The Journal of

Economic Perspectives

A journal of the
American Economic Association

Spring 2014

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 R. R. Donnelley Company, Jefferson City, Missouri, 65109, USA.

THE JOURNAL OF ECONOMIC PERSPECTIVES (ISSN 0895-3309), Spring 2014, Vol. 28, 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 http://www.vanderbilt.edu/AEA/.
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 28 • Number 2 • Spring 2014

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 28, Number 2—Spring 2014—Pages 3–28

Big Data: New Tricks for Econometrics†

Hal R. Varian

Computers are now involved in many economic transactions and can capture data associated with these transactions, which can then be manipulated and analyzed. Conventional statistical and econometric techniques such as regression often work well, but there are issues unique to big datasets that may require different tools.

First, the sheer size of the data involved may require more powerful data manipulation tools. Second, we may have more potential predictors than appropriate for estimation, so we need to do some kind of variable selection. Third, large datasets may allow for more flexible relationships than simple linear models. Machine learning techniques such as decision trees, support vector machines, neural nets, deep learning, and so on may allow for more effective ways to model complex relationships.

In this essay, I will describe a few of these tools for manipulating and analyzing big data. I believe that these methods have a lot to offer and should be more widely known and used by economists. In fact, my standard advice to graduate students these days is go to the computer science department and take a class in machine learning. There have been very fruitful collaborations between computer scientists and statisticians in the last decade or so, and I expect collaborations between computer scientists and econometricians will also be productive in the future.

■ *Hal Varian is Chief Economist, Google Inc., Mountain View, California, and Emeritus Professor of Economics, University of California, Berkeley, California. His email address is hal@ischool.berkeley.edu.*

† To access the Appendix and disclosure statements, visit
http://dx.doi.org/10.1257/jep.28.2.3 doi=10.1257/jep.28.2.3

Tools to Manipulate Big Data

Economists have historically dealt with data that fits in a spreadsheet, but that is changing as new more-detailed data becomes available (see Einav and Levin 2013, for several examples and discussion). If you have more than a million or so rows in a spreadsheet, you probably want to store it in a relational database, such as MySQL. Relational databases offer a flexible way to store, manipulate, and retrieve data using a Structured Query Language (SQL), which is easy to learn and very useful for dealing with medium-sized datasets.

However, if you have several gigabytes of data or several million observations, standard relational databases become unwieldy. Databases to manage data of this size are generically known as "NoSQL" databases. The term is used rather loosely, but is sometimes interpreted as meaning "not only SQL." NoSQL databases are more primitive than SQL databases in terms of data manipulation capabilities but can handle larger amounts of data.

Due to the rise of computer-mediated transactions, many companies have found it necessary to develop systems to process billions of transactions per day. For example, according to Sullivan (2012), Google has seen 30 trillion URLs, crawls over 20 billion of those a day, and answers 100 billion search queries a month. Analyzing even one day's worth of data of this size is virtually impossible with conventional databases. The challenge of dealing with datasets of this size led to the development of several tools to manage and analyze big data.

A number of these tools are proprietary to Google, but have been described in academic publications in sufficient detail that open-source implementations have been developed. Table 1 contains both the Google name and the name of related open-source tools. Further details can be found in the Wikipedia entries associated with the tool names.

Though these tools can be run on a single computer for learning purposes, real applications use large clusters of computers such as those provided by Amazon, Google, Microsoft, and other cloud-computing providers. The ability to rent rather than buy data storage and processing has turned what was previously a fixed cost of computing into a variable cost and has lowered the barriers to entry for working with big data.

Tools to Analyze Data

The outcome of the big-data processing described above is often a "small" table of data that may be directly human readable or can be loaded into an SQL database, a statistics package, or a spreadsheet. If the extracted data is still inconveniently large, it is often possible to select a subsample for statistical analysis. At Google, for example, I have found that random samples on the order of 0.1 percent work fine for analysis of business data.

Once a dataset has been extracted, it is often necessary to do some exploratory data analysis along with consistency and data-cleaning tasks. This is something

Table 1
Tools for Manipulating Big Data

Google name	*Analog*	*Description*
Google File System	Hadoop File System	This system supports files so large that they must be distributed across hundreds or even thousands of computers.
Bigtable	Cassandra	This is a table of data that lives in the Google File System. It too can stretch over many computers.
MapReduce	Hadoop	This is a system for accessing and manipulating data in large data structures such as Bigtables. MapReduce allows you to access the data in parallel, using hundreds or thousands of machines to extract the data you are interested in. The query is "mapped" to the machines and is then applied in parallel to different shards of the data. The partial calculations are then combined ("reduced") to create the summary table you are interested in.
Sawzall	Pig	This is a language for creating MapReduce jobs.
Go	None	Go is flexible open-source, general-purpose computer language that makes it easier to do parallel data processing.
Dremel, BigQuery	Hive, Drill, Impala	This is a tool that allows data queries to be written in a simplified form of of Structured Query Language (SQL). With Dremel it is possible to run an SQL query on a petabtye of data (1,000 terabytes) in a few seconds.

of an art, which can be learned only by practice, but data-cleaning tools such as OpenRefine and DataWrangler can be used to assist in data cleansing.

Data analysis in statistics and econometrics can be broken down into four categories: 1) prediction, 2) summarization, 3) estimation, and 4) hypothesis testing. Machine learning is concerned primarily with prediction; the closely related field of data mining is also concerned with summarization, and particularly with finding interesting patterns in the data. Econometricians, statisticians, and data mining specialists are generally looking for insights that can be extracted from the data. Machine learning specialists are often primarily concerned with developing high-performance computer systems that can provide useful predictions in the presence of challenging computational constraints. Data science, a somewhat newer term, is concerned with both prediction and summarization, but also with data manipulation, visualization, and other similar tasks. Note that terminology is not standardized in these areas, so these descriptions reflect general usage, not hard-and-fast definitions. Other terms used to describe computer-assisted data analysis include knowledge extraction, information discovery, information harvesting, data archaeology, data pattern processing, and exploratory data analysis.

Much of applied econometrics is concerned with detecting and summarizing relationships in the data. The most common tool used for summarization is (linear) regression analysis. As we shall see, machine learning offers a set of tools that can usefully summarize various sorts of nonlinear relationships in the data. We will focus on these regression-like tools because they are the most natural for economic applications.

In the most general formulation of a statistical prediction problem, we are interested in understanding the conditional distribution of some variable y given some other variables $x = (x_1, \ldots, x_P)$. If we want a point prediction, we can use the mean or median of the conditional distribution.

In machine learning, the x-variables are usually called "predictors" or "features." The focus of machine learning is to find some function that provides a good prediction of y as a function of x. Historically, most work in machine learning has involved cross-section data where it is natural to think of the data being independent and identically distributed (IID) or at least independently distributed. The data may be "fat," which means lots of predictors relative to the number of observations, or "tall" which means lots of observations relative to the number of predictors.

We typically have some observed data on y and x, and we want to compute a "good" prediction of y given new values of x. Usually "good" means it minimizes some loss function such as the sum of squared residuals, mean of absolute value of residuals, and so on. Of course, the relevant loss is that associated with *new* out-of-sample observations of x, not the observations used to fit the model.

When confronted with a prediction problem of this sort an economist would think immediately of a linear or logistic regression. However, there may be better choices, particularly if a lot of data is available. These include nonlinear methods such as 1) classification and regression trees (CART); 2) random forests; and 3) penalized regression such as LASSO, LARS, and elastic nets. (There are also other techniques, such as neural nets, deep learning, and support vector machines, which I do not cover in this review.) Much more detail about these methods can be found in machine learning texts; an excellent treatment is available in Hastie, Tibshirani, and Friedman (2009), which can be freely downloaded. Additional suggestions for further reading are given at the end of this article.

General Considerations for Prediction

Our goal with prediction is typically to get good *out-of-sample predictions*. Most of us know from experience that it is all too easy to construct a predictor that works well in-sample but fails miserably out-of-sample. To take a trivial example, n linearly independent regressors will fit n observations perfectly but will usually have poor out-of-sample performance. Machine learning specialists refer to this phenomenon as the "overfitting problem" and have come up with several ways to deal with it.

First, since simpler models tend to work better for out-of-sample forecasts, machine learning experts have come up with various ways to penalize models for excessive complexity. In the machine learning world, this is known as "regularization," and we will describe some examples below. Economists tend to prefer simpler models for the same reason, but have not been as explicit about quantifying complexity costs.

Second, it is conventional to divide the data into separate sets for the purpose of training, testing, and validation. You use the training data to estimate a model, the validation data to choose your model, and the testing data to evaluate how well your chosen model performs. (Often validation and testing sets are combined.)

Third, if we have an explicit numeric measure of model complexity, we can view it as a parameter that can be "tuned" to produce the best out of sample predictions. The standard way to choose a good value for such a tuning parameter is to use *k-fold cross-validation.*

1. Divide the data into k roughly equal subsets (folds) and label them by $s = 1, \ldots, k$. Start with subset $s = 1$.
2. Pick a value for the tuning parameter.
3. Fit your model using the $k - 1$ subsets other than subset s.
4. Predict for subset s and measure the associated loss.
5. Stop if $s = k$, otherwise increment s by 1 and go to step 2.

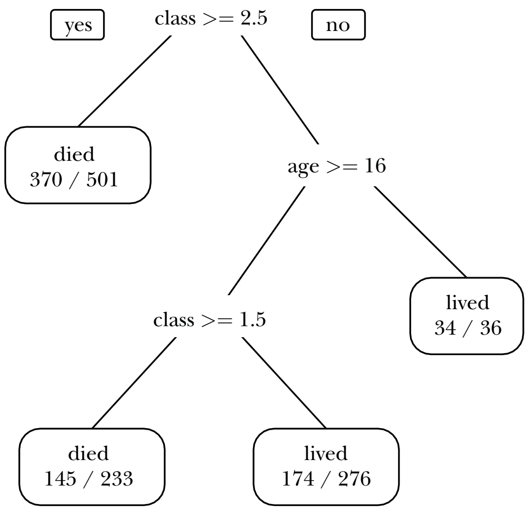
Common choices for k are 10, 5, and the sample size minus 1 ("leave one out"). After cross-validation, you end up with k values of the tuning parameter and the associated loss which you can then examine to choose an appropriate value for the tuning parameter. Even if there is no tuning parameter, it is prudent to use cross-validation to report goodness-of-fit measures since it measures out-of-sample performance, which is generally more meaningful than in-sample performance.

The test-train cycle and cross-validation are very commonly used in machine learning and, in my view, should be used much more in economics, particularly when working with large datasets. For many years, economists have reported in-sample goodness-of-fit measures using the excuse that we had small datasets. But now that larger datasets have become available, there is no reason not to use separate training and testing sets. Cross-validation also turns out to be a very useful technique, particularly when working with reasonably large data. It is also a much more realistic measure of prediction performance than measures commonly used in economics.

Classification and Regression Trees

Let us start by considering a discrete variable regression where our goal is to predict a 0–1 outcome based on some set of features (what economists would call explanatory variables or predictors). In machine learning, this is known as a

Figure 1
A Classification Tree for Survivors of the *Titanic*



Note: See text for interpretation.

classification problem. A common example would be classifying email into "spam" or "not spam" based on characteristics of the email. Economists would typically use a generalized linear model like a logit or probit for a classification problem.

A quite different way to build a classifier is to use a decision tree. Most economists are familiar with decision trees that describe a sequence of decisions that results in some outcome. A tree classifier has the same general form, but the decision at the end of the process is a choice about how to classify the observation. The goal is to construct (or "grow") a decision tree that leads to good out-of-sample predictions.

Ironically, one of the earliest papers on the automatic construction of decision trees (Morgan and Sonquist 1963) was coauthored by an economist. However, the technique did not really gain much traction until 20 years later in the work of Breiman, Friedman, Olshen, and Stone (1984). Nowadays this prediction technique is known as "classification and regression trees," or "CART."

To illustrate the use of tree models, I used the **R** package **rpart** to find a tree that predicts *Titanic* survivors using just two variables: age and class of travel.[1] The resulting tree is shown in Figure 1, and the rules depicted in the tree are shown in Table 2. The rules fit the data reasonably well, misclassifying about 30 percent of the observations in the testing set.

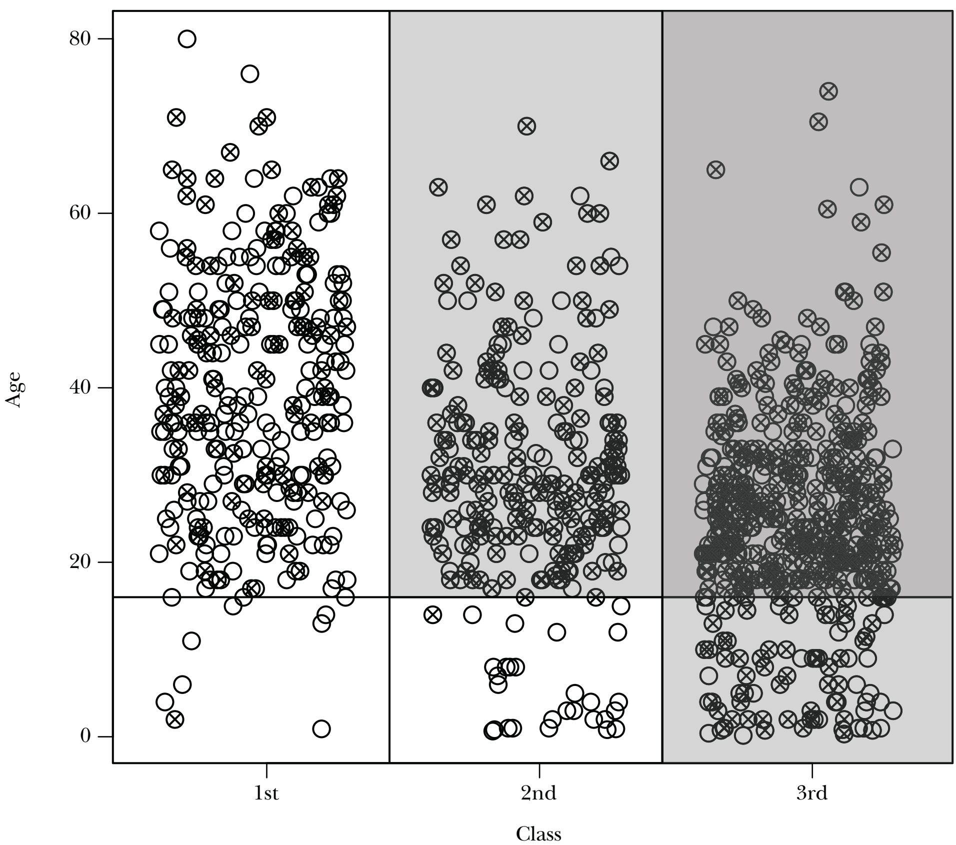
This classification can also be depicted in the "partition plot" (Figure 2), which shows how the tree divides up the space of age and class pairs into rectangular

[1] All data and code used in this paper can be found in the online Appendix available at http://e-jep.org.

Table 2
Tree Model in Rule Form

Features	*Predicted*	*Actual/Total*
Class 3	Died	370/501
Class 1–2, younger than 16	Lived	34/36
Class 2, older than 16	Died	145/233
Class 1, older than 16	Lived	174/276

Figure 2
The Simple Tree Model Predicts Death in Shaded Region
(empty circles indicate survival; circles with x*'s indicate death)*



regions. Of course, the partition plot can only be used for two variables, while a tree representation can handle an arbitrarily large number.

It turns out that there are computationally efficient ways to construct classification trees of this sort. These methods generally are restricted to binary trees (two branches

Table 3
Logistic Regression of Survival versus Age

Coefficient	*Estimate*	*Standard error*	*t value*	*p value*
Intercept	0.465	0.0350	13.291	0.000
Age	−0.002	0.001	−1.796	0.072

Note: Logistic regression relating survival (0 or 1) to age in years.

at each node). They can be used for classification with multiple outcomes ("classification trees") or with continuous dependent variables ("regression trees").

Trees tend to work well for problems where there are important nonlinearities and interactions. As an example, let us continue with the *Titanic* data and create a tree that relates survival to age. In this case, the rule generated by the tree is very simple: predict "survive" if age < 8.5 years. We can examine the same data with a logistic regression to estimate the probability of survival as a function of age, with results reported in Table 3.

