news-operation-in-the-suburbs.

Taber, Charles S., and Milton Lodge. 2006. "Motivated Skepticism in the Evaluation of Political Beliefs." *American Journal of Political Science* 50(3): 755–69.

Townsend, Tess. 2016. "Meet the Romanian Trump Fan behind a Major Fake News Site." *Inc.* http://www. inc.com/tess-townsend/ending-fed-trump-facebook.html.

Uscinski, Joseph E., Casey Klofstad, and Matthew D. Atkinson. 2016. "What Drives Conspiratorial Beliefs? The Role of Informational Cues and Predispositions." *Political Research Quarterly* 69(1): 57–71.

Wingfield, Nick, Mike Isaac, and Katie Benner. 2016." Google and Facebook Take Aim at Fake News Sites." *New York Times,* November 14.

Yeager, David S., Jon A. Krosnick, LinChiat Chang, Harold S. Javitz, Matthew S. Levendusky, Alberto Simpser, and Rui Wang. 2011. "Comparing the Accuracy of RDD Telephone Surveys and Internet Surveys Conducted with Probability and Non-Probability Samples." *Public Opinion Quarterly* 75(4): 709–47.

Zhao, Junzhou, Pinghui Wang, Jing Tao, Xiaobo Ma, and Xiaohong Guan. 2013. "A Peep on the Interplays between Online Video Websites and Online Social Networks." ariXiv:1305.4018.

Zimdars, Melissa. 2016. "False, Misleading, Clickbait-y, and Satirical 'News' Sources." http://d279m997dpfwgl.cloudfront.net/wp/2016/11/Resource-False-Misleading-Clickbait-y-and-Satirical-%E2%80%9CNews%E2%80%9D-Sources-1.pdf.

Journal of Economic Perspectives—Volume 31, Number 2—Spring 2017—Pages 237–256

Yuliy Sannikov: Winner of the 2016 Clark Medal

Susan Athey and Andrzej Skrzypacz

Yuliy Sannikov is an extraordinary theorist who has developed methods that offer new insights in analyzing problems that had seemed well-studied and familiar: for example, decisions that might bring about cooperation and/or defection in a repeated-play prisoner's dilemma game, or that affect the balance of incentives and opportunism in a principal–agent relationship. His work has broken new ground in methodology, often through the application of stochastic calculus methods. The stochastic element means that his work naturally captures situations in which there is a random chance that monitoring, communication, or signaling between players is imperfect. Using calculus in the context of continuous-time games allows him to overcome tractability problems that had long hindered research in a number of areas. Previous models often abstracted from crucial economic forces in the name of tractability, but Sannikov's methods allow models to include the most important forces and thus deliver results that are much more relevant and intuitive. Sannikov's remarkable research agenda has substantially altered the toolbox available for studying dynamic games.

His early training focused on mathematics. In high school, Sannikov won three gold medals at the International Mathematical Olympiads on behalf of his native Ukraine. He studied mathematics at Princeton as an undergraduate, where he was influenced by economist Dilip Abreu and mathematician Yakov Sinai. He completed his PhD in economics at the Stanford Graduate School of Business, where he was

■ *Susan Athey is Economics of Technology Professor and Andrzej Skrzypacz is the Theodore J. Kreps Professor of Economics, both at the Stanford Graduate School of Business, Stanford, California.Their email addresses are athey@stanford.edu and skrz@stanford.edu.*

† For supplementary materials such as appendices, datasets, and author disclosure statements, see the article page at
https://doi.org/10.1257/jep.31.2.237 doi=10.1257/jep.31.2.237

Yuliy Sannikov

advised by Robert Wilson and Andrzej Skrzypacz. Beyond his immediate advisors, a number of scholars inspired and contributed to his thinking, including Michael Harrison, Peter DeMarzo, Thomas Sargent, Darrell Duffie, and Paul Milgrom.

The excellence of Sanikov's work is widely recognized. He received the John Bates Clark medal in 2016, which is awarded annually by the American Economic Association "to that American economist under the age of forty who is judged to have made the most significant contribution to economic thought and knowledge." In 2015, the American Finance Association awarded the Fischer Black Prize to Sannikov, honoring an individual researcher under age 40 "for a body of work that best exemplifies the Fischer Black hallmark of developing original research that is relevant to finance practice."

Sannikov's research introduces not only technical advances, but also qualitatively new ideas, which is perhaps the greatest accolade one can bestow on any researcher. A hallmark of his papers is that they take the existing literature to a new level, opening up new lines of inquiry, and allowing qualitatively different types of insights to be derived. This essay offers an overview of Sannikov's research in several areas. We begin with his work using continuous-time approaches, rather than the better-known discrete-time approaches, in the analysis of dynamic games and dynamic contracting. We also sketch some of his more recent work that tackled more complex models in the design of securities, market microstructure, and the role of financial crises in macroeconomics, where the greater tractability available through Sannikov's approaches has an enormous impact in rigorous theoretical

Table 1
Selected Research Papers by Yuliy Sannikov

1. “Games with Imperfectly Observable Actions in Continuous Time.” 2007. *Econometrica* 75(5): 1285–1329.
2. “A Continuous-Time Version of the Principal–Agent Problem,” 2008. *Review of Economic Studie*s 75(3): 957–84.
3. “Optimal Security Design and Dynamic Capital Structure in a Continuous-Time Agency Model,” (with Peter M. DeMarzo). 2006. *Journal of Finance* 61(6): 2681–2724.
4. “Learning, Termination and Payout Policy in Dynamic Incentive Contracts,” (with Peter M. DeMarzo). Forthcoming. *Review of Economic Studies.*
5. “Moral Hazard and Long-Run Incentives.” 2014. Working paper 3430, Stanford Graduate School of Business, https://www.gsb.stanford.edu/faculty-research/working-papers/moral-hazard-long-run-incentives.
6. “Reputation in Continuous-Time Games,” (with Eduardo Faingold). 2011. *Econometrica* 79(3): 773–876.
7. “Impossibility of Collusion under Imperfect Monitoring with Flexible Production,” (with Andrzej Skrzypacz). 2007. *American Economic Review* 97(5): 1794–1823.
8. “The Role of Information in Games with Frequent Actions,” (with Andrzej Skrzypacz). 2010. *Econometrica* 78(3): 847–82.
9. “A Macroeconomic Model with a Financial Sector,” (with Markus K. Brunnermeier). 2014. *American Economic Review* 104(2): 379–421.
10. “The I Theory of Money,” (with Markus Brunnermeier). 2014. Working paper 3431, Stanford Graduate School of Business. https://www.gsb.stanford.edu/faculty-research/working-papers/i-theory-money.

analysis of some difficult and long-standing problems. Table 1 lists a selection of Sannikov’s research papers. In this essay, we will refer to Sannikov’s papers by their number in the table, while referring to other papers using the familiar author-date method.

Overview of Sannikov’s Introduction of Continuous-Time Methodology into Dynamic Games

Game theory has traditionally been done (mostly) with discrete-time models, rather than models in which agents make choices continuously. Discrete time makes it more straightforward to define the set of admissible strategies so they map unambiguously to payoffs, and to apply definitions of perfect equilibria. For example, suppose Alice and Bob play a repeated version of the well-known Battle of the Sexes game. In this game, two players are trying to meet at either the opera or the hockey game. Although Alice prefers the opera and Bob prefers the hockey game, they both prefer to be at the same event rather than ending up at different places. Suppose they follow simple strategies of rotating where they meet: going to the opera in even periods and hockey games in odd periods. This is a well-defined pair of strategies in discrete time with a clear outcome. But in continuous time, there are no longer odd and even periods. Moreover, it is not immediately clear how to define this kind of game in continuous time to allow for infinitely frequent changes of actions as a function of time and (even more problematic than in this example), as a function of opponent's past actions.

Or in a different game (one that is known as the Rubinstein bargaining game), suppose that Chris and Dale bargain to split a dollar. Chris makes all the offers and Dale can accept or reject offers. Rejections extend the game. Players prefer early agreements because delay implies some costs (that is, players discount the future relative to the present). In discrete time, if Dale expects a positive amount in the best equilibrium for him, Chris would offer slightly less in the first period and because of discounting, Dale would accept that first offer, making Dale's expectation irrational. Therefore, the unique subgame perfect Nash equilibrium is that Chris offers to take the whole dollar and Dale accepts (Dale is indifferent to accept that offer since he expects never to be offered any amount). But in continuous time, it appears that any split of the dollar is an equilibrium: if Dale expects Chris to offer 80 percent after every history of the game, it is sequentially optimal for Dale to reject any lesser offer—because in continuous time the future arrives without any costly delay.

Of course, there have been some previous applications of continuous-time methods in strategic environments. For example, the "war of attrition" is a game in which players would benefit from outwaiting an opponent, but in which waiting is also costly, and so players need to select a time to stop waiting. Another exception involves differential games, in which the game itself evolves over time. A classic example is the "homicidal chauffeur" problem, in which the driver of a car tries to chase down a jogger. The car is faster, but the jogger is more maneuverable, which sets the conditions for the dynamics of the game to evolve. In the area of contracting, the seminal Holmstrom and Milgrom (1987) paper on "aggregation and linearity" used a continuous-time model to provide a rationale for simple linear contracts in an environment where agents can choose their effort as a function of time, and because of the continuous-time assumption, are able to observe the results of their efforts immediately, not after a delay.