The tree model suggests that age is an important predictor of survival, while the logistic model says it is barely important. This discrepancy is explained in Figure 3 where we plot survival rates by age bins. Here we see that survival rates for the youngest passengers were relatively high, and survival rates for older passengers were relatively low. For passengers between these two extremes, age didn't matter very much. So what mattered for survival is not so much age, but whether the passenger was a child or elderly. It would be difficult to discover this pattern from a logistic regression alone.[2]

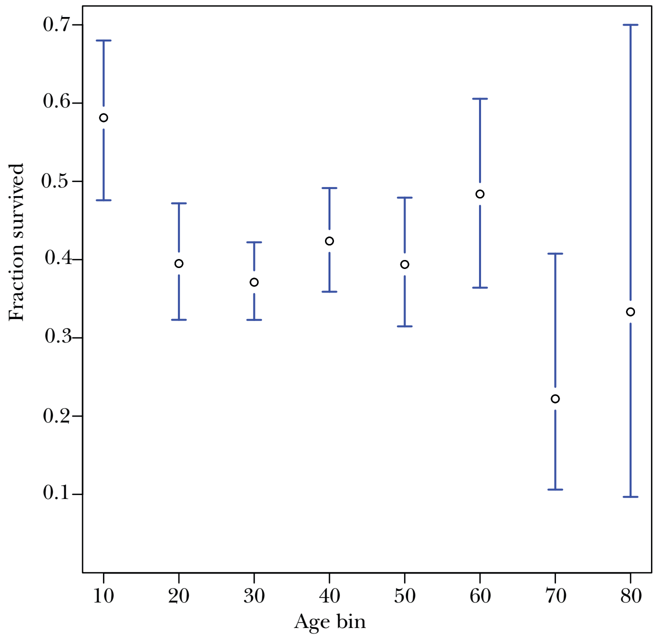
Trees also handle missing data well. Perlich, Provost, and Simonoff (2003) examined several standard datasets and found that "logistic regression is better for smaller data sets and tree induction for larger data sets." Interestingly enough, trees tend *not* to work very well if the underlying relationship really is linear, but there are hybrid models such as RuleFit (Friedman and Popescu 2005) that can incorporate both tree and linear relationships among variables. However, even if trees may not improve on predictive accuracy compared to linear models, the age example shows that they may reveal aspects of the data that are not apparent from a traditional linear modeling approach.

Pruning Trees

One problem with trees is that they tend to overfit the data. Just as a regression with n observations and n variables will give you a good fit in-sample, a tree with many branches will also fit the training data well. In either case, predictions using new data, such as the test set, could be very poor.

[2] It is true that if you *knew* that there was a nonlinearity in age, you could use age dummies in the logit model to capture this effect. However the tree formulation made this nonlinearity immediately apparent.

Figure 3
***Titanic* Survival Rates by Age Group**



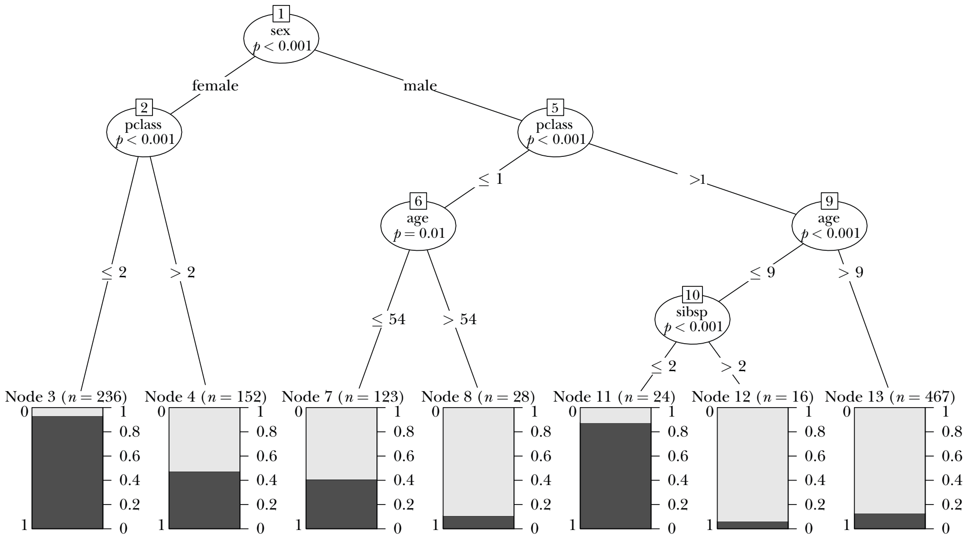
Notes: The figure shows the mean survival rates for different age groups along with confidence intervals. The age bin 10 means "10 and younger," the next age bin is "older than 10 through 20," and so on.

The most common solution to this problem is to "prune" the tree by imposing a cost for complexity. There are various measures of complexity, but a common one is the number of terminal nodes (also known as "leafs"). The cost of complexity is a tuning parameter that is chosen to provide the best out-of-sample predictions, which is typically measured using the 10-fold cross-validation procedure mentioned earlier.

A typical tree estimation session might involve dividing your data into ten folds, using nine of the folds to grow a tree with a particular complexity, and then predict on the excluded fold. Repeat the estimation with different values of the complexity parameter using other folds and choose the value of the complexity parameter that minimizes the out-of-sample classification error. (Some researchers recommend being a bit more aggressive and advocate choosing the complexity parameter that is one standard deviation lower than the loss-minimizing value.)

Of course, in practice, the computer program handles most of these details for you. In the examples in this paper, I mostly use default choices to keep things simple, but in practice these defaults will often be adjusted by the analyst. As with any other statistical procedure, skill, experience, and intuition are helpful in coming up with a good answer. Diagnostics, exploration, and experimentation are just as useful with these methods as with regression techniques.

Figure 4
A ctree for Survivors of the *Titanic*
(black bars indicate fraction of the group that survived)



Note: See text for interpretation.

There are many other approaches to creating trees, including some that are explicitly statistical in nature. For example, a "conditional inference tree," or ctree for short, chooses the structure of the tree using a sequence of hypothesis tests. The resulting trees tend to need very little pruning (Hothorn, Hornik, and Zeileis 2006). An example for the *Titanic* data is shown in Figure 4.

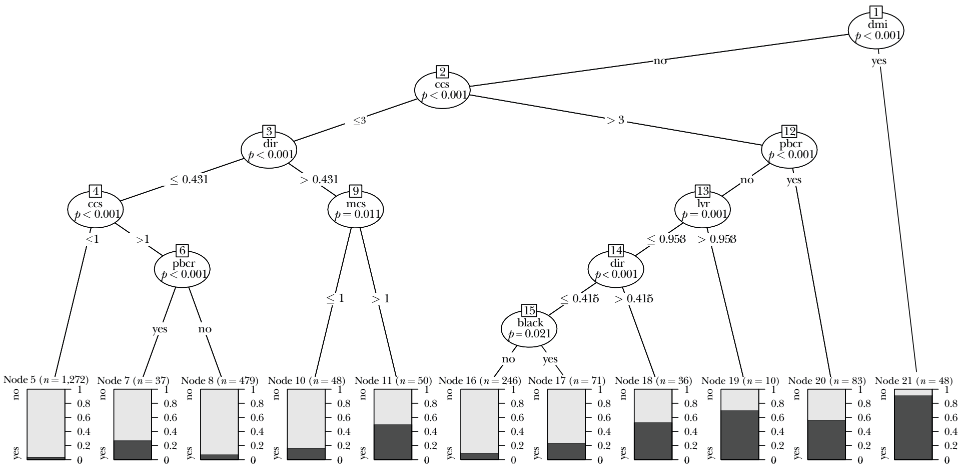
The first node divides by gender. The second node then divides by class. In the right-hand branches, the third node divides by age, and a fourth node divides by the number of siblings plus spouse aboard. The bins at the bottom of the figure show the total number of people in that leaf and a graphical depiction of their survival rate. One might summarize this tree by the following principle: "women and children first . . . particularly if they were traveling first class." This simple example again illustrates that classification trees can be helpful in summarizing relationships in data, as well as predicting outcomes.[3]

An Economic Example Using Home Mortgage Disclosure Act Data

Munnell, Tootell, Browne, and McEneaney (1996) examined mortgage lending in Boston to see if race played a significant role in determining who was approved for a mortgage. The primary econometric technique was a logistic regression where

[3] For two excellent tutorials on tree methods that use the *Titanic* data, see Stephens and Wehrley (2014).

Figure 5
Home Mortgage Disclosure Act (HMDA) Data Tree



Notes: Figure 5 shows a conditional tree estimated using the **R** package **party**. The black bars indicate the fraction of each group who were denied mortgages. The most important determinant of this is the variable "dmi," or "denied mortgage insurance." Other variables are: "dir," debt payments to total income ratio; "hir," housing expenses to income ratio; "lvr," ratio of size of loan to assessed value of property; "ccs," consumer credit score; "mcs," mortgage credit score; "pbcr," public bad credit record; "dmi," denied mortgage insurance; "self," self-employed; "single," applicant is single; "uria," 1989 Massachusetts unemployment rate applicant's industry; "condominium," unit is condominium; "black," race of applicant black; and "deny," mortgage application denied.

race was included as one of the predictors. The coefficient on race showed a statistically significant negative impact on probability of getting a mortgage for black applicants. This finding prompted considerable subsequent debate and discussion; see Ladd (1998) for an overview.

Here I examine this question using the tree-based estimators described in the previous section. The data consists of 2,380 observations of 12 predictors, one of which was race. Figure 5 shows a conditional tree estimated using the **R** package **party**.

The tree fits pretty well, misclassifying 228 of the 2,380 observations for an error rate of 9.6 percent. By comparison, a simple logistic regression does slightly better, misclassifying 225 of the 2,380 observations, leading to an error rate of 9.5 percent. As you can see in Figure 5, the most important variable is "dmi" = "denied mortgage insurance." This variable alone explains much of the variation in the data. The race variable ("black") shows up far down the tree and seems to be relatively unimportant.

One way to gauge whether a variable is important is to exclude it from the prediction and see what happens. When this is done, it turns out that the accuracy of the tree-based model doesn't change at all: exactly the same cases are misclassified. Of course, it is perfectly possible that there was racial discrimination elsewhere in the mortgage process, or that some of the variables included are highly correlated with race. But it is noteworthy that the tree model produced by standard procedures that omits race fits the observed data just as well as a model that includes race.

Boosting, Bagging, Bootstrap

There are several useful ways to improve classifier performance. Interestingly enough, some of these methods work by *adding* randomness to the data. This seems paradoxical at first, but adding randomness turns out to be a helpful way of dealing with the overfitting problem.

Bootstrap involves choosing (with replacement) a sample of size n from a dataset of size n to estimate the sampling distribution of some statistic. A variation is the "m out of n bootstrap" which draws a sample of size m from a dataset of size $n > m$.

Bagging involves averaging across models estimated with several different bootstrap samples in order to improve the performance of an estimator.

Boosting involves repeated estimation where misclassified observations are given increasing weight in each repetition. The final estimate is then a vote or an average across the repeated estimates.[4]

Econometricians are well-acquainted with the bootstrap but rarely use the other two methods. Bagging is primarily useful for nonlinear models such as trees (Friedman and Hall 2007). Boosting tends to improve predictive performance of an estimator significantly and can be used for pretty much any kind of classifier or regression model, including logits, probits, trees, and so on.

It is also possible to combine these techniques and create a "forest" of trees that can often significantly improve on single-tree methods. Here is a rough description of how such "random forests" work.

Random Forests

Random forests is a technique that uses multiple trees. A typical procedure uses the following steps.

1. Choose a bootstrap sample of the observations and start to grow a tree.
2. At each node of the tree, choose a random sample of the predictors to make the next decision. Do not prune the trees.
3. Repeat this process many times to grow a forest of trees.
4. In order to determine the classification of a new observation, have each tree make a classification and use a majority vote for the final prediction.

This method produces surprisingly good out-of-sample fits, particularly with highly nonlinear data. In fact, Howard and Bowles (2012) claim "ensembles of decision trees (often known as 'Random Forests') have been the most successful general-purpose algorithm in modern times." They go on to indicate that

[4] Boosting is often used with decision trees, where it can dramatically improve their predictive performance.

"the algorithm is very simple to understand, and is fast and easy to apply." See also Caruana and Niculescu-Mitzil (2006) who compare several different machine learning algorithms and find that ensembles of trees perform quite well. There are a number of variations and extensions of the basic "ensemble of trees" model such as Friedman's "Stochastic Gradient Boosting" (Friedman 2002).

One defect of random forests is that they are a bit of a black box—they don't offer simple summaries of relationships in the data. As we have seen earlier, a single tree can offer some insight about how predictors interact. But a forest of a thousand trees cannot be easily interpreted. However, random forests can determine which variables are "important" in predictions in the sense of contributing the biggest improvements in prediction accuracy.

Note that random forests involves quite a bit of randomization; if you want to try them out on some data, I strongly suggest choosing a particular seed for the random number generator so that your results can be reproduced. (See the online supplement for examples.)

I ran the random forest method on the HMDA data and found that it misclassified 223 of the 2,380 cases, a small improvement over the logit and the ctree. I also used the importance option in random forests to see how the predictors compared. It turned out that "dmi" was the most important predictor and race was second from the bottom, which is consistent with the ctree analysis.

Variable Selection

Let us return to the familiar world of linear regression and consider the problem of variable selection. There are many such methods available, including stepwise regression, principal component regression, partial least squares, Akaike information criterion (AIC) and Bayesian information criterion (BIC) complexity measures, and so on. Castle, Qin, and Reed (2009) describe and compare 21 different methods.

LASSO and Friends

Here we consider a class of estimators that involves penalized regression. Consider a standard multivariate regression model where we predict y_t as a linear function of a constant, b_0, and P predictor variables. We suppose that we have standardized all the (nonconstant) predictors so they have mean zero and variance one.

Consider choosing the coefficients $(b_1, \ldots, b_P)$ for these predictor variables by minimizing the sum of squared residuals plus a penalty term of the form

$$\lambda \sum_{p=1}^{P} \left[(1 - \alpha) |b_p| + \alpha |b_p|^2 \right].$$

This estimation method is called *elastic net regression*; it contains three other methods as special cases. If there is no penalty term $(\lambda = 0)$, this is *ordinary least squares*. If $\alpha = 1$, so that there is only the quadratic constraint, this is *ridge regression*.

If $\alpha = 0$, this is called the *LASSO*, an acronym for "least absolute shrinkage and selection operator."

These penalized regressions are classic examples of regularization. In this case, the complexity is the number and size of predictors in the model. All of these methods tend to shrink the least squares regression coefficients towards zero. The LASSO and elastic net typically produces regressions where some of the variables are set to be exactly zero. Hence this is a relatively straightforward way to do variable selection.

It turns out that these estimators can be computed quite efficiently, so doing variable selection on reasonably large problems is computationally feasible. They also seem to provide good predictions in practice.

Spike-and-Slab Regression

Another approach to variable selection that is novel to most economists is spike-and-slab regression, a Bayesian technique. Suppose that you have P possible predictors in some linear model. Let γ be a vector of length P composed of zeros and ones that indicate whether or not a particular variable is included in the regression.

We start with a Bernoulli prior distribution on γ; for example, initially we might think that all variables have an equally likely chance of being in the regression. Conditional on a variable being in the regression, we specify a prior distribution for the regression coefficient associated with that variable. For example, we might use a Normal prior with mean 0 and a large variance. These two priors are the source of the method's name: the "spike'" is the probability of a coefficient being nonzero; the "slab" is the (diffuse) prior describing the values that the coefficient can take on.

Now we take a draw of γ from its prior distribution, which will just be a list of variables in the regression. Conditional on this list of included variables, we take a draw from the prior distribution for the coefficients. We combine these two draws with the likelihood in the usual way, which gives us a draw from posterior distribution on both probability of inclusion and the coefficients. We repeat this process thousands of times using a Markov Chain Monte Carlo (MCMC) technique which gives us a table summarizing the posterior distribution for γ (indicating variable inclusion), β (the coefficients), and the associated prediction of y. We can summarize this table in a variety of ways. For example, we can compute the average value of γ_p which shows the posterior probability that the variable p is included in the regressions.

An Economic Example: Growth Regressions

We illustrate these different methods of variable selection using data from Sala-i-Martín (1997). This exercise involved examining a dataset of 72 counties and 42 variables in order to see which variables appeared to be important predictors of economic growth. Sala-i-Martín (1997) computed at all possible subsets of regressors of manageable size and used the results to construct an importance measure he called CDF(0). Ley and Steel (2009) investigated the same question using Bayesian

Table 4
Comparing Variable Selection Algorithms: Which Variables Appeared as Important Predictors of Economic Growth?