With a few such exceptions duly noted, it's fair to say that before Sannikov's work, the use of continuous-time models in game theory applications was highly limited.[1] A widely recognized problem with this approach was that in repeated interactions where actions are perfectly observed, writing the model in continuous (or almost continuous) time typically creates a situation in which agents who deviate from an equilibrium path in a dynamic game can be immediately punished, which

[1] Outside of the game theory context, continuous-time models had of course been used extensively in several areas of economics, for example in the areas of asset pricing and real option theory. Most of these applications have been in the area of competitive markets or single-person decision problems, rather than strategic interactions.

eliminates any possible benefits of such deviation, and drastically reduces usefulness of the continuous-time formulation. Other related problems with continuous-time modeling are described in Simon and Stinchcombe (1989). Many researchers have proposed fixes to these problems, but it is fair to say that—until the arrival of Sannikov's work—none of these approaches made a broad impact on work in applied theory.

The key early insight in Sannikov's work was that if we introduce imperfect monitoring, a feature of many real-world environments, it becomes possible to define strategies (or contracts) as a function of the imperfect signals that agents observe. If an agent deviates, then in a situation of imperfect monitoring, it takes time for other agents to become confident of the deviation. With this assumption, opponents no longer can react instantly to a deviation by an agent. By making the models more realistic for many applications, Sannikov managed to also achieve new tractability, which then allowed him and other researchers to provide new robust economic intuition.

Sannikov's [2] first paper on strategic interactions in continuous time started when he was a second-year PhD student. He had been studying dynamic contracting models in Thomas Sargent's class, and was presenting a new analysis of the problem studied by Phelan and Townsend (1991), which was an analysis of optimal dynamic contracting with moral hazard and risk-averse agents. The weakness of that paper was that while the result was very general, analytical characterizations were not available; instead, the optimal contract needed to be computed, and it was hard to deliver economic intuition for some features of the computed solution. More specifically, the setting did not rely on the simplifications of the Holmstrom and Milgrom (1987) approach mentioned earlier, and thus the optimal contracts were much more complicated than Holmstrom and Milgrom's linear contracts. Sannikov's insight was that if we take the continuous-time limit of the optimality conditions in Phelan and Townsend (1991), and assume that the noise in output can be described as Brownian motion, the problem becomes drastically simpler.

The key element of the intuition is that in static problems where agents need to be motivated to take an action through incentive contracts, the optimal contract calls for large rewards or punishments when unusual outputs are realized. Indeed, these models predict that contracts should be highly nonlinear, a prediction that is not borne out in many real-world settings. The optimality of large rewards and punishments also arises in a discrete-time formulation of a dynamic game: signals that are very informative about an agent's action call for large changes in continuation payoffs ("continuation payoffs" are the expected present value of equilibrium payoffs in the continuation of the game). As a result, the trade-off between incentives and risk becomes convex. The optimal resolution of this trade-off at a given point of time for a given continuation payoff depends on the principal's entire value function over all continuation payoffs.

In contrast, if information about an agent's performance arrives continuously, as in Holmstrom and Milgrom (1987) and [2], then the way in which the optimal contract changes with the realization of signals today leads to only small changes in payoffs in the future (otherwise, the risk of sharp decreases in future payoffs would expose the agent to a level of risk that outweighs the benefits in terms of incentives). As a result, the trade-off between incentives and risk is local: it depends only on how the value function changes locally, and the optimal contract/set of achievable payoffs can be computed much more easily, as a solution to a differential equation. Unlike in discrete time, where a solution requires knowing what is sequentially credible for large deviations, we only need to know what is feasible locally. Sannikov demonstrated in his work how this insight can be used to provide an alternative analysis of strategic dynamic interactions, one that leads to a more unified understanding of optimal contracts. The focus on local interactions, in turn, leads to a better understanding of long-run properties of the optimal contract through stochastic differential equations (that is, a differential equation that involves both a deterministic and a random element), which are a staple of his analysis.

After demonstrating the value of using continuous-time methods to analyze limits of previously more complicated problems, Sannikov developed formalism to state a general class of problems directly in continuous time, so that a wide range of applications can directly rely on his general framework.

Foundations of Repeated Games and Economics of Relationships

Repeated games are important in the social sciences because they are the basic model we use to discuss a variety of repeated strategic interactions in which agents face tension between opportunistic and cooperative behavior. Applications range from social dilemmas to team production to tacit collusion. In a series of papers, Sannikov has advanced our understanding of equilibria in repeated games and through that work provided new insights on the economics of ongoing relationships.

Sannikov's dissertation [1] was a technical breakthrough, applying the mathematical tools of stochastic calculus to analyze repeated games. At a time when the vast majority of the literature used discrete-time models, the paper showed how continuous-time tools allowed otherwise intractable games to be analyzed elegantly and neatly. The paper characterizes the set of equilibrium values obtainable in perfect public equilibria of a repeated game with imperfect public monitoring. The key simplifying assumption is that agents choose actions in continuous time and observe public signals about opponent play via a continuous information process. The monitoring technology is modeled as a pair (one for each player) of Brownian motions with a drift controlled by the instantaneous actions of agents.

In this setup, Sannikov showed, it is possible to characterize the boundary of the set of equilibrium payoffs as the solution to an ordinary differential equation. His characterization applies for any fixed discount factor, although the situation in which agents are somewhat impatient is more challenging and more interesting than the analysis of what happens in the limit as players become more patient. In the limit, as players in these games get very patient, tradeoffs between efficiency and incentives often disappear, and all feasible payoffs can be obtained (that is, a "folk theorem" holds). In contrast, with less-patient players, real tradeoffs arise, and a characterization of the set of equilibrium payoffs highlights those tradeoffs. This work was immediately recognized as being path-breaking, a truly significant advance in a literature that had for years been characterized by incremental improvements.[2] In the class of games that Sannikov analyzes, the ordinary differential equation that describes the boundary of the set allows direct interpretation of how dynamic rewards and punishments are used to provide incentives. As Sannikov shows, in his class of games, in the equilibria with payoffs on the boundary of the achievable set, players always remain on that boundary and move continuously in response to the public news.

This technical result has applied consequences. For example, it allows us to understand better the costs to a cartel of not using direct monetary payments between its members and instead using future rewards and punishments—for example, in the form of changes in future market shares. In the optimal collusive equilibrium, if the cartel members adjust their actions frequently and monitoring is via a gradual information process, incentives are provided by small transfers of continuation payoffs in response to the observed realizations of the public signals. The more sensitive are continuation payoffs to these signals, the more high-powered are the incentives faced by firms. The higher the noise in monitoring is, the lower the power of these incentives. When incentives are provided via future reallocation of market shares, typically it reduces somewhat the future total surplus of the cartel. For example, in a repeated prisoners' dilemma, to transfer a future payoff from player 2 to player 1, we need to sometimes let player 1 defect while player 2 is cooperating, which reduces the sum of payoffs over time. This cost of reduced payoffs over time could be avoided if the cartel members engage directly in monetary transfers, for example via product purchases from each other. Harrington (2006) notes that this is commonly done by explicit cartels. Since explicit transfers have their own costs and risks for the cartel (for example, they can increase the risk of detection), for small deviations from allocated market shares, it may be optimal to provide incentives via changes in promised continuation profits (for example, letting firms catch up with their promised sales). This could explain why cartels only meet and settle-up infrequently after unusually asymmetric outcomes.[3]

[2] The previously known method for finding the set of (perfect public) equilibria of such games has been the famous "APS method," named for the Abreu, Pearce, and Stacchetti (1990). This solution method involves iterating on a set of candidate equilibrium payoffs and finding the largest fixed point of the "APS" operator (where this operator takes as input a candidate set of payoffs that can be attained in the future, and returns the set of payoffs that can be obtained if the candidate set is available in future states). The APS method has been widely used in applied work. Despite its success, because the set of equilibrium payoffs is described as a fixed point of an operator, it is often hard to provide analytical characterizations of that set or to say much about the optimal strategies supporting the equilibrium.

[3] Geometrically, this result corresponds to the set of equilibrium payoffs being strictly convex. As Sannikov shows, in the optimal equilibrium there is a one-to-one mapping between the curvature of the set and the cost of using high-power incentives: with curvature, local movements on the boundary create a downward net drift proportional to the sensitivity of payoffs to the observed outcomes. The more curved is the set of continuation payoffs, the higher the incentives to engage in direct transfers. Since curvature is small around the symmetric payoffs, monetary transfers become particularly useful after unbalanced histories that push continuation payoffs away from symmetry.

The property that incremental information affects incentives via transfers of continuation payoffs does not always hold in repeated games in discrete time and/or with other monitoring structures. In some dynamic relationships, it is optimal to react to news by radically changing continuation play. Sannikov's work shows that if actions are frequent and information arrives in small increments, the play in any optimal equilibrium evolves continuously and it is never optimal to use the threat of immediate "price wars" to sustain collusion.

This economic insight helps reconcile theory and empirical evidence. On the one hand, many real-life cartels provide incentives via transfers either through direct transfers or trading of market shares (for example, see Harrington 2006), rather than via the famous price-war mechanism (Green and Porter 1984). On the other hand, when we teach about the economics of relationships by describing cooperative equilibria in repeated prisoners' dilemmas, we introduce the "grim trigger strategy"—if one player defects once, the other player defects forever—as a way of ensuring ongoing cooperation. But most cartels (and other cooperative efforts) clearly do not operate with a grim trigger strategy where one defection ends cooperation for all time. Sannikov's work identifies an important class of situations where players act frequently and information comes continuously for which price wars do not work, because they do not provide incentives without excessively destroying value.

Why might price wars not be useful in providing incentives? In [1], that conclusion emerges magically from continuous-time modeling and the martingale representation theorem, but we may wish to see what economic intuition emerges from a concrete discrete-time setting. This is developed in Sannikov's papers [7] and [8], coauthored with Andrzej Skrzypacz. Those papers study repeated games in discrete time (using standard tools) and then take the limit of those games as agents move closer to continuous time by adjusting their actions more and more frequently. They characterize how continuous-time information can and cannot be used to provide dynamic incentives.[4]

As an example, consider two symmetric firms repeatedly choosing quantities (or team members providing effort) and suppose that market prices depend on the sum of quantities and noise. Thus, prices affect the firms' profits and they also serve as signals about otherwise unobserved quantities produced by other firms. Firms can use prices to test the hypothesis that they are maintaining collusive quantities (q, q) against an alternative that one of the firms deviates and chooses quantity $q' > q$ instead. When noise takes the form of a Brownian motion, the key property of information flow from prices is that the likelihood ratio for any such test is evolving continuously. Thus, in settings of [7] and [8], information flow is proportional to the length of the discrete time period Δ between points at which players have the opportunity to adjust their actions. Longer periods of observation provide more accurate signals. This is the

[4] The exercise is similar to that of Abreu, Milgrom, and Pearce (1991), but while that paper considered only strongly symmetric equilibria with Poisson information structures, Sannikov's work considers all pure-strategy perfect public equilibria and a mixture of Poisson and Brownian news (in [8]).

kind of monitoring technology we would get if we aggregated information as in the continuous-time formulation in [1] over periods with a finite length.

Suppose firms wanted to hold down quantities produced and collude, but they know that individual firms will be tempted to deviate secretly and to produce higher quantities. If there were no noise in prices, it would be an equilibrium strategy to maintain high prices by the grim trigger strategy: produce at the lower collusive level as long as you see no deviation, but start a price war once you see evidence that any other producer is deviating. But with noise, deviations are not directly observable. Instead, firms have to use prices to conduct statistical inference of when to start a punishment phase: for example, the solution of Green and Porter (1984) is to trigger the price war when prices happen to be unusually low.

When firms aggregate information over a discrete-time period of length Δ to decide whether to trigger a price war, two quantities are crucial to the effectiveness of collusion. The first is the *likelihood difference* of the test, which at an intuitive level is the amount by which a deviation would increase the probability of the price war. The likelihood difference is key to the strength of incentives. The second is the probability of type-I errors, which in this case is the risk of triggering a price war even when nobody has actually deviated, and the price was low only because of randomness in demand. These errors are the costs of providing incentives via price wars; they quantify the benefits of collusion that are sacrificed per period to provide incentives.

When information flow is continuous, as in [7], any test with a likelihood difference of order Δ (the length of the period) must have a large probability of type-I error when Δ is small (about of order $\Delta^{1/2}$). Thus, if collusion brings a gain in payoffs on the order of Δ per time period (of length Δ) relative to static Nash, the cost of providing incentives is on the order of $\Delta^{1/2}$. Since costs outweigh the benefits when Δ is small, players are not able to obtain payoffs higher than the static Nash equilibrium. This observation motivated the "Impossibility of Collusion" in the title of [7]. Of course, the takeaway from this paper is not that collusion is impossible, but why many cartels work by transferring payoffs, often at the increased risk of detection, rather than using price wars. The key to determining whether the threat of price wars can be effective in providing incentives depends on the probability of type-I errors compared to the length of the period.

To sum up, there is a tension between the discrete-time theory of collusion, following the work of Green and Porter (1984) and Abreu, Pearce, and Stacchetti (1986), which focuses on the role of price wars in sustaining collusion, and empirical evidence, like that in Harrington (2006), which shows that many cartels prefer to use transfers and/or trade market shares. This tension is reconciled in [7] and [8] by showing that threats of sudden price wars are ineffective when information arrives continuously and players can react to it quickly. This same observation arises in the characterization of equilibrium payoffs of [1] via an ordinary differential equation, which summarizes the cost of incentive provision via transfers of continuation values by the curvature of the equilibrium payoff set. In this characterization, price wars may happen on the equilibrium path, but only as a last resort after the possibilities to transfer payoffs have been exhausted.

These results raise some interesting follow-up questions. First, one may wonder about the Mirrlees critique of principal–agent settings—an observation that when the actions of an agent affect the mean of a normally distributed output variable, it is possible to approach first-best outcomes by imposing extreme punishments on unlikely tail events. This critique has a particularly stark manifestation in the setting of Holmstrom and Milgrom (1987). In the continuous-time model analyzed by that paper, the optimal (second-best) contract is linear—there is a real trade-off between risk and incentives. In contrast, if in the same model, periods instead had length Δ even for a very small Δ, it would be possible to approach the first-best outcome by highly nonlinear contracts that impose extreme punishments after unlikely tail events. Why does the same observation not apply to repeated games, given that [1] is cast directly in continuous time as in Holmstrom and Milgrom, while [7] and [8] are cast in discrete time, taking the time period Δ between actions to 0? It turns out that in repeated games, the Mirrlees critique does not apply for several reasons. The transfers are limited by the set of feasible payoffs, so it is not possible to use very large punishments with arbitrarily low type-I errors. More importantly, even if incentives are provided only via transfers of continuation payoffs, as in [1] and [8], it turns out that large transfers will also cause large efficiency losses (because of the curvature of the set of continuation payoffs).

Second, it matters how we take the limit of games towards continuous time—in particular how we model the monitoring technology as Δ gets smaller. There are many discrete-time processes that converge in some sense to the Brownian motion as Δ gets small. Yet for monitoring purposes, they have very different properties, and small details can have a big impact on the equilibrium. For an extreme example, we can construct two seemingly similar processes with binary signals in discrete time that converge to the same Brownian motion (with drift that depends on actions). But in the first process, deviations change only the probability distribution over the binary steps, while in the second process they also change the domain of the steps. Games with the first type of process will have very similar properties to the ones described in [1], [7], and [8]. Games with the second process will allow perfect monitoring, where collusion is always easy for small Δ. Fudenberg and Levine (2007, 2009) as well as Sadzik and Stacchetti (2015) provide other examples, in which discrete-time signal processes converge to their continuous-time counterparts, but the informativeness of the signals may or may not converge. Hence, one has to be careful in interpreting games with frequent actions. In [7] and [8], Sannikov creates per-period distributions by integrating continuous-time processes instead of taking limits of discrete-time processes, and this distinction matters for the results.[5]

[5] One can also use this reasoning to compare Poisson and Brownian monitoring technologies. This involves again comparing the *likelihood differences* and *probabilities of type I error* along the lines we described above. Roughly, if deviations of players from the equilibrium path strategies increase the arrival rate of signals (the "bad news" case), *probabilities of type I error* go down to zero at the rate of Δ, so price wars can sometimes help, while if the deviations reduce the arrival rate (the "good news" case), price wars are even less useful than in the Brownian case.

Finally, it is important to ask whether these results have any empirical support. As we mentioned before, there is indirect evidence from real-life cartels that try to collect data on individual behavior of its members and provide incentives via transfers instead of price wars. For example, Cabral (2005) describes early problems of the lysine cartel, which tried to sustain collusion without identifying potential deviators but later introduced an additional system for monitoring market shares. The facts described in these papers are consistent with the main takeaway from this area of Sannikov's research. Those models offer a much better understanding of why in economic relationships that suffer from imperfect monitoring, it is important *not* to use strategies like the grim trigger or price war threats, but to develop more complicated schemes based on trading favors. Several authors have also provided evidence that Sannikov's models may be useful for understanding experimental data (for example, Bigoni, Potters, and Spagnolo 2012), but the literature on games with frequent interactions is still in a relatively early stage. One of the problems is that in discrete-time repeated games with imperfect monitoring, it has been commonly observed that subjects tend to underreact to the introduction of noise, so it is possible that difficulties of experimental subjects in performing statistical inference would swamp the game-theoretic considerations.

Applications to Dynamic Incentives

Sannikov's models have had very fruitful applications in contract theory. In the application in [2], discussed earlier in this paper, Sannikov analyzes a continuous-time model in which an agent controls the drift of a diffusion process and the paper characterizes the optimal contract in this environment. The modeling is beautiful, and it provides new insights that were not available with existing models. As in traditional models, eliciting higher effort from an agent comes at the cost of exposing the agent to additional risk, so the underlying tradeoff is familiar. However, the dynamics are much richer in this modeling framework. For example, one result is that agents eventually "retire," either because they have had a series of bad luck leading their utility to be so low that it is too expensive to provide additional incentives because the agent can't be hurt any further; or else because they have had a series of good luck leading the utility to be so high that additional rewards are not effective relative to their cost to the principal.

These insights can be taken in a variety of directions. For example, a classic problem in game theory looks at the situation of a "large" (long-lived) individual attempting to establish a reputation with a set of "small" (short-lived) players who react to the large player's behavior. In [6], Sannikov re-examines this problem through the lens of continuous time. The paper is able to characterize equilibria for a range of discount factors. A key qualitative insight from this work is that it is possible to better understand two important forces in models of reputation. The first is the standard "repeated game" force, where the large player is disciplined by the possibility of future punishments. The second is the "reputation" force, where the large player is disciplined by the extent to which today's action affects tomorrow's beliefs. In [6], only the belief force is at play because with Brownian-noise imperfect monitoring, it is not possible to provide incentives via coordinated punishments of the short-lived players for the same reasons as it is hard to sustain collusion with the threat of price wars. Thus, the equilibria will depend only the tradeoff between the reputation of the large player and how it is affected by beliefs. In this setting, the best equilibria turn out to involve mixed strategies that are called "Markov equilibria," in that strategies depend only the current value of the payoff-relevant state variable, rather than the entire history of the game.