Predictor	*Bayesian model averaging*	*CDF(0)*	*LASSO*	*Spike-and-Slab*
GDP level 1960	1.000	1.000	-	0.9992
Fraction Confucian	0.995	1.000	2	0.9730
Life expectancy	0.946	0.942	-	0.9610
Equipment investment	0.757	0.997	1	0.9532
Sub-Saharan dummy	0.656	1.000	7	0.5834
Fraction Muslim	0.656	1.000	8	0.6590
Rule of law	0.516	1.000	-	0.4532
Open economy	0.502	1.000	6	0.5736
Degree of capitalism	0.471	0.987	9	0.4230
Fraction Protestant	0.461	0.966	5	0.3798

Source: The table is based on that in Ley and Steel (2009); the data analyzed is from Sala-i-Martín (1997).
Notes: We illustrate different methods of variable selection. This exercise involved examining a dataset of 72 counties and 42 variables in order to see which variables appeared to be important predictors of economic growth. The table shows ten predictors that were chosen by Sala-i-Martín (1997) using a CDF(0) measure defined in the 1997 paper; Ley and Steel (2009) using Bayesian model averaging, LASSO, and spike-and-slab regressions. Metrics used are not strictly comparable across the various models. The "Bayesian model averaging" and "Spike-and-Slab" columns are posterior probabilities of inclusion; the "LASSO" column just shows the ordinal importance of the variable or a dash indicating that it was not included in the chosen model; and the CDF(0) measure is defined in Sala-i-Martín (1997).

model averaging, a technique related to, but not identical with, spike-and-slab. Hendry and Krolzig (2004) examined an iterative significance test selection method.

Table 4 shows ten predictors that were chosen by Sala-i-Martín (1997) using his two million regressions, Ley and Steel (2009) using Bayesian model averaging, LASSO, and spike-and-slab. The table is based on that in Ley and Steel (2009) but metrics used are not strictly comparable across the various models. The "Bayesian model averaging" and "spike-slab" columns show posterior probabilities of inclusion; the "LASSO" column just shows the ordinal importance of the variable or a dash indicating that it was not included in the chosen model; and the CDF(0) measure is defined in Sala-i-Martín (1997).

The LASSO and the Bayesian techniques are very computationally efficient and would likely be preferred to exhaustive search. All four of these variable selection methods give similar results for the first four or five variables, after which they diverge. In this particular case, the dataset appears to be too small to resolve the question of what is "important" for economic growth.

Variable Selection in Time Series Applications

The machine learning techniques described up until now are generally applied to cross-sectional data where independently distributed data is a plausible assumption. However, there are also techniques that work with time series. Here we describe an estimation method that we call Bayesian Structural Time Series (BSTS) that seems to work well for variable selection problems in time series applications.

Our research in this area was motivated by Google Trends data, which provides an index of the volume of Google queries on specific terms. One might expect that queries on "file for unemployment" might be predictive of the actual rate of filings for initial claims, or that queries on "Orlando vacation" might be predictive of actual visits to Orlando. Indeed, in Choi and Varian (2009, 2012), Goel, Hofman, Lahaie, Pennock, and Watts (2010), Carrière-Swallow and Labbé (2011), McLaren and Shanbhoge (2011), Artola and Galan (2012), Hellerstein and Middeldorp (2012), and other papers, many researchers have shown that Google queries do have significant short-term predictive power for various economic metrics.

The challenge is that there are billions of queries so it is hard to determine exactly which queries are the most predictive for a particular purpose. Google Trends classifies the queries into categories, which helps a little, but even then we have hundreds of categories as possible predictors so that overfitting and spurious correlation are a serious concern. Bayesian Structural Time Series is designed to address these issues. We offer a very brief description here; more details are available in Scott and Varian (2013a, 2013b).

Consider a classic time series model with *constant* level, linear time trend, and regressor components:

$$yt = \mu + bt + \beta xt + et.$$

The "local linear trend" is a stochastic generalization of this model where the level and time trend can vary through time.

Observation: $y_t = \mu_t + z_t + e_{1t} =$ level $+$ regression

State variable 1: $\mu_t = \mu_{t-1} + b_{t-1} + e_{2t} =$ random walk $+$ trend

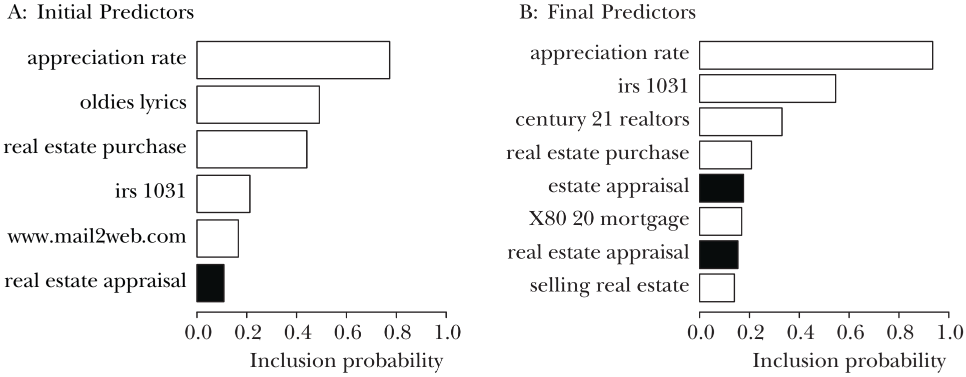
State variable 2: $z_t = \beta x_t =$ regression

State variable 3: $b_t = b_{t-1} + e_{3t} =$ random walk for trend

It is easy to add an additional state variable for seasonality if that is appropriate. The parameters to estimate are the regression coefficients β and the variances of (e_{it}) for $i = 1, \ldots 3$. We can then use these estimates to construct the optimal forecast based on techniques drawn from the literature on Kalman filters.

For the regression, we use the spike-and-slab variable choice mechanism described above. A draw from the posterior distribution now involves a draw of variances of (e_{1t}, e_{2t}, e_{3t}) a draw of the vector γ that indicates which variables are in the regression, and a draw of the regression coefficients β for the included variables. The draws of μ_t, b_t, and β can be used to construct estimates of y_t and forecasts for y_{t+1}. We end up with an (estimated) posterior distribution for each parameter of

Figure 6
An Example Using Bayesian Structural Time Series (BSTS)
(finding Google queries that are predictors of new home sales)



Source: Author using HSN1FNSA data from the St. Louis Federal Reserve Economic Data.
Notes: Consider the nonseasonally adjusted data for new homes sold in the United States, which is (HSN1FNSA) from the St. Louis Federal Reserve Economic Data. This time series can be submitted to Google Correlate, which then returns the 100 queries that are the most highly correlated with the series. We feed that data into the BSTS system, which identifies the predictors with the largest posterior probabilities of appearing in the housing regression; these are shown in Figure 6A. In these figures, black bars indicate a negative relationship, and white bars indicate a positive relationship. Two predictors, “oldies lyrics” and “www.mail2web” appear to be spurious so we remove them and re-estimate, yielding the results in Figure 6B.

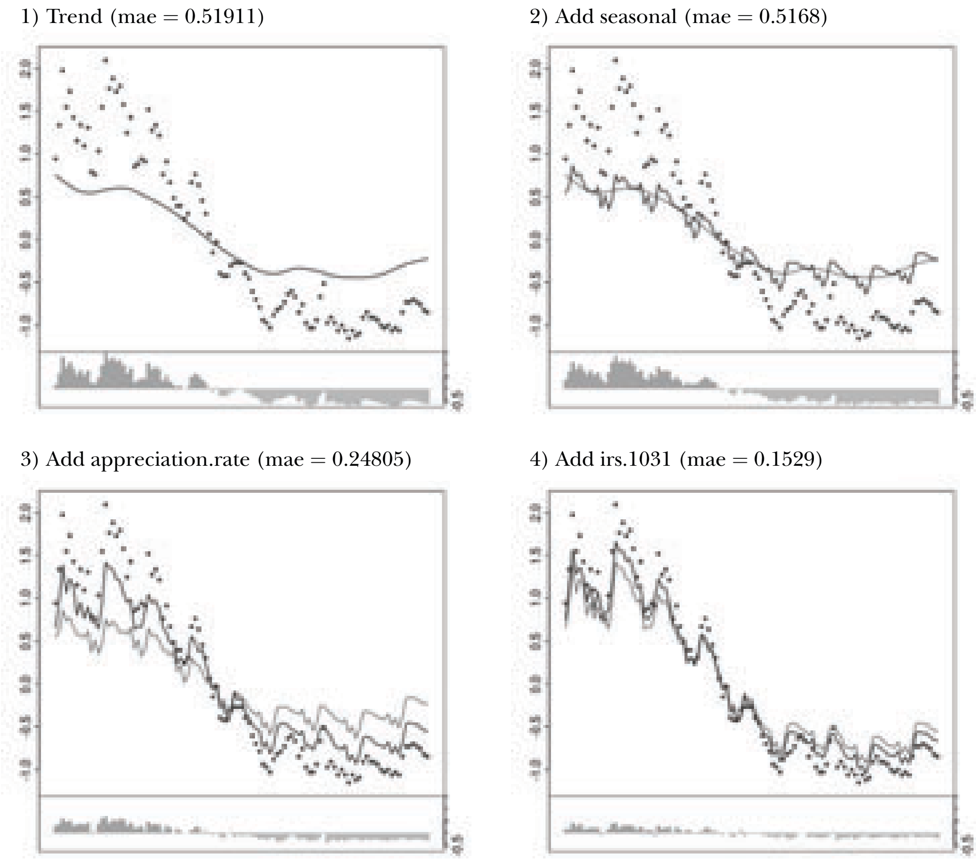
interest. If we seek a point prediction, we can average over these draws, which is essentially a form of Bayesian model averaging.

As an example, consider the nonseasonally adjusted data for new homes sold in the United States, which is (HSN1FNSA) from the St. Louis Federal Reserve Economic Data. This time series can be submitted to Google Correlate, which then returns the 100 queries that are the most highly correlated with the series. We feed that data into the BSTS system, which identifies the predictors with the largest posterior probabilities of appearing in the housing regression; these are shown in Figure 6A.In these figures, black bars indicate a negative relationship and white bars indicate a positive relationship. Two predictors, “oldies lyrics” and “www.mail2web” appear to be spurious so we remove them and re-estimate, yielding the results in Figure 6B.

The fit is shown in Figure 7, which shows the incremental contribution of the trend, seasonal, and two of the regressors. Even with only two predictors, queries on “appreciation rate” and queries on “irs 1031,” we get a pretty good fit.[5]

[5] IRS section 1031 has to do with deferring capital gains on certain sorts of property exchange.

Figure 7
Fit for the Housing Regression: Incremental Contribution of Trend, Seasonal, and Two Regressors



Source: Author using (HSN1FNSA) data from the St. Louis Federal Reserve.
Notes: The plots show the impact of the trend, seasonal, and a few individual regressors. Data has been standardized to have mean zero and variance 1. The residuals are shown on the bottom. The abbreviation "mae" stands for "mean absolute error."

Econometrics and Machine Learning

There are a number of areas where there would be opportunities for fruitful collaboration between econometrics and machine learning. I mentioned above that most machine learning uses independent and identically distributed data. However, the Bayesian Structural Time Series model shows that some of these techniques can be adopted for time series models. It is also possible to use machine learning techniques to look at panel data, and there has been some work in this direction.

However, the most important area for collaboration involves causal inference. Econometricians have developed several tools for causal inference such as instrumental variables, regression discontinuity, difference-in-differences, and various forms of natural and designed experiments (Angrist and Krueger 2001). Machine learning work has, for the most part, dealt with pure prediction. In a way, this is ironic, since theoretical computer scientists, such as Pearl (2009a, b) have made significant contributions to causal modeling. However, it appears that these theoretical advances have not as yet been incorporated into machine learning practice to a significant degree.

Causality and Prediction

As economists know well, there is a big difference between correlation and causation. A classic example: there are often more police in precincts with high crime, but that does not imply that increasing the number of police in a precinct would increase crime.

The machine learning models we have described so far have been entirely about prediction. If our data were generated by policymakers who assigned police to areas with high crime, then the observed relationship between police and crime rates could be highly predictive for the *historical* data but not useful in predicting the causal impact of explicitly *assigning* additional police to a precinct.

To enlarge on this point, let us consider an experiment (natural or designed) that attempts to estimate the impact of some policy, such as adding police to precincts. There are two critical questions.

1) How will police be assigned to precincts in both the experiment and the policy implementation? Possible assignment rules could be 1) random, 2) based on perceived need, 3) based on cost of providing service, 4) based on resident requests, 5) based on a formula or set of rules, 6) based on asking for volunteers, and so on. Ideally the assignment procedure in the experiment will be similar to that used in the policy. Developing accurate predictions about which precincts will receive additional police under the proposed policy based on the experimental data can clearly be helpful in predicting the expected impact of the policy.

2) What will be the impact of these additional police in both the experiment and the policy? As Rubin (1974) and many subsequent authors have emphasized, when we want to estimate the *causal* impact of some treatment we need to compare the outcome with the intervention to what *would have happened* without the intervention. But this counterfactual cannot be observed, so it must be predicted by some model. The better predictive model you have for the counterfactual, the better you will be able to estimate the causal effect, a rule that is true for both pure experiments and natural experiments.

So even though a predictive model will not necessarily allow one to conclude anything about causality by itself, such models may help in estimating the causal impact of an intervention when it occurs.

To state this in a slightly more formal way, consider the identity from Angrist and Pischke (2009, p. 11):

$$\text{observed difference in outcome} = \text{average treatment effect on the treated} + \text{selection bias.}$$

If you want to model the average treatment effect as a function of other variables, you will usually need to model both the observed difference in outcome and the selection bias. The better your predictive model for those components, the better your estimate of the average treatment effect will be. Of course, if you have a true randomized treatment–control experiment, selection bias goes away and those treated are an unbiased random sample of the population.

To illustrate these points, let us consider the thorny problem of estimating the causal effect of advertising on sales (Lewis and Rao 2013). The difficulty is that there are many confounding variables, such as seasonality or weather, that cause both increased ad exposures and increased purchases by consumers. For example, consider the (probably apocryphal) story about an advertising manager who was asked why he thought his ads were effective. "Look at this chart," he said. "Every December I increase my ad spend and, sure enough, purchases go up." Of course, in this case, seasonality can be included in the model. However, generally there will be other confounding variables that affect both exposure to ads and the propensity of purchase, which make causal interpretations of observed relationships problematic.

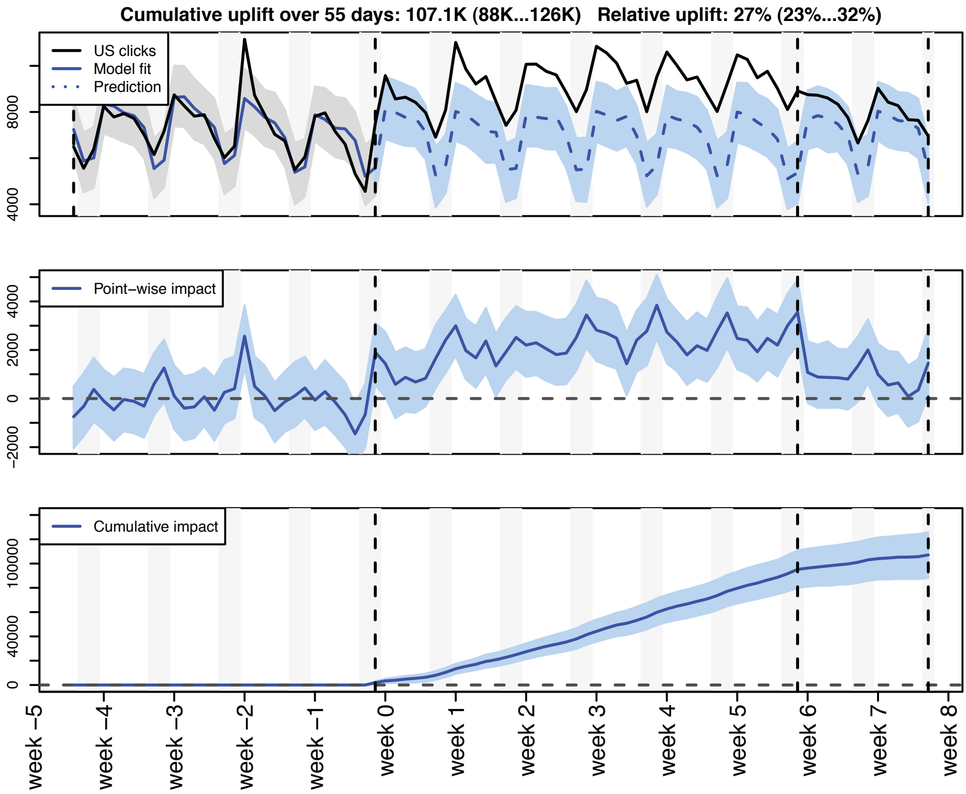
The ideal way to estimate advertising effectiveness is, of course, to run a controlled experiment. In this case the control group provides an estimate of the counterfactual: what would have happened without ad exposures. But this ideal approach can be quite expensive, so it is worth looking for alternative ways to predict the counterfactual. One way to do this is to use the Bayesian Structural Time Series (BSTS) method described earlier.

Suppose a given company wants to determine the impact of an advertising campaign on visits to its website. It first uses BSTS (or some other technique) to build a model predicting the time series of visits as a function of its past history, seasonal effects, and other possible predictors such as Google queries on its company name, its competitors' names, or products that it produces. Since there are many possible choices for predictors, it is important to use some variable selection mechanism such as those described earlier.

It next runs an ad campaign for a few weeks and records visits during this period. Finally, it makes a forecast of what visits *would have been* in the absence of the ad campaign using the model developed in the first stage. Comparing the actual visits to the counterfactual visits gives us an estimate of the causal effect of advertising.

Figure 8, shows the outcome of such a procedure. It is based on the approach proposed in Brodersen, Gallusser, Koehler, Remy, and Scott (2013), but the covariates are chosen automatically from Google Trends categories using Bayesian Structural Time Series (BSTS). Panel A shows the actual visits and the prediction

Figure 8
Actual and Predicted Website Visits



Source: This example is based on the approach proposed in Brodersen, Gallusser, Koehler, Remy, and Scott (2013), but the covariates are chosen automatically from Google Trends categories using Bayesian Structural Time Series (BSTS).
Notes: Suppose a given company wants to determine the impact of an advertising campaign on its website visits. Panel A shows the actual visits and the prediction of what the visits would have been without the campaign based on the BSTS forecasting model. Panel B shows the difference between actual and predicted visits, and Panel C shows the cumulative difference.

of what the visits would have been without the campaign based on the BSTS forecasting model. Panel B shows the difference between actual and predicted visits, and Panel C shows the cumulative difference. It is clear from this figure that there was a significant causal impact of advertising, which can then be compared to the cost of the advertising to evaluate the campaign.

This procedure does not use a control group in the conventional sense. Rather it uses a general time series model based on trend extrapolation, seasonal effects, and relevant covariates to forecast what would have happened without the ad campaign.

A good predictive model can be better than a randomly chosen control group, which is usually thought to be the gold standard. To see this, suppose that you run an ad campaign in 100 cities and retain 100 cities as a control. After the experiment is over, you discover the weather was dramatically different across the cities in the study. Should you add weather as a predictor of the counterfactual? Of course! If weather affects sales (which it does), then you will get a more accurate prediction of the counterfactual and thus a better estimate of the causal effect of advertising.

Model Uncertainty

An important insight from machine learning is that averaging over many small models tends to give better out-of-sample prediction than choosing a single model.

In 2006, Netflix offered a million dollar prize to researchers who could provide the largest improvement to their existing movie recommendation system. The winning submission involved a "complex blending of no fewer than 800 models," though they also point out that "predictions of good quality can usually be obtained by combining a small number of judiciously chosen methods" (Feuerverger, He, and Khatri 2012). It also turned out that a blend of the best- and second-best submissions outperformed either of them.

Ironically, it was recognized many years ago that averages of macroeconomic model forecasts outperformed individual models, but somehow this idea was rarely exploited in traditional econometrics. The exception is the literature on Bayesian model averaging, which has seen a steady flow of work; see Steel (2011) for a survey.

However, I think that model uncertainty has crept into applied econometrics through the back door. Many papers in applied econometrics present regression results in a table with several different specifications: which variables are included in the controls, which variables are used as instruments, and so on. The goal is usually to show that the estimate of some interesting parameter is not very sensitive to the exact specification used.

One way to think about it is that these tables illustrate a simple form of model uncertainty: how an estimated parameter varies as different models are used. In these papers, the authors tend to examine only a few representative specifications, but there is no reason why they couldn't examine many more if the data were available.

In this period of "big data," it seems strange to focus on *sampling uncertainty,* which tends to be small with large datasets, while completely ignoring *model uncertainty,* which may be quite large. One way to address this is to be explicit about examining how parameter estimates vary with respect to choices of control variables and instruments.

Summary and Further Reading

Since computers are now involved in many economic transactions, big data will only get bigger. Data manipulation tools and techniques developed for small datasets will become increasingly inadequate to deal with new problems. Researchers in machine learning have developed ways to deal with large datasets and economists interested in dealing with such data would be well advised to invest in learning these techniques.

I have already mentioned Hastie, Tibshirani, and Friedman (2009), who provide detailed descriptions of all the methods discussed here but at a relatively advanced level. James, Witten, Hastie, and Tibshirani (2013) describe many of the same topics at an undergraduate-level, along with **R** code and many examples. (There are several economic examples in the book where the tension between predictive modeling and causal inference is apparent.) Murphy (2012) examines machine learning from a Bayesian point of view.

Venables and Ripley (2002) offer good discussions of these topics with emphasis on applied examples. Leek (2013) presents a number of YouTube videos with gentle and accessible introductions to several tools of data analysis. Howe (2013) provides a somewhat more advanced introduction to data science that also includes discussions of SQL and NoSQL databases. Wu and Kumar (2009) give detailed descriptions and examples of the major algorithms in data mining, while Williams (2011) provides a unified toolkit. Domingos (2012) summarizes some important lessons including "pitfalls to avoid, important issues to focus on and answers to common questions."

■ *Thanks to Jeffrey Oldham, Tom Zhang, Rob On, Pierre Grinspan, Jerry Friedman, Art Owen, Steve Scott, Bo Cowgill, Brock Noland, Daniel Stonehill, Robert Snedegar, Gary King, Fabien Curto-Millet, and the editors of this journal for helpful comments on earlier versions of this paper. The author works for Google, and Google had the right to review this paper before publication.*

References

Angrist, Joshua D., and Alan B. Krueger. 2001. "Instrumental Variables and the Search for Identification: From Supply and Demand to Natural Experiments." *Journal of Economic Perspectives* 5(4): 69–85.

Angrist, Joshua D., and Jörn-Steffen Pischke. 2009. *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton University Press.

Artola, Concha, and Enrique Galan. 2012. "Tracking the Future on the Web: Construction of Leading Indicators Using Internet Searches." Documentos Ocasionales 1203T, Bank of Spain. http://www.bde.es/webbde/SES/Secciones/Publicaciones/PublicacionesSeriadas/DocumentosOcasionales/12/Fich/do1203e.pdf.

Breiman, Leo, Jerome H. Friedman, R. A. Olshen, and Charles J. Stone. 1984. *Classification and Regression Trees*. Wadsworth and Brooks/Cole, Monterey.

Brodersen, Kay H., Fabian Gallusser, Jim Koehler, Nicolas Remy, and Steven L. Scott. 2013. "Inferring Causal Impact Using Bayesian Structural Time-Series Models." http://research.google.com/pubs/pub41854.html.

Carrière-Swallow, Yan, and Felipe Labbé. 2011. "Nowcasting with Google Trends in an Emerging Market." *Journal of Forecasting* 32(4): 289–98.

Caruana, Rich, and Alexandru Niculescu-Mizil. 2006. "An Empirical Comparison of Supervised Learning Algorithms." In *Proceedings of the 23rd*

International Conference on Machine Learning, Pittsburgh, PA. Available at: http://www.autonlab.org/icml2006/technical/accepted.html.

Castle, Jennifer L., Xiaochuan Qin, and W. Robert Reed. 2009. "How to Pick the Best Regression Equation: A Review and Comparison of Model Selection Algorithms." Working Paper 13/2009, Department of Economics, University of Canterbury. http://www.econ.canterbury.ac.nz/RePEc/cbt/econwp/0913.pdf.

Choi, Hyunyoung, and Hal Varian. 2009. "Predicting the Present with Google Trends." http://google.com/googleblogs/pdfs/google_predicting_the_present.pdf.

Choi, Hyunyoung, and Hal Varian. 2012. "Predicting the Present with Google Trends." *Economic Record* 88(1): 2–9.

Domingos, Pedro. 2012. "A Few Useful Things to Know about Machine Learning." *Communications of the ACM* 55(10): 78–87.

Einav, Liran, and Jonathan D. Levin. 2013. "The Data Revolution and Economic Analysis." Technical report, NBER Innovation Policy and the Economy Conference, 2013. NBER Working Paper 19035.

Feuerverger, Andrey, Yu He, and Shashi Khatri. 2012. "Statistical Significance of the Netflix Challenge." *Statistical Science* 27(2): 202–231.

Friedman, Jerome. 2002. "Stochastic Gradient Boosting." *Computational Statistics & Data Analysis* 38(4): 367–78.

Friedman, Jerome, and Peter Hall. 2007. "On Bagging and Nonlinear Estimation." *Journal of Statistical Planning and Inference* 137(3): 669–83.

Friedman, Jerome H., and Bogdan E. Popescu. 2005. "Predictive Learning via Rule Ensembles." Technical report, Stanford University. http://www-stat.stanford.edu/~jhf/ftp/RuleFit.pdf

Goel, Sharad, Jake M. Hofman, Sébastien Lahaie, David M. Pennock, and Duncan J. Watts. 2010. "Predicting Consumer Behavior with Web Search." *PNAS* 107(41).

Hastie, Trevor, Robert Tibshirani, and Jerome Friedman. 2009. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd edition. Springer-Verlag.

Hellerstein, Rebecca, and Menno Middeldorp. 2012. "Forecasting with Internet Search Data." *Liberty Street Economics* Blog of the Federal Reserve Bank of New York, January 4. http://libertystreeteconomics.newyorkfed.org/2012/01/forecasting-with-internet-search-data.html.

Hendry, David F., and Hans-Martin Krolzig. 2004. "We Ran One Regression." *Oxford Bulletin of Economics and Statistics* 66(5): 799–810.

Hothorn, Torsten, Kurt Hornik, and Achim Zeileis. 2006. "Unbiased Recursive Partitioning: A Conditional Inference Framework." *Journal of Computational and Graphical Statistics* 15(3): 651–74.

Howard, Jeremy, and Mike Bowles. 2012. "The Two Most Important Algorithms in Predictive Modeling Today." Strata Conference presentation, February 28. http://strataconf.com/strata2012/public/schedule/detail/22658.

Howe, Bill. 2013. Introduction to Data Science. A course from the University of Washington. https://class.coursera.org/datasci-001/lecture/index.

James, Gareth, Daniela Witten, Trevor Hastie, and Robert Tibshirani. 2013. *An Introduction to Statistical Learning: With Applications in R.* New York: Springer.

Ladd, Helen F. 1998. "Evidence on Discrimination in Mortgage Lending." *Journal of Economic Perspectives* 12(2): 41–62.

Leek, Jeff. 2013. Data Analysis. Videos from the course. http://blog.revolutionanalytics.com/2013/04/coursera-data-analysis-course-videos.html.

Lewis, Randall A., and Justin M. Rao. 2013. "On the Near Impossibility of Measuring the Returns to Advertising." Unpublished paper, Google, Inc. and Microsoft Research. http://justinmrao.com/lewis_rao_nearimpossibility.pdf.

Ley, Eduardo, and Mark F. J. Steel. 2009. "On the Effect of Prior Assumptions in Bayesian Model Averaging with Applications to Growth Regression." *Journal of Applied Econometrics* 24(4): 651–74.

McLaren, Nick, and Rachana Shanbhoge. 2011. "Using Internet Search Data as Economic Indicators." *Bank of England Quarterly Bulletin* 51(2): 134–40.

Morgan, James N., and John A. Sonquist. 1963. "Problems in the Analysis of Survey Data, and a Proposal." *Journal of the American Statistical Association* 58(302): 415–34.

Munnell, Alicia H., Geoffrey M. B. Tootell, Lynne E. Browne, and James McEneaney. 1996. "Mortgage Lending in Boston: Interpreting HMDA Data." *American Economic Review* 86(1): 25–53.

Murphy, Kevin P. 2012. *Machine Learning: A Probabilistic Perspective.* MIT Press.

Pearl, Judea. 2009a. *Causality: Models, Reasoning, and Inference*, 2nd edition. Cambridge University Press.

Pearl, Judea. 2009b. "Causal Inference in Statistics: An Overview." *Statistics Surveys* 3: 96–146.

Perlich, Claudia, Foster Provost, and Jeffrey S. Simonoff. 2003. "Tree Induction vs. Logistic Regression: A Learning-Curve Analysis." *Journal of Machine Learning Research* 4: 211–55.

Rubin, Donald B. 1974. "Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies." *Journal of Educational Psychology* 66(5): 689–701.

Sala-i-Martín, Xavier. 1997. "I Just Ran Two Million Regressions." *American Economic Review* 87(2): 178–83.

Scott, Steve, and Hal Varian. 2013a. "Bayesian Variable Selection for Nowcasting Economic Time Series." NBER Working Paper 19567.

Scott, Steve, and Hal Varian. 2013b. "Predicting the Present with Bayesian Structural Time Series." NBER Working Paper 19567.

Steel, Mark F. J. 2011. "Bayesian Model Averaging and Forecasting." *Bulletin of E.U. and U.S. Inflation and Macroeconomic Analysis*, 200: 30–41.

Stephens, Revor, and Curt Wehrley. 2014. "Getting Started with R." *Kaggle*, September 28. https://www.kaggle.com/c/titanic-gettingStarted/details/new-getting-started-with-r.

Sullivan, Danny. 2012. "Google: 100 Billion Searches per Month, Search to Integrate Gmail, Launching Enhanced Search App for iOS." Search Engine Land. http://searchengineland.com/google-search-press-129925.

Venables, W. N., and B. D. Ripley. 2002. *Modern Applied Statistics with S,* 4th edition. New York: Springer.

Williams, Graham. 2011. *Data Mining with Rattle and R*. New York: Springer.

Wu, Xindong, and Vipin Kumar, eds. 2009. *The Top Ten Algorithms in Data Mining*. CRC Press.

Journal of Economic Perspectives—Volume 28, Number 2—Spring 2014—Pages 29–50

High-Dimensional Methods and Inference on Structural and Treatment Effects†

Alexandre Belloni, Victor Chernozhukov, and Christian Hansen

Data with a large number of variables relative to the sample size—"high-dimensional data"—are readily available and increasingly common in empirical economics. High-dimensional data arise through a combination of two phenomena.

First, the data may be inherently high dimensional in that many different characteristics per observation are available. For example, the US Census, the Current Population Survey, the Survey of Income and Program Participation, the National Longitudinal Survey of Youth, and the American Housing Survey collect information on hundreds of individual characteristics. Economists are also increasingly using scanner datasets that record transaction-level data for households across a wide range of products, or text data where counts of words in documents may be used as variables. In both of these latter examples, there may be thousands or tens of thousands of available variables per observation.

Second, even when the number of available variables is relatively small, researchers rarely know the exact functional form with which the small number of variables enters the model of interest. Researchers are thus faced with a large set of potential variables formed by different ways of interacting and transforming the underlying variables.

■ *Alexandre Belloni is Associate Professor of Decision Sciences, Fuqua School of Business, Duke University, Durham, North Carolina. Victor Chernozhukov is Professor of Economics, Massachusetts Institute of Technology, Cambridge, Massachusetts. Christian Hansen is Wallace W. Booth Professor of Econometrics and Statistics, University of Chicago Booth School of Business, Chicago, Illinois. Their email addresses are abn5@duke.edu, vchern@mit.edu, and chansen1@chicagobooth.edu.*

† To access the Appendix, visit
http://dx.doi.org/10.1257/jep.28.2.29

doi=10.1257/jep.28.2.29

There are many statistical methods available for constructing prediction models in the presence of high-dimensional data; for example, see Hastie, Tibshirani, and Friedman (2009) for a review. These methods tend to do a good job at prediction, which is what they are designed for, but they can often lead to incorrect conclusions when inference about model parameters such as regression coefficients is the object of interest (Leeb and Pötscher 2008a, b).