Other authors have been building on Sannikov's framework. For example, Bohren (2016) analyzes a market where a firm or worker provides a sequence of customers with a product or service. The worker's effort noisily impacts a payoff-relevant state variable, such as a measure of the worker's rating or a firm's quality. Again, all equilibria are Markov. The model is applied to analyze the optimal design of a platform-based rating application. The paper observes that the modeling "demonstrates the power of the continuous-time setting to deliver sharp insights and a computationally tractable equilibrium characterization in a rich class of dynamic games." It illustrates the power of Sannikov's foundational work.

Dynamic Contracts, Security Design, and Firm Financing

Continuous-time methods have been central in finance at least since the pioneering work of Black and Scholes (1973) and Merton (1973), but their use has been primarily in the area of asset pricing. Sannikov's methods for analyzing dynamic contracting models extend this technology to corporate finance questions, enabling researchers to analyze the dynamics of optimal executive compensation and security design. As noted at the start, Sannikov received the 2015 Fisher Black Prize in finance in recognition of the importance and usefulness of his contributions.

A pair of papers on finance that Sannikov wrote with Peter DeMarzo [3, 4] exemplify this work. The first paper [3] studies how optimal dynamic contracts can be implemented with a standard capital structure consisting of a credit line, long-term debt, and equity. In the model, a cash-constrained risk-neutral agent undertakes a project that experiences fluctuations in cash flows and thus requires financing by a third party. The agent's effort shifts the mean of cash flows (which can be alternatively interpreted as the agent refraining from diverting cash for private benefit), and the resulting output is observed in real time by the principal/investors. Outside investors are risk-neutral, deep-pocketed, and more patient than the agent. Investors enter into a dynamic contract with the agent in which they commit to payments and a termination rule, both as a function of the project's realized performance. To provide incentives, the optimal contract needs to give the agent some "skin in the game"—which means that total consumption of the agent needs to be sensitive to observed cash flows. This sensitivity is achieved by adjusting the agent's continuation payoff up or down in response to earnings surprises. If performance is good, the agent receives direct payments once the continuation payoff exceeds a threshold level. If performance is bad, the agent is terminated once his continuation payoff falls below a cutoff. Thus, the agent is rewarded for good performance both by accelerating the timing of payouts and reducing the risk of termination. Although termination is inefficient (the project is productive), the possibility of termination is needed to provide incentives because of the agent's limited liability (his continuation payoff cannot be reduced below his outside option). The optimal contract defers payments to the agent to reduce the risk of termination, but must balance this against the agent's relative impatience. The paper first describes the optimal evolution of the agent's continuation payoff and consumption and then describes how the optimal contract can be implemented using standard securities. Despite the complications of dynamic contracting, there exists a relatively simple implementation: the agent gets debt financing to start the project and can draw on a credit line to cover bad outcomes, but only up to a limit—and when the agent hits that limit, the project is terminated. On the other hand, once the agent pays off the credit line, the agent *chooses* to pay dividends to equity holders (which include himself).

While the qualitative implementation of the dynamic incentive contract in terms of a simple capital structure had previously been shown in a discrete-time setting by DeMarzo and Fishman (2007), the continuous-time model with Brownian shocks introduced in [3] provides a major simplification of the problem. It develops a more complete characterization of the optimal contract via an ordinary differential equation, which allows a straightforward calculation of payoffs, security valuations, and comparative statics. For example, the use of credit versus long-term debt varies with features of the environment such as volatility: the less volatile are the project cash flows, the smaller is the credit line given to the agent and hence the smaller room for error the agent has. The paper also shows that the optimal contract may require the firm to hold a compensating cash balance while borrowing (at a higher rate) through the credit line. Thus, the model provides a justification for behavior that might seem irrational absent incentive problems—the cash balance ensures that the firm will have cash flow in future states where investors may not be willing to provide funds.

An additional feature of the implementation is that the contract allows the agent to determine the firm's payout policy and choose which securities to pay off first. The agent chooses to pay off the credit line before paying dividends, but once the credit line is paid off, the agent will pay dividends rather than "hoard cash" (that is, increase the cash balance or pay off long-term debt). The paper shows that volatility affects primarily the mix of securities used by firms—more volatile firms use a larger credit line relative to long-term debt—but has a much smaller impact on the total credit available.

Finally, despite the presence of leverage, the usual conflicts between debt and equity need not arise; that is, neither equity-holders nor the agent have the incentive to increase risk, or to increase dividends to induce default on debt, or to contribute more capital to postpone default. This surprising result arises from the endogenous nature of the optimal contract. For example, because the agent is terminated and his access to funds is cut off once the credit line reaches its limit, the agent has no incentive to take on a high degree of risk. The usual intuition that firms close to bankruptcy have an incentive to take risk that might impose large losses on creditors is overturned because, with continuous Brownian-motion driven cashflows, creditors can stop the agent before the losses become too severe. While this result depends on the continuous nature of the shocks in the model, it reveals that whether asset substitution or excessive risk taking is a first-order problem in firm financing or other dynamic moral hazard problems depends on the nature of the risk: increases in continuous volatility are less important than "tail risk"—the agent taking actions that with small probability can cause dramatic losses. This line of reasoning has been developed for example by DeMarzo, Livdan, and Tchistyi (2014), Biais, Mariotti, Rochet, and Villeneuve (2010), and Varas (2013).

The follow-up study [4] enriches the model of [3] to allow for dynamic learning about the profitability of the project on behalf of both the principal and the agent. Since the agent's effort affects output, which is the source of the principal's learning, the agent has an incentive to manipulate the principal's belief about the project as well as about his effort. In particular, by shirking and reducing output today, the agent obtains an immediate private benefit and lowers the principal's expectations for the output that agent should deliver in the future. This paper develops a rich characterization of how such incentives must be controlled in the optimal contract.

The resulting model produces a very natural "life cycle" of firm dynamics. In its early stages, the firm is financially constrained, with no payouts and the potential for inefficient termination if performance is sufficiently poor. If early-stage termination is avoided, however, the project "matures" and the firm pays dividends (of which the agent receives a share). Notably, dividends are based on expected future earnings and thus are much smoother than the firm's realized earnings. The intuition for this dividend-smoothing is that surprises in current earnings will change beliefs about the firm's future prospects, making it optimal for the firm to absorb the shock to earnings via its cash reserves (for example, a positive surprise raises expectations and makes the project more valuable, so higher reserves are warranted as greater insurance against future uncertainty).

The paper also highlights that features of the optimal contract depend on whether the information the agent can manipulate is project-specific or whether it reflects the agent's general ability (in which case the agent can continue to benefit from this information in his next job, even if he is fired). In the latter case, compensating the agent with equity is sufficient, and once the firm matures, there is no longer any risk of inefficient termination. But in the former case, the paper shows that if the project does poorly in the early stages, even if the firm survives to maturity, the contract is permanently affected—dividend payments are lower, and the agent faces a permanently higher threshold for termination. These long-run distortions are actually optimal because they make it easier to provide incentives (and lower the risk of termination) in the critical early phase of the project.

In [5], Sannikov analyzes optimal contracts in an environment where the agent's effort has a long-run impact on the stochastic process of output rather than just shifting the mean of contemporaneous payoffs. Although this type of setting is obviously much more realistic than typical models where effort only affects current output, dynamic models like this have resisted analysis in the past due to tractability challenges. The paper provides a characterization of the optimal contract, which has some interesting features. First, the agent's exposure to firm risk in the optimal contract is dictated by the degree of control over current outcomes that arises through current as well as past actions. As a result, risk exposure starts small but adjusts towards a target level of risk exposure over time. Second, the contract includes consumption-smoothing features, so that incentive effects of current performance are distributed over time, both on the positive side and on the negative side, to give the most bang-for-the-buck in terms of providing incentives. Third, due to participation constraints (specifically, limited liability), pay-for-performance has bounded sensitivity, and an agent is terminated if performance is too poor.

Financial Frictions in Macroeconomic Models

Sannikov has been developing a new line of research in macroeconomics with Markus Brunnermeier [9]. This research continues the themes from his previous work, where his modeling approaches are simultaneously more realistic and more tractable than previous models. The inherent complexity of macroeconomic models enables even greater value-added in terms of being able to capture important macroeconomic phenomenon.

After the recent financial crisis, distortions in the financial sector have received renewed attention when modeling fluctuations in aggregate economic activity. The idea that financial distortions play a central role in at least some crisis periods, and that they play a role in the propagation of other types of macroeconomic disturbances as well, is not new. For example, in work that started in the 1980s, Bernanke and Gertler introduced a "financial accelerator" mechanism into dynamic stochastic general-equilibrium models of aggregate fluctuations (for a survey, see Bernanke, Gertler, and Gilchrist 1999). In order for models to be tractable, the models focused on the analysis of fluctuations of the key variable (the net worth of leveraged investors who undertake risky investments) around a constant long-run, steady-state value of aggregate output, and exogenous shocks had to be small enough not to move too far away from this steady state. This in turn allows the dynamics to be represented as approximately linear, which facilitates characterization of equilibrium investor behavior.

Although the tractability implied by linearized representations of equilibrium dynamics have allowed financial accelerator mechanisms to be studied in a variety of contexts, the simplifications also rule out some very important forces and associated phenomena. In particular, the models do not allow the possibility that severe financial crises can occur except as highly unlikely occurrences where exogenous shocks are very large. In addition, the assumptions of the models imply that shocks have the same effect no matter what the underlying state of the economy when the shock occurs, so that the models do not admit the concept of a fragile state of the financial sector of the economy. Thus, it is difficult for these models to capture some of the key stylized facts of the financial crisis.