The goal of this paper is to provide an overview of how innovations in "data mining" can be adapted and modified to provide high-quality inference about model parameters. These data mining methods are relevant for learning about economic parameters where they are motivated, for example, by a desire to control properly for confounding variables. Note that here we use the term "data mining" in a modern sense which denotes a principled search for "true" predictive power that guards against false discovery and overfitting, does not erroneously equate in-sample fit to out-of-sample predictive ability, and accurately accounts for using the same data to examine many different hypotheses or models.

The key concept underlying the analysis of high-dimensional data is that dimension reduction or "regularization" is necessary to draw meaningful conclusions. The need for regularization can easily be seen when one considers an example where there are exactly as many variables (plus a constant) as there are observations. In this case, the ordinary least squares estimator will fit the data perfectly, returning an R^2 of one. However, using the estimated model is likely to result in very poor forecasting properties out-of-sample because the model estimated by least squares is overfit: the least-squares fit captures not only the signal about how predictor variables may be used to forecast the outcome, but also fits the noise that is present in the given sample, and is not useful for forming out-of-sample predictions. Producing a useful forecasting model in this simple case requires regularization; that is, the estimates must be constrained so that overfitting is avoided and useful out-of-sample forecasts can be obtained.

We begin with a discussion of "approximately sparse" regression models in high-dimensional data. These models are characterized by having many potential predictor/control variables of which only a few are important for predicting the outcome. The challenge in this case is to obtain good out-of-sample forecasts of outcome (and/or treatment) variables without assuming that the researcher knows which of the many available variables actually correspond to the important predictors. We then turn to the issue of model selection with high-dimensional data when the goal is learning about specific model parameters. We show how methods designed for forecasting in approximately sparse regression models can be used in this context. To illustrate these ideas, we apply them to examples from three papers in the empirical literature: estimating the effect of eminent domain on house prices, estimating the effect of abortion on crime, and estimating the effect of institutions on economic output. Our focus is not to rework these studies in any complete way but to show how one can bring high-dimensional analysis into this work and how this introduction can influence the findings of the analysis.

Approximately Sparse Regression Models

To fix ideas, suppose we are interested in forecasting outcome y_i with controls w_i according to the model

$$y_i = g(w_i) + \zeta_i,$$

where the expected value of the error terms ζ_i given w_i is equal to zero. Further, suppose we have a sample of $i = 1, \ldots, n$ independent observations. To avoid overfitting and produce useful out-of-sample forecasts, we will generally need to restrict or regularize the function $g(\cdot)$.

There are many regularization approaches that produce the needed dimension reduction. Perhaps the simplest and most widely applied approach is the researcher making an ad hoc decision. Typically, applied researchers assume that they need only a small number of controls, which are chosen based on economic intuition and a modeling framework. Moreover, the researcher often assumes the controls enter the model in a simple fashion, usually linearly, allowing for the usual set of simple transformations and forming a small number of interaction terms. This approach has intuitive appeal and at some level is unavoidable. A researcher will always have to start by imposing some dimension reduction. However, it does leave one wondering whether the correct variables and functional forms were chosen.

Nonparametric methods, such as traditional series/sieve expansions, are also available. In this framework, one assumes that the model depends only on a small number of variables in a smooth but potentially unknown way and then uses a series of transformations of these variables in estimation; for example, see Newey (1997) and Chen (2007). Practical implementation of a nonparametric estimator requires that the researcher has selected an initial set of variables and a pre-specified set of series terms containing transformations of these variables. While more flexible than parametrics, traditional nonparametrics has a number of important limitations. Most importantly, it is again assumed that the most important terms for predicting y_i are contained within a pre-specified set of transformed variables determined by the researcher that is quite small relative to the number of observations.

In this paper, we focus on an approach to regularization that treats $g(w_i)$ as a high-dimensional, approximately linear model.[1] Specifically, we assume that

$$g(w_i) = \sum_{j=1}^{p} \beta_j x_{i,j} + r_{p,i}.$$

[1] There is also work on high-dimensional nonlinear models. For example, van de Geer (2008) and Belloni, Chernozhukov, and Wei (2013) consider high-dimensional generalized linear regression, and Belloni and Chernozhukov (2011), Belloni, Chernozhukov, and Kato (2013), and Kato (2011) consider quantile regression. We consider only linear models here for simplicity. The basic insights from the high-dimensional linear models extend to nonlinear settings though the theoretical analysis and practical computation is more complicated. High-dimensional linear models can also encompass many interesting settings and can accommodate flexible functional approximation just as nonparametric series estimators can.

The variables $x_i = (x_{i,1}, \ldots, x_{i,p})'$ may simply be the elementary regressors w_i or may be made of transformations of these elementary regressors as in series modeling. In contrast to series modeling, we allow the number of these variables p to be larger than the sample size n. The final term $r_{p,i}$ is an approximation error. As with series, it is assumed that $r_{p,i}$ is small enough relative to sampling error in a well-defined sense (Bickel, Ritov, and Tsybakov 2009; Belloni, Chen, Chernozhukov, and Hansen 2012). Without further restrictions on the model, practical inference in this kind of high-dimensional linear model remains impossible since $p \geq n$ is allowed.

A structure that has played an important role in the literature is *approximate sparsity* of the high-dimensional linear model. Approximate sparsity imposes a restriction that only s variables among all of $x_{i,j}$, where s is much smaller than n, have associated coefficients β_j that are different from 0, while permitting a nonzero approximation error $r_{p,i}$. Thus, estimators for this kind of model attempt to learn the identities of the variables with large nonzero coefficients, while simultaneously estimating these coefficients.[2]

Note that the approximately sparse high-dimensional linear model structure includes as a special case both the traditional parametric and nonparametric model. The approximately sparse high-dimensional model generalizes these approaches by allowing the researcher to consider many explanatory variables and to use the data to learn which of the many variables are the most important. This setting thus encompasses many usual approaches to data analysis and accommodates the realistic scenario where a researcher does not know a priori exactly which variables should be included in a model.

An appealing method for estimating the parameters of sparse high-dimensional linear models is the Least Absolute Shrinkage and Selection Operator (LASSO), introduced by Frank and Friedman (1993) and Tibshirani (1996), where coefficients are chosen to minimize the sum of the squared residuals plus a penalty term that penalizes the size of the model through the sum of absolute values of the coefficients. In our discussion and empirical examples, we use a variant of the LASSO estimator that we proposed in Belloni, Chen, Chernozhukov, and Hansen (2012) defined as

$$\hat{\beta} = \arg\min_b \sum_{i=1}^{n} \left(y_i - \sum_{j=1}^{p} x_{i,j}\, b_j \right)^2 + \lambda \sum_{j=1}^{p} |b_j| \gamma_j,$$

where $\lambda > 0$ is the "penalty level" and γ_j are the "penalty loadings." The penalty loadings are chosen to insure basic equivariance of coefficient estimates to rescaling of $x_{i,j}$ and can also be chosen to address heteroskedasticity, clustering, and non-normality in model errors. For more discussion, see Belloni, Chen, Chernozhukov, and Hansen (2012), Belloni, Chernozhukov, Hansen, and Kozbur (2014), and the online appendix for this paper at http://e-jep.org.

[2] Much of the high-dimensional linear model literature assumes that the model is exactly sparse, so the approximation error is identically 0. The approximately sparse model is strictly more general in that it allows for a nonzero approximation error in the analysis.

The penalty level, λ, controls the degree of penalization. Practical choices for λ that provably guard against overfitting are provided in Belloni, Chen, Chernozhukov, and Hansen (2012). (See also Belloni, Chernozhukov, Fernandéz-Val, and Hansen 2013; Belloni, Chernozhukov, Hansen, and Kozbur 2014.) It is also common to choose λ by cross-validation in prediction contexts, though it is important to note that this choice may not immediately equate to good performance when prediction is not the end goal.

The penalty function in the LASSO is special in that it has a kink at 0, which results in a sparse estimator with many coefficients set exactly to zero. Thus, the LASSO estimator may be used for variable selection by simply selecting the variables with nonzero estimated coefficients. A large part of the appeal of the LASSO estimator relative to other selection methods is that the LASSO problem is a convex optimization problem and highly efficient computational algorithms exist for its solution. LASSO-type estimators have also been shown to have appealing properties under plausible assumptions that allow for approximation errors, heteroskedasticity, clustering and fixed effects, and non-normality (Bickel, Ritov, and Tsybakov 2009; Belloni, Chen, Chernozhukov, and Hansen 2012; Belloni, Chernozhukov, Hansen, and Kozbur 2014; Gautier and Tsybakov 2011).

Finally, it is important to note that the nonzero coefficients that are part of the solution to the LASSO problem tend to be substantially biased towards zero. An appealing method to alleviate this bias is to employ the Post-LASSO estimator as in Belloni and Chernozhukov (2013) and Belloni, Chen, Chernozhukov, and Hansen (2012). The Post-LASSO estimator works in two steps. First, LASSO is applied to determine which variables can be dropped from the standpoint of prediction. Then, coefficients on the remaining variables are estimated via ordinary least squares regression using only the variables with nonzero first-step estimated coefficients. The Post-LASSO estimator is convenient to implement and, as we show in Belloni and Chernozhukov (2013) and Belloni, Chen, Chernozhukov, and Hansen (2012), works as well as and often better than LASSO in terms of rates of convergence and bias.

Model Selection When the Goal is Causal Inference

Using LASSO as a method for penalized estimation of the coefficients of a sparse linear model is useful for obtaining forecasting rules and for estimating which variables have a strong association to an outcome in a sparse framework. However, naively using the results obtained from such a procedure to draw inferences about model parameters can be problematic.

Part of the difficulty in drawing inferences after regularization or model selection is that these procedures are designed for forecasting, not for inference about model parameters. This observation suggests that more desirable inference properties may be obtained if one focuses on model selection over the predictive parts of the economic problem—the reduced forms and first-stages—rather than using model selection in the structural model directly.

The more difficult problem with doing inference following model selection is that model selection mistakes may occur. If one could be sure that the variable selector would always choose exactly all of the variables with nonzero coefficients, one could simply use the data to select this set of variables and then use the selected set coupled with any conventional procedure to do estimation and inference about parameters of interest. The validity of this approach is delicate because it relies on perfect model selection. Once one allows for the realistic scenario where some variables may have small but nonzero effects, it is likely there will be model selection mistakes in which such variables are not selected. The omission of such variables then generally contaminates estimation and inference results based on the selected set of variables. This problem is not restricted to the high-dimensional setting but is present even in low-dimensional settings when model selection is considered. This intuition is formally developed in Leeb and Pötscher (2008a, b). Because model selection mistakes seem inevitable in realistic settings, it is important to develop inference procedures that are robust to such mistakes.

An element in the provision of this robustness that has been employed recently is to focus on a small set of parameters of interest over which no model selection will be done, leaving model selection or regularization to be done only over "nuisance" parts of the problem. The estimation for the main parameters is then carried out using estimating equations that are *orthogonal* or *immune* to small perturbations in the nuisance parts. In Belloni, Chen, Chernozhukov, and Hansen (2012) and Belloni, Chernozhukov, and Hansen (2013, forthcoming), we provide an approach that does this in a canonical instrumental variable model; see also Ng and Bai (2009). In Belloni, Chernozhukov, and Hansen (2013) and Belloni, Chernozhukov, and Hansen (forthcoming), we provide an approach for inference about coefficients in a partially linear model, or about average treatment effects in a heterogeneous treatment effects model with binary treatment; see also Farrell (2013). In addition to showing how to obtain valid inference following model selection in canonical econometric models, these papers develop basic intuition helpful in understanding how inference following regularization may be performed outside of these models. Thus, we outline the approaches of these papers below.

Providing formal results for doing inference about parameters following model selection for other models relevant in applied economics is a topic of ongoing research. For example, in Belloni, Chernozhukov, Fernandéz-Val, and Hansen (2013), we consider the estimation of heterogeneous treatment effects with endogenous receipt of treatment and present a general framework for econometric models where the orthogonality condition is explained in detail.

Inference with Selection among Many Instruments

Consider the linear instrumental variables model with potentially many instruments

$$y_i = \alpha d_i + \varepsilon_i$$

$$d_i = z_i' \Pi + r_i + v_i,$$

where $E[\varepsilon_i \mid z_i] = E[v_i \mid z_i, r_i] = 0$ but $E[\varepsilon_i v_i] \neq 0$, leading to endogeneity. In this setting, d_i is a scalar endogenous variable of interest, z_i is a p-dimensional vector of instruments where the number of instruments p may be much larger than the number of observations,[3] and r_i is an approximation error. Allowing for a small number of included exogenous variables is straightforward by defining the variables in the model as residuals after partialing these variables out, and we suppress this case for simplicity. The results in Belloni, Chen, Chernozhukov, and Hansen (2012) also allow for a nonscalar but finite-dimensional treatment vector.

One approach to estimation and inference about α in this context is to select a small number of instruments from z_i to use in a conventional two-stage least squares estimation. In Belloni, Chen, Chernozhukov, and Hansen (2012), we provide a set of formal conditions under which conventional inference from the two-stage least squares estimator based on instruments selected by LASSO or another variable selection procedure is valid for learning about the parameter of interest, α. The key features that allow this can be illustrated by noting that this model cleanly fits into the heuristic outline for doing valid inference after using high-dimensional methods provided above. The parameter of interest, α, is finite-dimensional and there is no selection over whether d_i will be included in the model. The variable selection component of the problem is limited to the first-stage equation relating the endogenous variable to the instruments, which is a pure predictive relationship. Finally, the structure of the problem is such that model selection mistakes in which a valid instrument with a small but nonzero coefficient is left out of the first-stage will not substantially affect the second-stage estimator of α as long as other instruments with large coefficients are selected. In other words, the second-stage instrumental variable estimate is *orthogonal* or *immune* to variable selection errors where instruments with small, nonzero coefficients are mistakenly excluded from estimation.

Inference with Selection among Many Controls

As a more complex example, consider a linear model where a treatment variable, d_i, is taken as exogenous after conditioning on control variables:

$$y_i = \alpha d_i + x_i'\theta_y + r_{yi} + \zeta_i,$$

where $E[\zeta_i \mid d_i, x_i, r_{yi}] = 0$, x_i is a p-dimensional vector of controls where $p \gg n$ is allowed, r_{yi} is an approximation error, and the parameter of interest is α, the effect of the treatment on the outcome.

Before turning to a procedure that provides high-quality estimates and inferential statements about α, it is useful to discuss some intuitive benchmarks that do

[3] In the instrumental variables setting, there are many papers that examine the properties of various instrumental variables estimators under many-instrument asymptotics where the number of instruments p is allowed to increase with the sample size n in such a way that $p < n$ and $p/n \to \rho < 1$; see, for example, Bekker (1994), Chao and Swanson (2005), Hansen, Hausman, and Newey (2008), and Hausman, Newey, Woutersen, Chao, and Swanson (2012). These approaches do not apply when $p \geq n$ and tend to perform poorly when $p/n \approx 1$.

not work. Considering such cases builds intuition concerning features that must be guarded against in applying high-dimensional methods in this and related contexts.

One naive approach would attempt to select control variables by applying LASSO to the equation above, forcing the treatment variable to remain in the model by excluding α from the LASSO penalty. One could then try to estimate and do inference about α by applying ordinary least squares with y_i as the outcome, and d_i and any selected control variables as regressors. The problem with this approach can be seen by noting that LASSO and any other high-dimensional modeling device targets *prediction,* not learning about specific model parameters. From the standpoint of prediction, any variable that is highly correlated to the treatment variable will tend to be dropped since including such a variable will tend not to add much predictive power for the outcome given that the treatment is already in the model. Of course, the exclusion of a variable that is highly correlated to the treatment will lead to substantial *omitted-variables bias* if the coefficient in θ_y associated with the variable is nonzero. Such omissions will happen routinely in any procedure that looks just at the equation above.

There are two problems with the above naive approach. First, it ignores a key component to understanding omitted-variables bias, the relationship between the treatment variable and the controls. To aid in learning about this relationship, we introduce an additional "reduced form" relation between the treatment and controls:

$$d_i = x_i' \theta_d + r_{di} + v_i,$$

where $E[v_i | x_i, r_{di}] = 0$. The other problem is that the naive approach is based on a "structural" model where the target is to learn the treatment effect given controls, not an equation representing a forecasting rule for y_i given d_i and x_i. It is thus useful to transform the first equation of this section to a reduced form, *predictive* equation by substituting the equation introduced for d_i into the "structural" equation yielding the reduced form system:

$$y_i = x_i' (\alpha\theta_d + \theta_y) + (\alpha r_{di} + r_{yi}) + (\alpha v_i + \zeta_i) = x_i'\pi + r_{ci} + \varepsilon_i$$

$$d_i = x_i' \theta_d + r_{di} + v_i,$$

where $E[\varepsilon_i | x_i, r_{ci}] = 0$, r_{ci} is a composite approximation error, and the second equation is the same as above. Both of these equations represent predictive relationships, which may be estimated using high-dimensional methods.

Before turning to the recommended procedure, let us mention the second set of naive procedures that use only one of these two equations for selection. The problem with working with only one of the above equations is that single-equation approaches rely on there being no errors in variable selection. To see the problem with such an approach, note that applying a variable selection method to say the first equation for forecasting y_i with x_i will tend to select variables with large entries in coefficient vector π but will tend to miss variables with moderately sized coefficients. However, missing variables that have strong predictive power for d_i, namely variables with large coefficients in θ_d, may lead to substantive *omitted-variables bias* in the estimator of α if the coefficients on these variables in θ_y are moderately sized. Intuitively, any such variable has a moderate direct effect on the outcome that will be incorrectly misattributed to the effect of the treatment when this variable is strongly related to the treatment and the variable is not included in the regression. Similarly, if one applied a variable selection method to only the second equation for predicting d_i, one would potentially miss variables that have moderate-sized coefficients in predicting d_i but large direct effects on y_i. Such an omission may again lead to non-negligible omitted-variables bias.

To guard against such model selection mistakes, it is important to consider *both* equations for selection: we apply variable selection methods to *each* of the two reduced form equations and then use all of the selected controls in estimation of α. Thus, variable selection is used to select a set of variables that are useful for predicting y_i, say x_{yi}, and a set of variables that are useful for predicting d_i, say x_{di}. We then estimate α by ordinary least squares regression of y_i on d_i and the *union* of the variables selected for predicting y_i and d_i, contained in x_{yi} and x_{di}. We thus make sure we use variables that are important for either of the two predictive relationships to guard against omitted-variables bias, discussed above, when estimating α.

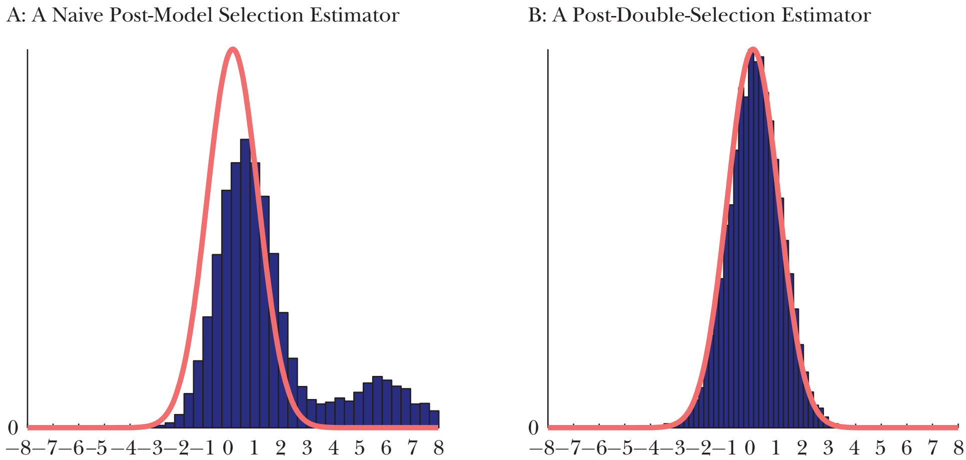
Using both variable selection steps immunizes the resulting procedure against the types of model selection mistakes discussed above for single-equation procedures. Specifically, using the variables selected in both reduced form equations ensures that any variables that have large effects in either the "structural" equation for y_i or the reduced form equation for d_i are included in the model. Any excluded variables are therefore at most mildly associated to y_i and d_i, which greatly limits the scope for omitted-variables bias. It is also noteworthy that the "double selection" procedure implicitly estimates the residuals ε_i and v_i and then regresses the estimates of ε_i on the estimates of v_i to construct an estimator of α, thereby providing a selection analog of Robinson's (1988) method for estimating the parameters of a partially linear model to the high-dimensional case.

In Belloni, Chernozhukov, and Hansen (2013, forthcoming), we provide formal conditions under which this "double selection" procedure will lead to valid inference about α even when selection mistakes are allowed, and provide substantial simulation evidence that the procedure works across a wide variety of models. Using both selection steps also enhances efficiency by finding variables that are strongly predictive of the outcome and may remove residual variance.[4]

As a concrete illustration of the potential pitfalls of naive procedures and the robustness of the "double selection" approach, we present results from a simulation

[4] That is, standard errors may go down, at least theoretically, after performing the variable selection steps if the selected variables reduce residual variance sufficiently to offset the increased variability due to including more variables. In fact, under homoskedasticity, the estimator is semi-parametrically efficient, achieving the efficiency bound of Robinson (1988).

Figure 1
The "Double Selection" Approach to Estimation and Inference versus a Naive Approach: A Simulation from Belloni, Chernozhukov, and Hansen (forthcoming)
(distributions of estimators from each approach)



Source: Belloni, Chernozhukov, and Hansen (forthcoming).
Notes: The left panel shows the sampling distribution of the estimator of α based on the first naive procedure described in this section: applying LASSO to the equation $y_i = d_i + x_i' \theta_y + r_{yi} + \zeta_i$ while forcing the treatment variable to remain in the model by excluding α from the LASSO penalty. The right panel shows the sampling distribution of the "double selection" estimator (see text for details) as in Belloni, Chernozhukov, and Hansen (forthcoming). The distributions are given for centered and studentized quantities.

exercise in this linear modeling context in Figure 1. Details underlying the simulation are as in Belloni, Chernozhukov, and Hansen (2013). The left panel shows the sampling distribution of the estimator of α based on the first naive procedure discussed in this section, while the right panel shows the sampling distribution of the "double selection" estimator. The second mode in the left panel is due to model selection mistakes where important variables are missed leading to badly biased estimates of α. This strong omitted-variables bias is absent from the distribution of the "double selection" estimator, which was specifically designed to reduce the influence of such mistakes as discussed above.

Some Empirical Examples

In this section, we provide three concrete examples of the use of these methods. An online Appendix available with this paper at http://e-jep.org provides implementation details.

Estimating the Impact of Eminent Domain on House Prices

We consider instrumental variable estimation of the effects of federal appellate court decisions regarding eminent domain on housing prices. Recall that eminent domain refers to the government's taking of private property. Federal court rulings that a government seizure was unlawful (pro-plaintiff rulings) thus uphold individual property rights and make future exercise of eminent domain more difficult due to the structure of the US legal system. A more detailed discussion of the economics of takings law (or eminent domain) and other institutional and econometric considerations can be found in Belloni, Chen, Chernozhukov, and Hansen (2012) and Chen and Yeh (2012).

The analysis of the effects of takings law is complicated by the possible endogeneity between takings law decisions and economic variables: for example, a taking may be less likely if real estate prices are low and sellers are eager to unload property. To address the potential endogeneity of takings law, we employ an instrumental variables strategy based on the identification argument of Chen and Sethi (2010) and Chen and Yeh (2012) that relies on the random assignment of judges to federal appellate panels. Because judges are randomly assigned to three-judge panels to decide appellate cases, the exact identity of the judges and their demographics are randomly assigned conditional on the distribution of characteristics of federal circuit court judges in a given circuit-year. Under this random assignment, the characteristics of judges serving on federal appellate panels can only be related to property prices through the judges' decisions; thus the judge's characteristics will plausibly satisfy the instrumental variable exclusion restriction.

Following this argument, we try to uncover the effect of takings law by estimating models of the form

$$\log(\mathit{Case\text{–}Shiller}_{ct}) = \alpha \cdot \mathit{TakingsLaw}_{ct} + \beta_c + \beta_t + \gamma_c t + W_{ct}'\delta + \varepsilon_{ct}$$

using the characteristics of judges actually assigned to cases as instruments for $\mathit{TakingsLaw}_{ct}$. In this equation, $\mathit{Case\text{–}Shiller}_{ct}$ is the average of the Case–Shiller home price index within circuit court c at time t; $\mathit{TakingsLaw}_{ct}$ represents the number of pro-plaintiff appellate takings decisions in federal circuit court c and year t; W_{ct} are included exogenous variables that include a dummy variable for whether there were relevant cases in that circuit-year, the number of takings appellate decisions, and controls for the distribution of characteristics of federal circuit court judges in a given circuit-year; and β_c, β_t, and $\gamma_c t$ are respectively circuit-specific effects, time-specific effects, and circuit-specific time trends. An appellate court decision is coded as pro-plaintiff if the court ruled that a taking was unlawful, thus overturning the government's seizure of the property in favor of the private owner. The parameter of interest, α, thus represents the effect of an additional decision upholding individual property rights on a measure of property prices. The sample size in this example is 183.

The argument given above suggests that judges' characteristics satisfy the instrumental variables exclusion restriction. Of course, to be valid instruments, the characteristics must also be useful for predicting judicial decisions. In the data, we observe a variety of demographic information about each judge, and the basic identification argument suggests that any set of characteristics of the three-judge panel will be unrelated to structural unobservables. Given the large number of instruments that could be constructed by considering all combinations of characteristics of three-judge panels, it is also infeasible to just use all possible instruments.

Thus, a sensible way to proceed is to use variable selection methods to find a set of good instruments from a large set of intuitively chosen potential instruments. Under the exclusion restriction, the ideal set of instruments provides a high-quality prediction of the endogenous variable—judicial decisions in this example. Forming high-quality predictions, which is of course different from obtaining a good in-sample fit, is exactly what LASSO and other data mining procedures are designed to do. Note that using LASSO with proper penalty parameters theoretically guarantees that any instruments selected are not simply spuriously correlated to the endogenous variable but have true predictive power. This guarantee means that LASSO could select no instruments at all as there may be no set of variables with sufficient predictive power to achieve the required standard.

Intuitively, reliably distinguishing true predictive power from spurious association becomes more difficult as more variables are considered. This intuition can be seen in the theory of high-dimensional variable selection methods, and the methods work best in simulations when selection is done over a collection of variables that is not overly extensive. It is therefore important that some persuasive economic intuition exists to produce a carefully chosen, well-targeted set of variables to be selected over even when using automatic variable selection methods.

In this example, we first did dimension reduction by intuitively selecting characteristics thought to have strong signals about judge preferences over government versus individual property rights. We chose to consider only gender, race, religion (Jewish, Catholic, Protestant, evangelical, not-religious), party affiliation, source of academic degrees (bachelor's degree from an in-state university, bachelor's degree from a public university, JD from a public university, has an LLM or SJD), and whether the judge had been elevated from a district court. For each of these baseline variables, we then constructed three new variables, counting the number of panels with one member with each characteristic, two members with each characteristic, and three members with each characteristic. To allow for nonlinearities, we included first-order interactions between all of the previously mentioned variables, a cubic polynomial in the number of panels with at least one Democrat, a cubic polynomial in the number of panels with at least one member with a JD from a public university, and a cubic polynomial in the number of panels with at least one member elevated from within the district. In addition to limiting the selection to be over this set of baseline variables, we did additional pre-processing to remove instruments that we thought likely to be irrelevant based on features of the instrument set alone. We removed any instrument where the standard deviation was extremely small and also removed one instrument from any pair of instruments that had a bivariate correlation exceeding .99 in absolute value. These instruments were removed as they were highly unlikely to have much power.[5] After these initial choices, we are left with a total of 147 instruments. The number of instruments plus the number of control variables is greater than the number of observations in this example, so conventional instrumental variables estimators using the full set of variables are not defined.

With this set of 147 instruments, we then estimate the first-stage relationship using LASSO. The estimated coefficients have just one nonzero element, the coefficient on the number of panels with one or more members with JD from a public university squared. Using this instrument gives a first-stage coefficient of 0.4495 with estimated standard error of 0.0511—that is, this variable appears to be a strong instrument. The second stage estimate using the LASSO-selected instrument is then 0.0648 with estimated standard error of 0.0240. This estimate is statistically significant at the usual levels, suggesting that a single additional judicial decision reinforcing individual property rights is associated with between 2 and 11 percent higher property prices with an average number of pro-plaintiff decisions per year of 0.19.

For comparison, we also experimented with an "intuitive" instrumental variable using the number of judicial panels with one or more Democrats. The political affiliation of judges is known to predict judicial decisions in several contexts, so one might hypothesize that this intuition carries over to judicial decisions regarding eminent domain. When we used the number of panels with one or more judges identified as Democrats as the single instrument, we found that one would not reject the hypothesis that this instrument is unrelated to the endogenous variable, the number of pro-plaintiff decisions, at any reasonable confidence level—which in turn suggests that this instrument is too weak to be useful.

We suspect that most analysts, relying on intuition about what might be a useful instrumental variable, would not have intuited that they should use the number of judicial panels with one or more members with JD from a public university squared. However, we find that one obtains a much stronger first-stage relationship in a two-stage least squares approach using this instrument selected by a formal variable selection method, relative to that obtained by an "intuitive" benchmark. This stronger first-stage in turn leads to a corresponding sensible and reasonably precise second-stage estimate. This substantive difference suggests that high-dimensional techniques may usefully complement researchers' intuition for choosing instruments and strengthen their ability to draw useful conclusions from the data.

Estimating the Effect of Legalized Abortion on Crime

Donohue and Levitt (2001) sought to estimate the effect of abortion on crime rates. Looking at state-level data, they find that higher rates of abortion in the years

[5] Note that selection based on characteristics of the instruments without reference to the endogenous variable or outcome cannot introduce bias as long as the instruments satisfy the instrumental variable exclusion restriction.

around 1970, as legal restrictions on abortion were eased in a number of states, are associated with lower rates of crime two decades later. However, there is a basic problem in estimating the causal impact of abortion on crime: state-level abortion rates during the earlier time period were not randomly assigned. It seems at least plausible that certain factors may be associated with both state-level abortion rates and state-level crime rates. Failing to control for these factors will then lead to omitted-variables bias in the estimated abortion effect.

To address these potential confounding factors, Donohue and Levitt (2001) estimate a differences-in-differences style model for state-level crime rates running from 1985 to 1997. Their basic specification is

$$y_{cit} = \alpha_c a_{cit} + w_{it}'\beta_c + \delta_{ci} + \gamma_{ct} + \varepsilon_{cit}.$$

The dependent variable, y_{cit}, indexes the crime-rate for crime type c (categorized between *violent, property,* and *murder*) in state i in year t. On the right-hand side, the independent variables are w_{it}, a measure of the abortion rate relevant for type of crime c (as determined by the ages of criminals when they tend to commit crimes); w_{it}, a set of variables to control for time-varying confounding state-level factors; δ_{ci}, state-specific effects that control for any time-invariant state-specific characteristics; and γ_{ct}, time-specific effects that control for national aggregate trends. For control variables w_{it}, Donohue and Levitt (2001) include the log of lagged prisoners per capita, the log of lagged police per capita, the unemployment rate, per-capita income, the poverty rate, the generosity of the Aid to Families with Dependent Children (AFDC) welfare program at time $t - 15$, a dummy for having a concealed weapons law, and beer consumption per capita. They present baseline estimation results based on this formulation as well as results from different models which vary the sample and set of controls in their tables IV and V. We refer the reader to the original paper for additional details, data definitions, and institutional background.

In this example, we take first-differences of the basic Donohue-Levitt formulation as our baseline. We use the same state-level data as Donohue and Levitt (2001) but delete Washington, DC,[6] which gives a sample with 50 cross-sectional observations and 12 time-series observations for a total of 600 observations. With these deletions, our baseline estimates using the same controls are quite similar to those reported in Donohue and Levitt (2001). Estimates of the effect of abortion on crime from this first-difference model are given in the first row of Table 1. These baseline results suggest that increases in abortion rates are strongly associated with decreases in crime rates; for example, an increase in the effective abortion rate of

[6] Removing Washington DC produces results similar to those in Donohue and Levitt (2001) without the need to introduce the weights used in Donohue and Levitt (2001) and is done for simplicity. This similarity between the weighted results and unweighted results excluding Washington DC is also discussed in Donohue and Levitt (2001).

Table 1
Effect of Abortion on Crime

	Type of crime					
	Violent		*Property*		*Murder*	
Estimator	*Effect*	*Std. error*	*Effect*	*Std. error*	*Effect*	*Std. error*
First-difference	−.157	.034	−.106	.021	−.218	.068
All controls	.071	.284	−.161	.106	−1.327	.932
Double selection	−.171	.117	−.061	.057	−.189	.177

Notes: This table reports results from estimating the effect of abortion on violent crime, property crime, and murder. The row labeled "First-difference" gives baseline first-difference estimates using the controls from Donohue and Levitt (2001). The row labeled "All controls" includes a broad set of controls meant to allow flexible trends that vary with state-level characteristics. The row labeled "Double selection" reports results based on the double selection method outlined in this paper and selecting among the variables used in the "All controls" results.