In [9], Brunnermeier and Sannikov analyze a simple dynamic stochastic general equilibrium model with a financial accelerator mechanism that is in many ways fairly standard. The paper departs from the literature by using continuous-time methods that allow the equilibrium to be characterized by solving an ordinary differential equation without requiring any linearization of the model's dynamics. This allows the dynamics to be characterized under assumptions that do not imply the existence of a "long-run steady state" near which the dynamics are locally mean-reverting. This approach allows the authors to answer an important criticism of the earlier generation of dynamic stochastic general equilibrium models with financial accelerator mechanisms, namely that the models were unable as a quantitative matter to generate large enough aggregate fluctuations in response to aggregate shocks of a realistic size.

Like the previous literature, the key state variable in the model is the net worth of leveraged investors as a share of total wealth. The continuous-time model makes it possible to solve for the amount of time this state variable spends far away from its long-run average value. In turn, this enables analysis of how model parameters affect the frequency and duration of spells where the economy is (endogenously) in a state such that financial distortions are much larger than usual. Financial innovations for risk sharing among individuals, such as derivative contracts that improve risk sharing, or securitization, can create conditions that make financial crises more likely.

The nonlinear solution for the model dynamics also makes it possible to analyze how both average behavior and the way that behavior should respond to further small shocks change depending on the current value of the net-worth state variable. This makes it possible to identify different "phases" of a "financial cycle," and to consider how desirable policy might differ depending on the phase. As a consequence of this aspect of their analysis, the model differentiates between two sources of risk in the economy. In the first, risk is high for exogenous reasons (future disturbances have high variance), while in the second, the risk is endogenous. An example of the second type of risky state is one that emerged because leveraged investors' net worth has declined in response to past shocks. Interestingly, an increase in exogenous risk can lead to a decrease in equilibrium leverage, decreasing the second type of risk. The authors refer to this as the "volatility paradox," and argue that the period of superficially low macroeconomic risk that extended from 1985–2005 helped create conditions that lead to an increase in investor leverage, creating (endogenously) the risk of financial crisis.

In a paper [10], which is not yet published, the authors use their framework to study the effects of monetary policy. In order to do so, the model is augmented with intermediaries that take deposits and make loans. The intermediaries have three functions: 1) monitoring projects; 2) creating a diversified portfolio of projects; and 3) investing in long-term assets and issuing short-term liabilities, thus transforming the maturity of financial instruments. The model establishes that the supply of credit and liquid assets will vary with the underlying state of the economy (similar to the previous model, the state is characterized by the net worth of intermediaries). When borrowers experience shocks and default on loans, banks both reduce lending and supply less "inside money," creating further defaults, so that shocks are amplified by the response of the intermediaries. In this context, monetary policy can affect the probability of financial crises through its effect on leveraged intermediaries. This contrasts with the more traditional mechanism where changes in real interest rates affect household savings and firm investments, and also from existing studies that emphasize the role of bank reserve requirements. The authors discuss the impact of different policy responses in states of the economy characterized by risks due to excess leverage of intermediaries. The authors argue for a policy of "stealth recapitalization" through interest rate cuts, as this type of policy helps strong institutions more than weak ones, because stronger institutions attained their position of strength by hedging their risk. The strong institution carries out the hedge by buying long-term bonds whose value increases when interest rates fall. In contrast, bank bailouts undermine the incentives of banks to manage risk.

While it is early to judge the eventual impact of the arguments in this paper, the approach does bring issues that have been highlighted by the recent financial crisis and the policy response to it to the forefront. The paper further illustrates the fruitfulness of the authors' modeling approach for addressing subtle issues of considerable importance for the applied literature.

Some Concluding Thoughts

A hallmark of Sannikov's research is that his innovative approaches allow both more realistic assumptions and more realistic conclusions. In many (though certainly not all) settings, imperfect observability of actions or other key economic variables is an empirical reality. It turns out that in such settings, the technical drawbacks of continuous-time models largely disappear. Further, the improved tractability of the setup enables richer, more realistic conclusions as well, and conclusions that do not rely on limiting analysis, like looking only at limits as players become perfectly patient.

One potential criticism of Sannikov's approach is the claim that in reality, strategic interaction never takes place in truly continuous time. The point is fair enough, as long as one also recognizes that discrete-time models of strategic interaction are also a considerable simplification of reality. A benefit of Sannikov's approach is that once analysis is completed, it is possible to find discrete-time analogs of the model and results, and then one can judge if the model is missing something of first-order importance and which insights are likely to be robust. The ability to toggle back and forth between discrete-time and continuous-time models, and to consider the difference between them, is a considerable advance.

Another possible drawback of Sannikov's general approach is that some functional form assumptions are typically required: for example, a common assumption is that the random variation in key economic quantities follows a Brownian motion. However, we believe that in this class of problems, these functional form assumptions are a small price to pay. Economic modeling always involves some simplifications of reality, and the assumptions we make need to be judged on their relative realism and on the power of insights they bring us. Many dynamic agency models become so complicated that they fail to deliver clear intuitions. Sannikov's work has demonstrated that while at first it may seem that continuous-time methods are more complicated than discrete-time methods, after some initial investment they often deliver huge improvements in tractability. His work and the literatures that it is inspiring are allowing researchers to develop new and clarifying intuitions for a variety of problems in which we want to capture agency issues in dynamic settings.

■ *We are grateful to Michael Woodford for his insights surrounding Sannikov's work on financial frictions and to Peter DeMarzo for reviewing a draft of this paper.*

References

Abreu, Dilip, Paul Milgrom, and David Pearce. 1991. "Information and Timing in Repeated Partnerships." *Econometrica: Journal of the Econometric Society* 59(6): 1713–33.

Abreu, Dilip, David Pearce, and Ennio Stacchetti. 1986. "Optimal Cartel Equilibria with Imperfect Monitoring." *Journal of Economic Theory* 39(1): 251–69.

Abreu, Dilip, David Pearce, and Ennio Stacchetti. 1990. "Toward a Theory of Discounted Repeated Games with Imperfect Monitoring." *Econometrica: Journal of the Econometric Society* 58(5): 1041–63.

Bernanke, Ben S., Mark Gertler, and Simon Gilchrist. 1999. "The Financial Accelerator in a Quantitative Business Cycle Framework." Chap. 21 in *Handbook of Macroeconomics,* vol. 1, edited by John B. Taylor and Michael Woodford, 1341–93. Elsevier.

Biais, Bruno, Thomas Mariotti, Jean-Charles Rochet, and Stéphane Villeneuve. 2010. "Large Risks, Limited Liability, and Dynamic Moral Hazard." *Econometrica* 78(1): 73–118.

Bigoni, Maria, Johannes (Jan) J. M. Potters, and Giancarlo Spagnolo. 2012. "Flexibility and Collusion with Imperfect Monitoring." CEPR Discussion Paper DP8877. Available at SSRN: http://ssrn.com/abstract=2034095.

Black, Fischer, and Myron Scholes. 1973. "The Pricing of Options and Corporate Liabilities." *Journal of Political Economy* 81(3): 637–54.

Bohren, Aislinn. 2016. "Using Persistence to Generate Incentives in a Dynamic Moral Hazard Problem." https://ssrn.com/abstract=2889017.

Cabral, Luís M. B. 2005. "Collusion Theory: Where to Go Next?" *Journal of Industry, Competition and Trade* 5(3–4): 99–206.

DeMarzo, Peter M., and Michael J. Fishman. 2007. "Optimal Long-Term Financial Contracting." *Review of Financial Studies* 20(6): 2079–2128.

DeMarzo, Peter M., Dmitry Livdan, and Alexei Tchistyi. 2014. "Risking Other People's Money: Gambling, Limited Liability, and Optimal Incentives." Working Paper 3149, Graduate School of Business, Stanford University.

Fudenberg, Drew, and David K. Levine. 2007. "Continuous Time Limits of Repeated Games with Imperfect Public Monitoring." *Review of Economic Dynamics* 10(2): 173–92.

Fudenberg, Drew, and David K. Levine. 2009. "Repeated Games with Frequent Signals." *Quarterly Journal of Economics* 124(1): 233–65.

Green, Edward J., and Porter, Robert H. 1984. "Noncooperative Collusion under Imperfect Price Information." *Econometrica: Journal of the Econometric Society* 52(1): 87–100.

Harrington, Joseph E., Jr. 2006. "How Do Cartels Operate?" In *Foundations and Trends in Microeconomics* 2(1): 1–105.

Holmstrom, Bengt, and Paul Milgrom. 1987. "Aggregation and Linearity in the Provision of Intertemporal Incentives." *Econometrica: Journal of the Econometric Society* 55(2): 303–28.

Merton, Robert C. 1973. "Theory of Rational Option Pricing." *Bell Journal of Economics and Management Science* 4(1): 141–83.

Phelan, Christopher, and Robert M. Townsend. 1991. "Computing Multi-Period, Information-Constrained Optima." *Review of Economic Studies* 58(5): 853–81.

Sadzik, Tomasz, and Ennio Stacchetti. 2015. "Agency Models with Frequent Actions." *Econometrica* 83(1): 193–237.

Simon, Leo K., and Maxwell B. Stinchcombe. 1989. "Extensive Form Games in Continuous Time: Pure Strategies." *Econometrica* 57(5): 1171–1214.