100 per 1,000 live births is associated with around a 15 percent reduction in violent crime. This association may be taken as causal under the assumption that all potential confounding factors not captured in w_{it} are either time-invariant or captured by a national trend.

Due to the inclusion of state and time effects, the baseline specification will control for any factors related to abortion and crime rates that are either time-invariant or vary only at the national level. While this formulation is fairly flexible, it produces valid estimates of the causal effect of abortion on crime rates *only if* time-varying state-specific factors that are correlated to both abortion and crime rates are captured by a small set of characteristics. An approach that is sometimes used to help alleviate such concerns is to include a set of state-specific linear time trends in the model to account for differences in state-specific trends that may be related to both the outcome and treatment variable of interest. However, this approach introduces many additional variables. Perhaps more importantly, the assumption of a linear state-specific trend is questionable in many circumstances as an approximation and certainly cannot capture the evolution of variables such as the crime rate or the abortion rate over any long time horizon.

Instead of using state-specific linear trends, we consider a generalization of the baseline model that allows for nonlinear trends interacted with observed state-specific characteristics and then use variable selection methods to find potentially important confounding variables. This approach allows us to consider quite flexible models without including so many additional variables that it becomes mechanically impossible to learn about the abortion effect. A key choice in using high-dimensional variable selection methods is the set of candidate variables to consider. For this example, our choice of these variables was motivated by our desire to accommodate a flexible trend that might offer a sensible model of the evolution of abortion or crime rates over a 12-year period. To accomplish this, we use the double-selection procedure outlined in the previous section with models of the form

$$\Delta y_{cit} = \alpha_c \Delta a_{cit} + z_{cit}'\beta_c + \tilde{\gamma}_{ct} + \Delta\varepsilon_{cit}$$

$$\Delta a_{cit} = z_{cit}'\Pi_c + \tilde{\kappa}_{ct} + \Delta v_{cit}.$$

In this formulation, $\Delta y_{cit} = y_{cit} - y_{cit-1}$ and Δa_{cit}, $\Delta\varepsilon_{cit}$, and Δv_{cit} are defined similarly; $\tilde{\gamma}_{ct}$ and $\tilde{\kappa}_{ct}$ are time effects; z_{cit} is a large set of controls; and we have introduced an equation for the abortion rate to make the relation to the earlier discussion clear. z_{itc} consists of 284 variables made up of the levels, differences, initial level, initial difference, and within-state average of the eight state-specific time-varying observables, the initial level and initial difference of the abortion rate relevant for crime type c, quadratics in each of the preceding variables, interactions of all the aforementioned variables with t and t^2, and the main effects t and t^2. This set of variables corresponds to a cubic trend for the level of the crime rate and abortion rate that is allowed to depend on observed state-level characteristics.

Because the set of variables we consider has fewer elements than there are observations, we can estimate the abortion effect after controlling for the full set of variables. Results from ordinary least squares regression of the differenced crime rate on the differenced abortion rate, a full set of time dummies, and the full set of variables in z_{itc} are given in the second row of Table 1. The estimated abortion effects are extremely imprecise with confidence intervals at the usual levels including implausibly large negative and implausibly large positive values for the abortion effect across all three outcomes. Of course, very few researchers would consider using 284 controls with only 600 observations because of exactly this issue.

The final row of Table 1 provides the estimated abortion effects based on the double-selection method of Belloni, Chernozhukov, and Hansen (forthcoming). At each stage of the process, we include the full set of time dummies without penalizing the parameters on these variables, which results in their selection in all cases, as we wish to allow for a flexible aggregate trend. In this example, we use LASSO to select variables from z_{cit} that are useful for predicting the change in crime rate Δy_{cit} and the change in the associated abortion rate. We then use the union of the set of selected variables, including time effects, as controls in a final ordinary least squares regression of Δy_{cit} on Δa_{cit}. In all equations, the selected variables suggest the presence of a nonlinear trend that depends on state-specific characteristics.[7] Looking at

[7] For violent crime, lagged prisoners, lagged police, lagged police × t, the initial income difference, the initial income difference × t, the initial beer consumption difference × t, average income, average income × t, and the initial abortion rate are selected in the abortion equation; and no variables are selected in the crime equation. For property crime, lagged prisoners, lagged police, lagged income, the initial income difference, the initial income difference × t, average income, and the initial abortion rate are selected in the abortion equation; and initial income squared × t^2 and average income squared × t^2 are selected in the crime equation. For the murder rate, lagged prisoners, lagged prisoners × t, lagged police × t, the initial income difference × t, average income × t, the initial abortion

the results, we see that estimated abortion effects are much more precise than the "kitchen sink" results that include all controls. However, the double-selection estimates for the effect of abortion on crime rates still produce 95 percent confidence intervals that encompass large positive and negative values.

It is interesting that one would draw qualitatively different conclusions from the estimates obtained using formal variable selection than from the estimates obtained using a small set of intuitively selected controls. Specifically, one would conclude that increases in abortion have a strong negative effect on crime rates using a small set of intuitively selected controls but would fail to reject the hypothesis that abortion is unrelated to crime rates at usual significance levels using estimates obtained using formal variable-selection. Of course, this comparison does not mean that the effect of the abortion rate provided in the first row of Table 1 is inaccurate for measuring the causal effect of abortion on crime. It does, however, imply that this conclusion is not robust to the presence of fairly parsimonious nonlinear trends. Foote and Goetz (2008) reach a similar conclusion based on an intuitive argument.[8]

Estimating the Effect of Institutions on Output

For our final example, we consider estimation of the effect of institutions on aggregate output following the work of Acemoglu, Johnson, and Robinson (2001). Estimating the effect of institutions on output is complicated by the clear potential for simultaneity between institutions and output: specifically, better institutions may lead to higher incomes, but higher incomes may also lead to the development of better institutions. To help overcome this simultaneity, Acemoglu, Johnson, and Robinson (2001) use mortality rates for early European settlers as an instrument for institution quality. The validity of this instrument hinges on the argument that settlers set up better institutions in places where they are more likely to establish long-term settlements; that where they are likely to settle for the long term is related to settler mortality at the time of initial colonization; and that institutions are highly persistent. The exclusion restriction for the instrumental variable is then motivated by the argument that GDP, while persistent, is unlikely to be strongly influenced by mortality in the previous century, or earlier, except through institutions.

In their paper, Acemoglu, Johnson, and Robinson (2001) note that their instrumental variable strategy will be invalidated if other factors are also highly persistent and related to the development of institutions within a country and to the country's GDP. A leading candidate for such a factor, as they discuss, is geography. Thus, they control for the distance from the equator in their baseline specifications and also consider specifications with different sets of geographic controls such as dummy variables for continents; see their table 4.

rate, and the initial abortion rate $\times$ t are selected in the abortion equation; and no variables are selected in the crime equation.

[8] See also Donohue and Levitt's (2008) response to Foote and Goetz (2008), which considers the same problem using a longer panel and finds similar results to Donohue and Levitt (2001) including state-specific linear time trends.

As a complement to these results, we consider using high-dimensional methods to aid in estimating the model

$$\log(GDPpercapita_i) = \alpha \cdot ProtectionfromExpropriation_i + x_i'\beta + \varepsilon_i.$$

We use the same set of 64 country-level observations as Acemoglu, Johnson, and Robinson (2001). *ProtectionfromExpropriation* is a measure of the strength of individual property rights that is used as a proxy for the strength of institutions, and x_i is a set of variables that are meant to control for geography. The underlying identifying assumption is the same as that employed in Acemoglu, Johnson, and Robinson (2001), which is that mortality risk is a valid instrument after controlling for geography. Acemoglu, Johnson, and Robinson (2001) address this by assuming that the confounding effect of geography is adequately captured by a linear term in distance from the equator or a set of dummy variables. The use of high-dimensional methods allow us to replace this assumption by the assumption that geography can be sufficiently controlled for by a small number of variables constructed from geographic information whose identities will be learned from the data.

To make use of high-dimensional methods, we note that the model in this example is equivalent to the three-equation system

$$\log(GDPpercapita_i) = \alpha \cdot ProtectionfromExpropriation_i + x_i'\beta + \varepsilon_i$$

$$ProtectionfromExpropriation_i = \pi_1 \cdot SettlerMortality_i + x_i'\Pi_2 + v_i$$

$$SettlerMortality_i = x_i'\gamma + u_i,$$

which yields three reduced form equations relating the structural variables to the controls:

$$\log(GDPpercapita_i) = x_i'\tilde{\beta} + \tilde{\varepsilon}_i$$

$$ProtectionfromExpropriation_i = x_i'\tilde{\Pi}_2 + \tilde{v}_i$$

$$SettlerMortality_i = x_i'\gamma + u_i.$$

We can thus select a set of control terms by carrying out variable selection for each of these three reduced form equations using the essential idea outlined in the discussion of selecting control variables. Valid estimation and inference for the parameter α can then proceed by conventional instrumental variable estimation using $SettlerMortality_i$ as an instrument for $ProtectionfromExpropriation_i$ with the union of variables selected from each reduced form as included control variables.

It is important that a set of baseline variables be selected before variable selection methods are applied. Our target is to control for geography, so we consider

Table 2
Effect of Institutions on Output

	Latitude	*All controls*	*Double selection*
First stage	−0.5372	−0.2182	−0.5429
	(0.1545)	(0.2011)	(0.1719)
Second stage	0.9692	0.9891	0.7710
	(0.2128)	(0.8005)	(0.1971)

Notes: In an exercise that follows the work of Acemoglu, Johnson, and Robinson (2001), this table reports results from estimating the effect of institutions, using settler mortality as an instrument. The row "First Stage" gives the first-stage estimate of the coefficient on settler mortality obtained by regressing "$ProtectionfromExpropriation_i$" on "$SettlerMortality_i$" and the set of control variables indicated in the column heading. The row "Second stage" gives the estimate of the structural effect of institutions on log(GDP per capita) using "$SettlerMortality_i$" as the instrument and controlling for variables as indicated in the column heading (see text for details). Each column reports the results for different sets of control variables. The column "Latitude" controls linearly for distance from the equator. The column "All controls" includes 16 controls defined in the main text and in footnote 9, and the column "Double selection" uses the union of the set of controls selected by LASSO for predicting GDP per capita, for predicting institutions, and for predicting settler mortality. Standard errors are in parentheses.

a flexible but still parsimonious set of variables constructed from geography. Specifically, we set x_i equal to the dummy variables for Africa, Asia, North America, and South America plus a cubic-spline in latitude (altogether, twelve variables for latitude).[9]

We report estimation results in Table 2. The first row of the table, labeled "First stage," gives the estimate of the coefficient on "*SettlerMortality*" from the first stage regression of "*ProtectionfromExpropriation*" on "*SettlerMortality*" and a set of control variables defined by the column headings, with the corresponding estimated standard error provided in parentheses below the coefficient estimate. The second row of the table, labeled "Second stage," gives the estimate of the structural effect of institutions on "log(*GDPpercapita*)" obtained by instrumental variables estimation of "log(*GDPpercapita*)" on "*ProtectionfromExpropriation*" using "*SettlerMortality*" as the instrument and controlling for variables as indicated in the column heading, with the estimated standard error again provided below the coefficient estimate in parentheses.

The first column of the table labeled "Latitude" gives baseline results that control linearly for latitude. These results correspond to the findings of Acemoglu, Johnson, and Robinson (2001), suggesting a strong positive effect of improved institutions on output with an underlying reasonably strong first-stage. This contrasts

[9] Specifically, we include latitude, latitude^2, latitude^3, $(\text{latitude}-.08)_+$, $(\text{latitude}-.16)_+$, $(\text{latitude}-.24)_+$, $((\text{latitude}-.08)_+)^2$, $((\text{latitude}-.16)_+)^2$, $((\text{latitude}-.24)_+)^2$, $((\text{latitude}-.08)_+)^3$, $((\text{latitude}-.16)_+)^3$, and $((\text{latitude}-.24)_+)^3$ where latitude denotes the distance of a country from the equator normalized to be between 0 and 1, the breakpoints in the latitude function were chosen by taking round numbers near the quartiles of latitude, and $f(a) = (a)_+$ is shorthand notation for $f(a) = (a)1(a > 0)$ where $1(\cdot)$ is the indicator function that returns 1 when the expression inside the parentheses is true and 0 otherwise.

strongly with the second column of the table, which gives results controlling for all 16 of the variables defined above and in footnote 9. Controlling for the full set of terms results in a very imprecisely estimated first-stage. The estimate of the effect of institutions is then unreliable given the weak first-stage.

The variable selection methods discussed in this paper are constructed to produce a reasonable trade-off between this perhaps overly flexible second case and the parsimonious first case by allowing flexible functions to be considered but only using terms which are useful for understanding the underlying reduced form relationships. The final column of Table 2 labeled "Double selection" controls for the union of the set of variables selected by running LASSO on each of the three reduced form equations. The same single variable, the dummy variable for Africa, is selected in the reduced form equations for GDP and mortality, and no variables are selected in the reduced form for the expropriation variable. Thus, the final column is simply the conventional instrumental variable estimate with the Africa dummy included as the single control variable. The results are qualitatively similar to the baseline results, though the first-stage is somewhat weaker and the estimated effect of institutions is slightly attenuated though still very strong and positive. The slightly weaker first-stage suggests that the intuitive baseline obtained by controlling linearly for latitude may be inadequate, though the results are not substantively altered in this case. Again, we believe these results suggest that high-dimensional techniques may usefully complement the sets of sensitivity analyses that researchers are already doing (such as those underlying table 4 of Acemoglu, Johnson, and Robinson 2001). High-dimensional techniques can add rigor to these exercises and thus potentially strengthen the plausibility of conclusions drawn in applied economic papers.

Conclusion

The high-dimensional methods discussed in this paper provide a useful addition to the standard tools used in applied economic research. They allow researchers to perform inference about economically interesting model parameters in settings with rich confounding information. In these settings, dimension reduction is important if one hopes to learn from the data. We have emphasized variable selection methods for achieving this dimension reduction and outlined procedures that provide valid inference in simple, canonical models allowing for inevitable variable selection mistakes. In Belloni, Chernozhukov, Fernandéz-Val, and Hansen (2013), we provide an extended treatment showing that valid post-selection inference is generally available when estimation is based on orthogonal estimating equations.

We hope readers will see that data mining done correctly is the opposite of "bad practice": it is an extremely useful tool that opens many doors in the analysis of interesting economic data. These tools allow researchers to add rigor and robustness to the "art" of variable or model selection in data analyses where the aim is to draw inferences about economically meaningful parameters. We have only skimmed the surface of the growing set of statistical methods appropriate to this setting and anticipate that there will be many developments that further extend these tools to cover an increasing set of economically relevant settings. Ultimately, the practice of considering high-dimensional data more openly coupled with appropriate methodology should strengthen the plausibility of inferences drawn from data and allow a deeper exploration of the structures underlying economic data.

References

Acemoglu, Daron, Simon Johnson, and James A. Robinson. 2001. "The Colonial Origins of Comparative Development: An Empirical Investigation." *American Economic Review* 91(5): 1369–1401.

Bekker, Paul A. 1994. "Alternative Approximations to the Distributions of Instrumental Variable Estimators." *Econometrica* 63(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, and Victor Chernozhukov. 2011. "ℓ_1-Penalized Quantile Regression for High Dimensional Sparse Models." *Annals of Statistics* 39(1): 82–130.

Belloni, Alexandre, and Victor Chernozhukov. 2013. "Least Squares after Model Selection in High-Dimensional Sparse Models." *Bernoulli* 19(2): 521–47.

Belloni, Alexandre, Victor Chernozhukov, Ivan Fernandéz-Val, and Christian Hansen. 2013. "Program Evaluation with High-dimensional Data." arXiv e-print, arXiv:1311.2645.

Belloni, Alexandre, Victor Chernozhukov, and Christian B. Hansen. 2013. "Inference for High-Dimensional Sparse Econometric Models." *Advances in Economics and Econometrics. 10th World Congress,* Vol. 3, edited by Daron Acemoglu, Manuel Arellano, and Eddie Dekel, 245–95. Cambridge University Press.