Varas, Felipe. 2013. "Contracting Timely Delivery with Hard to Verify Quality." August 26. Available at SSRN: http://ssrn.com/abstract=2317978.

Journal of Economic Perspectives—Volume 31, Number 2—Spring 2017—Pages 257–264

Recommendations for Further Reading

Timothy Taylor

This section will list readings that may be especially useful to teachers of undergraduate economics, as well as other articles that are of broader cultural interest. In general, with occasional exceptions, the articles chosen will be expository or integrative and not focus on original research. If you write or read an appropriate article, please send a copy of the article (and possibly a few sentences describing it) to Timothy Taylor, preferably by email at taylort@macalester.edu, or c/o *Journal of Economic Perspectives,* Macalester College, 1600 Grand Ave., St. Paul, MN 55105.

Smorgasbord

The OECD focuses on *Tackling Wasteful Health Care Spending*. From the "Foreword": "Across OECD countries, a significant share of health care system spending and activities are wasteful at best, and harm our health at worst. One in ten patients in OECD countries is unnecessarily harmed at the point of care. More than 10% of hospital expenditure is spent on correcting preventable medical mistakes or infections that people catch in hospitals. One in three babies is delivered by caesarean section, whereas medical indications suggest that C-section rates should be 15% at most. Meanwhile, the market penetration of generic pharmaceuticals—drugs with effects equivalent to those of branded products but typically sold at lower prices—ranges between 10–80% across OECD countries. And a third of OECD citizens consider the health sector to be corrupt or even extremely corrupt. At a

■ *Timothy Taylor is Managing Editor, Journal of Economic Perspectives, based at Macalester College, Saint Paul, Minnesota. He blogs at http://conversableeconomist.blogspot.com.*

† For supplementary materials such as appendices, datasets, and author disclosure statements, see the article page at
https://doi.org/10.1257/jep.31.2.257 doi=10.1257/jep.31.2.257

time when public budgets are under pressure worldwide, it is alarming that around one-fifth of health expenditure makes no or minimal contribution to good health outcomes. … Actions to tackle waste are needed in the delivery of care, in the management of health services, and in the governance of health care systems." January 2017, at http://www.oecd-ilibrary.org/social-issues-migration-health/tackling-wasteful-spending-on-health_9789264266414-en.

Arthur Turrell presents an overview of "Agent-Based Models: Understanding the Economy from the Bottom Up." "Agent-based models explain the behaviour of a system by simulating the behaviour of each individual 'agent' within it. These agents and the systems they inhabit could be the consumers in an economy, fish within a shoal, particles in a gas, or even galaxies in the Universe. … The agent-based approach to problem-solving began in the physical sciences but has now spread to many other disciplines including biology, ecology, computer science and epidemiology. … Despite being less widely used, agent-based models have produced many important insights in economics, including how the statistics observed in financial markets arise, and how business cycles occur. … With respect to central banks, there are three particularly promising areas of development for agent-based modelling. The first is the ongoing application of macroeconomic agent-based models to monetary policy. Several models which explicitly include central banks have now been established and are on hand to examine policy questions. The second is in modelling the banking and financial sector, to determine how financial stress is transmitted through the system as a whole. Third, researching the potential impact of the introduction of a central bank digital currency could be explored using an agent-based model." *Quarterly Bulletin*, Bank of England, 2016 Q4, pp. 173–188, at http://www.bankofengland.co.uk/publications/Pages/quarterlybulletin/2016/q4/a2.aspx.

Karl Alexander and Stephen L. Morgan have written "The Coleman Report at Fifty: Its Legacy and Implications for Future Research on Equality of Opportunity," which is an introductory essay for a 13-paper special issue on this topic in the *Russell Sage Foundation Journal of the Social Sciences*. "In thumbnail, EEO [*Equality of Educational Opportunity*] concluded that 1) differences across schools in average achievement levels were small compared to differences in achievement levels within schools; 2) the differences in achievement levels detected did not align appreciably with differences in school resources other than the socioeconomic makeup of the student body; and 3) family background factors afforded a much more powerful accounting of achievement differences than did any and all characteristics of the schools that children attended." The other papers discuss how evidence on these findings has evolved over time. September 2016, vol. 2, no. 5, pp. 1–16, http://www.rsfjournal.org/toc/rsf/2/5.

The *World Development Report 2017* from the World Bank has the theme "Governance and the Law." "Contrary to what many growth theories predict, there is no tendency for low- and middle-income countries to converge toward high-income countries. … As ideas and resources spread at an increasingly rapid rate across countries, policy solutions to promote further progress abound. However, policies that should be effective in generating positive development outcomes are often not adopted, are poorly implemented, or end up backfiring over time. Although the development community has focused a great deal of attention on learning what policies and interventions are needed to generate better outcomes, it has paid much less attention to learning why those approaches succeed so well in some contexts but fail to generate positive results in others. … Ultimately, confronting the challenges faced by today's developing countries—poor service delivery, violence, slowing growth, corruption, and the 'natural resource curse,' to name a few—requires rethinking the process by which state and nonstate actors interact to design and implement policies, or what this Report calls governance …" January 2017, https://openknowledge.worldbank.org/handle/10986/25880.

Michael D. Giandrea and Shawn A. Sprague discuss "Estimating the U.S. Labor Share." "Keynes and other economists had accepted as fact that the share of national output accruing to workers as compensation was relatively constant. In fact, this idea had become so well accepted that some economists even began using the supposed constancy as an issue to be addressed in theories of production and economic growth. The term 'Bowley's Law,' referring to a 19th-century economist who had compiled statistics on the issue, was even coined to describe this stability. … However, in the late 20th century—after many decades of relative stability—the labor share began to decline in the United States and many other economically advanced nations, and in the early 21st century it fell to unprecedented lows. … The material that follows reviews the BLS methodology for estimating the labor share, discusses the uses and limitations of this measure, and suggests potential improvements in that methodology." *Monthly Labor Review*, February 2017, https://www.bls.gov/opub/mlr/2017/article/estimating-the-us-labor-share.htm.

Josh Lerner and Antoinette Schoar discuss "Rise of the Angel Investor: A Challenge to Public Policy." "Angel investors are high-net-worth individuals, often (but not exclusively) former entrepreneurs and corporate executives, who make private investments in start-up companies with their own money. … Angels typically invest at the seed funding stage, making them among the first equity investors in a company beyond its founders. … Angels invested a total of $24.6 billion in 2015 with an average deal size of $345,390, according to the Center for Venture Research. … The Angel Capital Association (ACA) lists more than 300 U.S. groups in its database. The average ACA angel group in 2015 had 68 member angels and invested a total of nearly $2.5 million in 10.3 deals in 2007. At least between 10,000 and 15,000 angels are believed to belong to angel groups in the U.S. … [E]stimates suggest that the total size of angel investment has long surpassed venture capital investment in the U.S. and increasingly in some other countries as well. For instance, survey estimates suggest the projected size of the total angel market in the U.S. grew from $17.6 billion in 2009 to $24.1 billion in 2014. The estimated capital deployed by angel groups in Europe has almost doubled over the past five years, and in Canada, it almost tripled. … But despite their rapid growth, we know very little about the role that angels play internationally and the type of firms in which they invest." *Third Way*, September 23, 2016, http://www.thirdway.org/report/rise-of-the-angel-investor-a-challenge-to-public-policy.

Angles on Productivity and Growth

Gary Clyde Hufbauer and Zhiyao (Lucy) Lu make a case for “Increased Trade: A Key to Improving Productivity.” “[A] $1 billion increase in two-way trade increases potential GDP, through supply-side efficiencies, by $240 million. … Between 1990 and 2008, real US two-way trade in nonoil goods and services increased at an average rate of 5.86 percent a year. If two-way trade had increased at this pace after 2011, the real value of US two-way nonoil trade in 2014 would have been $308 billion greater than the observed value ($4.50 trillion versus $4.19 trillion). Based on the average dollar ratio of 0.24, the hypothetical increase in US two-way trade would have delivered a $74 billion increase in US GDP through supply-side efficiencies in 2014.” Peterson Institute for International Economics Policy Brief 16-15, October 2016, https://piie.com/system/files/documents/pb16-15.pdf.

Vincent Aussilloux, Agnès Bénassy-Quéré, Clemens Fuest, and Guntram Wolff discuss “Making the Best of the European Single Market.” “Applying the synthetic counterfactuals method to various EU enlargements, Campos et al (2014) find that ‘*per capita European incomes in the absence of the economic and political integration process would have been on average 12 per cent lower today, with substantial variations across countries, enlargements as well as over time*’. This average figure is within the range found in the limited and fragile literature on this issue (5 to 20 percent, depending on the study). … Still, trade between European countries is estimated to be about four times less than between US states once the influence of language and other factors like distance and population have been corrected for. For goods, non-tariff obstacles to trade are estimated to be around 45 percent of the value of trade on average, and for services, the order of magnitude is even higher. If the intensity of trade between member states could be doubled from a factor of 1/4 to a factor of 1/2 in order to narrow the gap with US states, it could translate into an average 14 percent higher income for Europeans …” *Policy Contribution*, Bruegel, Issue No. 3, 2017, http://bruegel.org/wp-content/uploads/2017/02/PC-03-2017-single-market-010217-.pdf.

Bart van Ark discusses “The Productivity Paradox of the New Digital Economy.” “This article has argued that there are good reasons to believe that the New Digital Economy is still in the installation phase producing only random and localized gains in productivity in certain industries and geographies. … [W]e do not expect large aggregate growth effects from the New Digital Economy any time soon …” “What’s more, we find that when looking at the top half of industries which represent the most intensive users of digital technology (measured by their purchases of ICT [information and communications technology] assets and services relative to GDP) have collectively accounted for the largest part of the slowdown in productivity growth in all three economies since 2007, namely for 60 per cent of the productivity slowdown in the United States, 66 per cent of the slowdown in Germany, and 54 per cent of the slowdown in the United Kingdom. In the United States the contribution of the most intensive ICT-using industries declined from 46 per cent to 26 per cent of aggregate productivity growth between both periods. … The fact that ICT intensive users account for a larger part of the slowdown than less-intensive ICT users is another indication that the difficulty of absorbing the technology effectively is part of the explanation for the productivity slowdown." *International Productivity Monitor*, Fall 2016, pp. 3–18, http://www.csls.ca/ipm/31/vanark.pdf.

Claudio Borio compares "Secular Stagnation or Financial Cycle Drag?" and opts for the latter. "The [secular stagnation] hypothesis is quite compelling in some respects, but even a cursory look at the facts raises some questions. The hypothesis was originally developed for the United States, a country that posted a large current account deficit even pre-crisis—hardly a symptom of domestic demand deficiency. True, US growth pre-crisis was not spectacular, but it was not weak either—recall how people hailed the Great Moderation, an era of outstanding performance. Likewise, the world as a whole saw record growth rates and low unemployment rates—again, hardly a symptom of global demand deficiency. Finally, recent declines in unemployment rates to historical averages—and, in some cases, such as the United States, close to estimates of full employment—point to supply, rather than demand, constraints on growth. At the same time, a number of specific pieces of evidence support the financial cycle drag hypothesis. First, there is plenty of evidence that banking crises, which occur during financial busts, cause very long-lasting damage to the economy. They result in permanent output losses, so that output may regain its pre-crisis long-term growth trend but evolves along a lower path. There is also evidence that recoveries are slower and more protracted. … Second, BIS research has found evidence that financial (credit) booms tend to undermine productivity growth, further helping to explain the post-crisis weakness … Third, measures of output gaps used in policymaking now show that output was indeed above potential pre-crisis. … The reason is simple: the symptom of unsustainable expansion was not rising inflation, which stayed low and stable, but the buildup of financial imbalances, in the form of unusually strong and persistent credit growth and property price increases." Keynote speech at the Economic Policy Conference of the National Association for Business Economics, March 5–7, 2017. http://www.bis.org/speeches/sp170307.pdf.

Reports about India

The *Economic Survey 2016–2017* from India's Ministry of Finance, where Arvind Subramanian is the Chief Economic Adviser, offers perspective on a wide range issues involving the economy of India. From the opening chapter, "Eight Interesting Facts About India": "New estimates based on railway passenger traffic data reveal annual work-related migration of about 9 million people, almost double what the 2011 Census suggests." "From 2009 to 2015, China's credit-to-GDP soared from about 142 percent to 205 percent and its growth decelerated. The contrast with India's indicators is striking." "Welfare spending in India suffers from misallocation … The districts accounting for the poorest 40% receive 29% of the total funding." "India has 7 taxpayers for every 100 voters ranking us 13th amongst 18 of our democratic G-20 peers." "India's share of working age to non-working age population will peak later and at a lower level than that for other countries but last longer." "As of 2011, India's openness—measured as the ratio of trade in goods and services to GDP has far overtaken China's … India's internal trade to GDP is also comparable to that of other large countries and very different from the caricature of a barrier-riddled economy." "Spatial dispersion in income is still rising in India in the last decade (2004–14), unlike the rest of the world and even China." "Evidence from satellite data indicates that Bengaluru and Jaipur collect only between 5% to 20% of their potential property taxes." January 2017, http://finmin.nic.in/indiabudget2017-2018/e_survey.asp.

V. Anantha Nageswaran and Gulzar Natarajan ask *Can India Grow? Challenges, Opportunities, and the Way Forward.* From the "Summary": "Despite India's impressive economic growth rates in the mid-2000s, the long-term magnitude and sustainability of this progress remains uncertain. … • India's high-growth phase of 2003–2008 had much to do with growth-friendly global economic conditions that have since run their course. • The country's domestic structural deficiencies—namely poor human resource capabilities; a narrow and predominantly informal industrial base; and a fragmented, low-productivity primary sector—keep a lid on growth and a floor on inflation. • India also faces formidable long-term headwinds due to premature deindustrialization, the limitations of a services-led growth model, the plateauing of global trade, stagnation in developed economies, and the costs associated with climate change. • The country's state capacity deficiencies amplify the effects of these constraints." Carnegie India, November 2016, http://carnegieendowment.org/files/CEIP_CanIndiaGrow_Final_.pdf.

Reports about Sub-Saharan Africa

A team of World Bank researchers led by Somik Vinay Lall, J. Vernon Henderson, and Anthony J. Venables have published *Africa's Cities: Opening Doors to the World.* "In principle, cities should benefit businesses and people through increased economic density. … In sum, the ideal city can be viewed economically as an efficient labor market that matches employers and job seekers through connections. The typical African city fails in this matchmaker role. A central reason for this failure—one that has not yet been sufficiently recognized—is that the city's land use is fragmented. Its transport infrastructure is insufficient, and too much of its development occurs through expansion rather than infill. … And without the economic density that gives rise to efficiency, Africa's cities do not seem to increase worker productivity. … Cities in Africa are costly for households, workers, and businesses. Because food and building costs are high, families can hardly remain healthy or afford decent housing. Because commuting by vehicle is not only slow but expensive, workers find it hard to take and keep jobs that match their skills. And the need for higher wages to pay higher living costs makes firms less productive and competitive, keeping them out of tradable sectors. As a result, African cities are avoided by potential regional and global investors and trading partners. … When urban costs drive nominal wages too high, firms will not be able to compete in the tradable sector and will produce only nontradables. … The reason why a firm in the nontradable sector can afford to pay higher wages —while a firm in the tradable sector cannot—is that the nontradable producer can raise its prices citywide. By doing so, it passes its own cost increases on to consumers in the urban market. But such price hikes make the cost of living in a city even higher, contributing to the workers' urban costs. This sequence can become a vicious cycle that keeps African cities out of the tradable sector and limits their economic growth." February 2017, https://www.worldbank.org/en/news/video/2017/02/09/africas-cities-opening-doors-to-the-world.

The *OECD–FAO Agricultural Outlook 2016–2025* devotes a chapter to "Agriculture in Sub-Saharan Africa: Prospects and Challenges for the Next Decade." "The high contribution of the agricultural sector to GDP also underlines the limited diversification of most African economies. On average, agriculture contributes 15% of total GDP, however it ranges from below 3% in Botswana and South Africa to more than 50% in Chad … Agriculture employs more than half of the total labour force and within the rural population, provides a livelihood for multitudes of small-scale producers. Smallholder farms constitute approximately 80% of all farms in SSA [sub-Saharan Africa] and employ about 175 million people directly. … [R]ecent surveys suggest that agriculture is also the primary source of livelihood for 10% to 25% of urban households. … The African model of agricultural growth differed significantly from that of Asia or South America. In Asia, growth was driven largely by intensification, whereas in South America, it was the result of significant improvement in labour productivity arising from mechanisation. By contrast, strong growth in SSA agricultural output has accrued predominantly from area expansion and intensification of cropping systems, as opposed to large-scale improvement in productivity. … [P]roductivity per agricultural worker has improved by a factor of only 1.6 in Africa over the past 30 years, compared to 2.5 in Asia. … Arguably the greatest challenge facing the agricultural sector in SSA is weak infrastructure including transportation networks, access to energy, irrigation systems and stockholding facilities." July 2016, http://www.oecd-ilibrary.org/agriculture-and-food/oecd-fao-agricultural-outlook-2016_agr_outlook-2016-en.

Discussion Starters

Anja Shortland explores "Governing Kidnap for Ransom: Lloyd's as a 'Private Regime.'" "Kidnapping is a major (if largely hidden) criminal market, with an estimated total turnover of up to US$1.5 billion a year. … Commercially, kidnap insurance is only viable under three (related) conditions. First, kidnaps should be nonviolent and detentions short—otherwise, individuals and firms withdraw from high-risk areas. Second, insurance premia must be affordable. Although insurance is only demanded if people are concerned about kidnapping, actual kidnaps must be rare, and ransoms affordable. Insurers struggle in kidnapping hotspots: High premia deter potential customers. … Third, ransoms and kidnap volumes must be predictable and premium income must cover (expected) losses. If kidnapping generates supernormal profits, more criminals enter the kidnap business. Premium ransoms quickly generate kidnapping booms. Insurers, therefore, have a common interest in ordering transactions and preventing ransom inflation. … [K] idnap insurance is indeed controlled by a single enterprise: Lloyd's of London. Yet within Lloyd's there are around 20 international syndicates underwriting kidnap for ransom insurance. The syndicates compete for business according to clear protocols regarding how insurance contracts are structured, how information is (discreetly) exchanged, and how ransom negotiations are conducted." *Governance,* April 2017, vol. 30, no. 2, pp. 283–299, http://onlinelibrary.wiley.com/doi/10.1111/gove.12255/full.

A group of 16 co-authors led by Michael A. Rees and including Alvin E. Roth offer a proposal for "Kidney Exchange to Overcome Financial Barriers to Kidney Transplantation." "Recent worldwide estimates suggest that 2–7 million people died prematurely in 2010 because they did not have access to renal replacement therapy (RRT). … Organ shortage is the major limitation to kidney transplantation in the developed world. Conversely, millions of patients in the developing world with end-stage renal disease die because they cannot afford renal replacement therapy—even when willing living kidney donors exist. This juxtaposition between countries with funds but no available kidneys and those with available kidneys but no funds prompts us to propose an exchange program using each nation's unique assets. Our proposal leverages the cost savings achieved through earlier transplantation over dialysis to fund the cost of kidney exchange between developed-world patient–donor pairs with immunological barriers and developing-world patient–donor pairs with financial barriers. By making developed-world health care available to impoverished patients in the developing world, we replace unethical transplant tourism with global kidney exchange—a modality equally benefitting rich and poor. We report the 1-year experience of an initial Filipino pair, whose recipient was transplanted in the United States with an American donor's kidney at no cost to him. The Filipino donor donated to an American in the United States through a kidney exchange chain. Follow-up care and medications in the Philippines were supported by funds from the United States." *American Journal of Transplantation,* March 2017, pp. 782–90.

Congratulations To The 2017 Elected AEA Officers

PRESIDENT-ELECT

OLIVIER
BLANCHARD

Robert M. Solow Professor of Economics, Emeritus, Massachusetts Institute of Technology, and Fred Bergsten Senior Fellow, Peterson Institute for International Economics

VICE-PRESIDENTS

ALAN B.
KRUEGER

Bendheim Professor of Economics and Public Affairs, Princeton University

VALERIE A.
RAMEY

Professor of Economics, University of California, San Diego

EXECUTIVE COMMITTEE

NICHOLAS
BLOOM

Eberle Professor of Economics, Stanford University

ERICA
FIELD

Professor of Economics and Global Health, Duke University

AMERICAN ECONOMIC ASSOCIATION

American Economic Association

www.vanderbilt.edu/AEA

More than 130 Years of Encouraging Economic Research

The *JOE Network* fully automates the hiring process for the annual economics job market cycle.

NEW!
CANDIDATE
VIDEOS
&
INTERVIEW
SCHEDULING

For:

JOB CANDIDATES

- Search and Save Jobs
- Create a Custom Profile
- Manage Your CV and Applications
- Get the Attention of Hiring Committees
- Apply for Multiple Jobs from One Site
- Request Reference Letters

EMPLOYERS

- Post and Manage Job Openings
- Search Candidate Profiles
- Manage Applications and Materials
- Collect Reference Letters
- Download Applicant Data
- Share Candidate Materials

FACULTY

- Manage Letter Requests
- Upload Custom or Default Letters
- Track Task Completion Status
- Assign Surrogate Access
- Minimize Time Investment

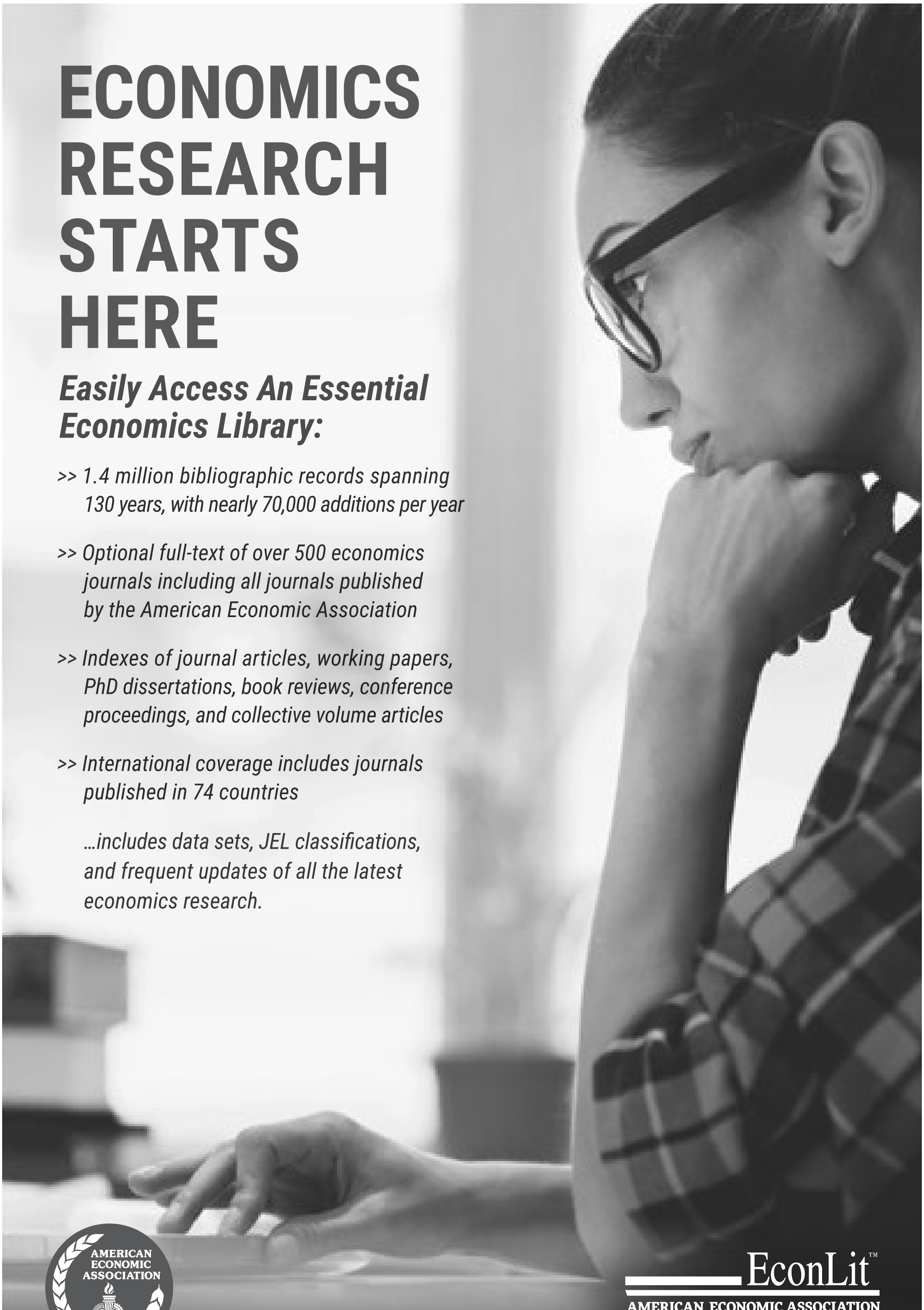
This hiring season, take advantage of the AEA's enhanced JOE Network targeted to the comprehensive needs of all participants in the annual economics job market cycle.

The ***JOE Network*** automates the hiring process. Users share materials, communicate confidentially, and take advantage of new features to easily manage their files and personal data. Everything is securely maintained and activated in one location. The JOE Network is accessible right from your desktop at the AEA website.

Experience the same great results with more features, more time savings, and a beginning-to-end process.

AMERICAN
ECONOMIC
ASSOCIATION

Try the JOE Network today! **www.aeaweb.org/JOE**



Compliments of the American Economic Association

Webcasts of Selected Sessions from the 2017 AEA Annual Meeting . . .

Now available on the AEA Website

January 6, 2017

- **Brexit: Six Months Later**
  Olivier Blanchard
  Jonathan Portes
  Andrew Lilico
  Karl Whelan

- **Nobels on Where is the World Economy Headed?**
  Presiding: *Dominick Salvatore*
  Where in the World Is the World Headed? *Angus Deaton*
  Seeking Political Keys for Economic Growth *Roger Myerson*
  How the Left and Right Are Failing the West *Edmund Phelps*
  Economic Risks Associated with Deep Change in Technology *Robert J. Shiller*
  New Divisions in the World Economy *Joseph E. Stiglitz*

- **AEA/AFA Joint Luncheon - Will the Market Fix the Market?**
  Eric Budish, introduced by *Alvin E. Roth*

- **Gender Agenda**
  Presiding: Muriel Niederle
  Quantifying the Disincentive Effects of Joint Taxation on Married Women's Labor Supply
  Alexander Bick and Nicola Fuchs-Schuendeln
  Long Hours and Women's Job Choices: Cross Country and Within United States Evidence
  Patricia Cortes and Jessica Pan
  The Expanding Gender Earnings Gap: Evidence from the LEHD-Census
  Claudia Goldin, Sari Pekkala Kerr, Claudia Olivetti, and Erling Barth
  Competitiveness and Education Choices *Muriel Niederle*
  Discussant: *Erik Hurst*

- **AEA Richard T. Ely Lecture: The Economist as Plumber: Large Scale Experiments to Inform the Details of Policy Making**
  Esther Duflo, introduced by Alvin E. Roth

The American Economic Association

Correspondence relating to advertising, business matters, permission to quote, or change of address should be sent to the AEA business office: aeainfo@vanderbilt.edu. Street address: American Economic Association, 2014 Broadway, Suite 305, Nashville, TN 37203. For membership, subscriptions, or complimentary *JEP* for your e-reader, go to the AEA website: http://www.aeaweb.org. Annual dues for regular membership are $20.00, $30.00, or $40.00, depending on income; for an additional fee, you can receive this journal, or any of the Association's journals, in print. Change of address notice must be received at least six weeks prior to the publication month.

FSC
www.fsc.org
MIX
Paper from
responsible sources
FSC™ C132124

The Journal of *Economic Perspectives*

Spring 2017, Volume 31, Number 2

Symposia

Recent Ideas in Econometrics

Are Measures of Economic Growth Biased?

Articles

AMERICAN ECONOMIC ASSOCIATION