Belloni, Alexandre, Victor Chernozhukov, and Christian Hansen. Forthcoming. "Inference on Treatment Effects after Selection amongst High-Dimensional Controls." *Review of Economic Studies.*

Belloni, Alexandre, Victor Chernozhukov, Christian Hansen, and Damian Kozbur. 2014. "Inference in High-Dimensional Panel Data." Unpublished paper; available from the authors by request.

Belloni, Alexandre, Victor Chernozhukov, and Kengo Kato. 2013. "Uniform Post Selection Inference for LAD Regression and Other Z-Estimation Problems." arXiv e-print, arXiv:1304.0282.

Belloni, Alexandre, Victor Chernozhukov, and Ying Wei. 2013. "Honest Confidence Regions for Logistic Regression with a Large Number of Controls." arXiv e-print arXiv:1304.3969.

Bickel, Peter J., Ya'acov Ritov, and Alexandre B. Tsybakov. 2009. "Simultaneous Analysis of Lasso and Dantzig Selector." *Annals of Statistics* 37(4): 1705–32.

Chao, John C., and Norman R. Swanson. 2005. "Consistent Estimation with a Large Number of Weak Instruments." *Econometrica* 73(5): 1673–92.

Chen, Daniel L., and Jasmin Sethi. 2010. "Insiders and Outsiders: Does Forbidding Sexual Harassment Exacerbate Gender Inequality?" Available at SSRN: http://ssrn.com/abstract=1119968.

Chen, Daniel L., and Susan Yeh. 2012. "Growth under the Shadow of Expropriation? The Economic Impacts of Eminent Domain." http://www.sole-jole.org/13463.pdf.

Chen, Xiaohong. 2007. "Large Sample Sieve Estimation of Semi-Nonparametric Models." Chap. 76 in *Handbook of Econometrics*, Vol. 6B, edited by James J. Heckman and Edward E. Leamer, 5559–5632. Elsevier.

Donohue, John J., III, and Steven D. Levitt. 2001. "The Impact of Legalized Abortion on Crime." *Quarterly Journal of Economics* 116(2): 379–420.

Donohue, John J., III, and Steven D. Levitt. 2008. "Measurement Error, Legalized Abortion, and the Decline in Crime: A Response to Foote and Goetz." *Quarterly Journal of Economics* 123(1): 425–40.

Farrell, Max H. 2013. "Robust Inference on Average Treatment Effects with Possibly More Covariates than Observations." arXiv e-print, arXiv:1309.4686.

Foote, Christopher L., and Christopher F. Goetz. 2008. "The Impact of Legalized Abortion on Crime: Comment." *Quarterly Journal of Economics* 123(1): 407–23.

Frank, Ildiko E., and Jerome H. Friedman. 1993. "A Statistical View of Some Chemometrics Regression Tools." *Technometrics* 35(2): 109–135.

Gautier, Eric, and Alexander B. Tsybakov. 2011. "High-Dimensional Instrumental Variables Regression and Confidence Sets." Available at arXiv.

Hansen, Christian, Jerry Hausman, and Whitney K. Newey. 2008. "Estimation with Many Instrumental Variables." *Journal of Business and Economic Statistics* 26(4): 398–422.

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

Hausman, Jerry, Whitney K. Newey, Tiemen Woutersen, John Chao, and Norman Swanson. 2012. "Instrumental Variable Estimation with Heteroskedasticity and Many Instruments." *Quantitative Economics* 3(2): 211–55.

Kato, Kengo. 2011. "Group Lasso for High Dimensional Sparse Quantile Regression Models." arXiv e-print, arXiv:1103.1458.

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

Leeb, Hannes and Benedikt M. Pötscher. 2008b. "Recent Developments in Model Selection and Related Areas." *Econometric Theory* 24(2): 319–22.

Newey, Whitney K. 1997. "Convergence Rates and Asymptotic Normality for Series Estimators." *Journal of Econometrics* 79(1): 147–68.

Robinson, Peter M. 1988. "Root-*N*-Consistent Semiparametric Regression." *Econometrica* 56(4): 931–54.

Serena Ng, and Jushan Bai. 2009. "Selecting Instrumental Variables in a Data Rich Environment." *Journal of Time Series Econometrics* 1(1).

Tibshirani, Rob. 1996. "Regression Shrinkage and Selection via the Lasso." *Journal of the Royal Statistical Society Series B* 58(1): 267–88.

van de Geer, Sara A. 2008. "High-Dimensional Generalized Linear Models and the LASSO." *Annals of Statistics* 36(2): 614–45.

Journal of Economic Perspectives—Volume 28, Number 2—Spring 2014—Pages 51–74

Political Campaigns and Big Data[†]

David W. Nickerson and Todd Rogers

The all-encompassing goal of political campaigns is to maximize the probability of victory. To that end, every facet of a campaign is evaluated by how many votes an activity will generate and at what cost. To perform this cost–benefit analysis, campaigns need accurate predictions about the preferences of voters, their expected behaviors, and their responses to campaign outreach. For instance, efforts to increase voter turnout are counterproductive if the campaign mobilizes people who support the opponent. Over the past six years, campaigns have become increasingly reliant on analyzing large and detailed datasets to create the necessary predictions. While the adoption of these new analytic methods has not radically transformed how campaigns operate, the improved efficiency gives data-savvy campaigns a competitive advantage. This has led the political parties to engage in an arms race to leverage ever-growing volumes of data to create votes. This paper describes the utility and evolution of data in political campaigns.

The techniques used as recently as a decade or two ago by political campaigns to predict the tendencies of citizens appear extremely rudimentary by current standards. At that time, citizens' likely support was gauged primarily by their party affiliations and the "performance" of the precincts in which they lived (that is, what

■ *David W. Nickerson is Associate Professor of Political Science, University of Notre Dame, South Bend, Indiana. He served as the "Director of Experiments" in the Analytics Department in the 2012 re-election campaign of President Barack Obama. Todd Rogers is Assistant Professor of Public Policy, Harvard Kennedy School of Government, Cambridge, Massachusetts. He co-founded the Analyst Institute, which uses field experiments and behavioral science insights to develop best practices in progressive political communications. Their email addresses are dnickers@nd.edu and Todd_Rogers@hks.harvard.edu.*

[†] To access the Appendix and disclosure statements, visit
http://dx.doi.org/10.1257/jep.28.2.51

doi=10.1257/jep.28.2.51

percentage of the precinct had voted for a given party in the recent past). Whether a person was predicted to turn out and vote was often based on the past four general elections; for example, it was not uncommon to hear phrases like "2 of 4 voter" or "3 of 4 voter" used in campaign targeting plans. Past donors would be recontacted and asked for a flat amount of money (or perhaps asked for their highest previous contribution if that information was available) and prior volunteer captains would be recontacted, but intermittent volunteers were unlikely to appear on any lists. Back then, a "numbers-driven campaign" implied that candidates and their advisors paid close attention to poll numbers and adjusted policies in response to surveys. A memorable example of this dynamic is the story of President Clinton's advisor Dick Morris fielding a poll to choose Jackson Hole, Wyoming, as the vacation spot for the president (Kuhn 2004). Presidential campaigns targeted states based on historical notions of which states could see the vote swing either way, combined with the realities of the campaign budget.

In retrospect, the reliance of political campaigns on such rough—although often useful—heuristics is puzzling. Campaigns a decade ago already possessed considerable information on citizens' preferences based on what they had collected directly from volunteers, donors, and their own polling. Voter registration rolls were available at the state level from Secretaries of State. Detailed census information was available. Why did campaigns take so long to realize the value of information resources they already possessed?

Part of the answer is technological: adequate storage and computing power required large investments and were beyond the infrastructure of nearly all campaigns and state parties. Even if an entrepreneurial campaign made that investment, much of the available data would not have been as reliable as it is today. States were not required to keep electronic copies of which citizens voted in each past election until 2002 with the passage of the Help America Vote Act of 2002 (42 U.S.C. § 15483), so using the data on voting in federal elections would have been onerous in many regions.

But perhaps the biggest impediment to wider adoption of data-driven campaigning was simply that statistical thinking—and the human capital that produces it—had not yet taken root in the world of political consulting. Campaign consultants generate most of their business through social networks and are judged by win/loss records. Political candidates are typically trained in nonquantitative fields like law, education, and medicine and are more focused on fundraising and voter outreach than the nitty-gritty of managing a campaign. There were certainly consultants specializing in campaign data analytics, and the development of "predictive scores" for voters existed as a niche business, but most campaign decisions did not rely on these approaches. There were too few people with the skills required to make a noticeable impact on how campaigns operated and too few decisionmakers equipped to appreciate the effect that a fuller use of information could have. At that time, mail vendors were on the cutting edge of using consumer data for modeling purposes and at least a decade ahead of the political campaign learning curve (Malchow 2003).

These impediments to data-driven campaigning have changed in recent years. The costs of purchasing, storing, managing, and analyzing data have decreased exponentially. The supply of quantitatively oriented political operatives and campaign data analysts has increased as predictive analytics has gained footholds in other sectors of the economy like banking, consulting, marketing, and e-commerce. To reduce the need for individual campaigns to spend scarce funds purchasing citizen information from commercial vendors, the national parties have decided to construct, maintain, and regularly augment their own voter databases (McAuliffe with Ketten 2008, pp. 280–87).

These conditions have provided fertile ground for analytically minded consultants to apply statistical tools to campaign activities and campaign data. Contemporary political campaigns amass enormous databases on individual citizens and hire data analysts to create models predicting citizens' behaviors, dispositions, and responses to campaign contact. This data-driven campaigning gives candidates and their advisers powerful tools for plotting electoral strategy. A political campaign has limited financial resources. It can use this data-driven approach to shape decisions about who the campaign should target, with a sense of how much such contact will affect voter preferences, behaviors like fundraising, or turnout at the polls. This technology allows campaigns to target their outreach tactically at particular individuals and then also to aggregate these predictive estimates up to the jurisdiction level to inform large-scale strategic decisions.

Given that campaigns view their analytic techniques as secret weapons to be kept out of the hands of opponents, the public discourse on campaign data has been largely speculative and somewhat hypothetical, ranging from hyping the performance of the tools (Scherer 2012) to alarmist concerns about the personal privacy of voters (Duhigg 2012). This paper describes the state of contemporary campaign data analytics. We begin by explaining why campaigns need data and the "predictive scores" that they seek to calculate. We then describe where that data comes from and the techniques used to analyze political data. We conclude by noting several challenges facing campaigns as data analytics become more widely used and increasingly accurate. The analytics revolution has not radically transformed campaigns in the manner that television did in the 1960s, but in a close political contest, data-driven campaigning can have enough effect to make the difference between winning and losing.

Why Do Campaigns Need Data?

Contemporary campaigns use data in a number of creative ways, but the primary purpose of political data has been—and will be for the foreseeable future—providing a list of citizens to contact. Campaigns need accurate contact information on citizens, volunteers, and donors. Campaigns would like to record which citizens engage in specific campaign-supporting actions like donating money, volunteering, attending rallies, signing petitions, or expressing support for candidates or issues in tracking polls. Indeed, the Federal Election Commission requires campaigns and coordinated committees to disclose the identity of all individuals who contribute more than $200 during the calendar year. These disclosure requirements mean that campaigns have a legal requirement, as well as financial incentive, to maintain good lists of donors.

Campaigns also use data to construct predictive models to make targeting campaign communications more efficient and to support broader campaign strategies. These predictive models result in three categories of "predictive scores" for each citizen in the voter database: behavior scores, support scores, and responsiveness scores.

Behavior scores use past behavior and demographic information to calculate explicit probabilities that citizens will engage in particular forms of political activity. The primary outcomes campaigns are concerned with include voter turnout and donations, but other outcomes such as volunteering and rally attendance are also of interest.

Support scores predict the political preferences of citizens. In the ideal world of campaign advisers, campaigns would contact all citizens and ask them about their candidate and issue preferences. However, in the real world of budget constraints, campaigns contact a subset of citizens and use their responses as data to develop models that predict the preferences of the rest of the citizens who are registered to vote. These support scores typically range from 0 to 100 and generally are interpreted to mean "if you sample 100 citizens with a score of X, X percent would prefer the candidate/issue." A support score of "0" means that no one in a sample of 100 citizens would support the candidate/issue, "100" means that everyone in the sample would support the candidate/issue, and "50" means that half of the sample would support the candidate/issue. Support scores only predict the preferences at the aggregate level, not the individual level. That is, people assigned support scores of 50 are not necessarily undecided or ambivalent about the candidate/issue and, in fact, may have strong preferences. But when citizens have support scores of 50, it means that it is difficult to predict their political preferences.

Responsiveness scores predict how citizens will respond to campaign outreach. While there are theoretical rationales as to who might be most responsive to blandishments to vote (Arceneaux and Nickerson 2009) and attempts at persuasion (Hillygus and Shields 2008), in general predicting which types of individuals will be most and least responsive to particular direct communications in a given electoral context is difficult. Campaigns can use fully randomized field experiments to measure the average response to a campaign tactic (Gerber and Green 2000; Green and Gerber 2008; Nickerson and Rogers 2010; Arceneaux and Nickerson 2010; Nickerson 2005; Nickerson, Friedrichs, and King 2006; Bryan, Walton, Rogers, and Dweck 2011; Gerber and Rogers 2009; Bailey, Hopkins, and Rogers 2013; Rogers and Nickerson 2013). The results of these experiments can then be analyzed to detect and model heterogeneous treatment effects (in this case, predictive scores). The estimated model can then be used to predict treatment responsiveness for the entire target population and guide future targeting decisions (Issenberg 2012a, b, c).

Some of the results of these experiments can only be used to inform decisions in future elections: for example, the results of most voter turnout experiments necessarily come after Election Day. But other experiments can be conducted during the election cycle to improve efficiency in real time; for example, lessons from experiments evaluating the efficacy of treatments aimed at increasing observable behaviors like donations and volunteering can be put to immediate use. Similarly, the persuasiveness of campaign communications can be gauged through randomized experiments that measure voter preferences through post-treatment polling of the treatment and control groups. The types of citizens found to be especially responsive to the campaign treatment in these pilot experiments—as reflected in the responsiveness score—can be targeted during a larger rollout of the campaign treatment. Conversely, citizens who are unresponsive, or are predicted to respond negatively, can be avoided by the campaign.

Campaigns are primarily concerned with the practical question of how accurately predictive scores forecast the behaviors, preferences, and responses of individual citizens, not with testing an academic theory. As a result, the variables included in the construction of these scores often have thin theoretical justifications. That said, a variable in a dataset that is found to predict an outcome of interest but has no theoretical rationale for the relationship is more likely to prove to be spurious when validated against an "out-of-sample" dataset. For instance, the analyst may discover that people between the ages of 37 and 43 are more likely to support Republicans than older and younger age groups. However, there is no particular reason to suspect that this six-year cohort is especially conservative, suggesting that the finding could be a sample-specific fluke that would not generalize to the overall population. Successful predictive scores need not be based on theories or imply causal relationships, but campaign data analysts must still think critically and creatively about what variables sensibly relate to their outcomes of interest to generate predictive scores with the external validity required by campaigns.

Where Does Campaign Data Come From?

Procuring and maintaining large databases of citizens with up-to-date information from multiple sources may seem straightforward, but it is a nontrivial logistical hurdle and requires substantial financial commitment. After all, people frequently change residences and contact information (Nickerson 2006a). Campaigns also need to track their own behavior to limit awkward interactions with citizens who have been contacted multiple times previously.

In the recent past, campaigns struggled to manage and integrate the various sources of their data. The data collected by those working on digital communications rarely linked with the data collected by those working on field operations—meaning canvassing, phone calls, volunteer recruitment, and so on—or fundraising. One of the most heralded successes of the 2012 campaign to re-elect President Obama was the creation of *Narwhal*, a program that merged data collected from these digital, field, and financial sources into one database (Gallagher 2012; Madrigal 2012). As a result, the Obama re-election campaign began with a ten terabyte database (BigData-Startups 2013) that grew to be over 50 terabytes by the end of the election (Burt 2013).

The foundation of voter databases is the publicly available official voter files maintained by Secretaries of State, which ensure that only eligible citizens actually cast ballots and that no citizen votes more than once.[1] The official voter file contains a wide range of information. In addition to personal information such as date of birth and gender,[2] which are often valuable in developing predictive scores, voter files also contain contact information such as address and phone. More directly relevant to campaigns, certain details about past electoral participation are also recorded on official voter files. *Who* citizens vote for is secret, but *whether* citizens vote is reflected in official voter files—as is the method used to vote: for example, in person on Election Day or by use of absentee or another form of early voting. This information concerning past vote history unsurprisingly tends to be the most important data in the development of voter turnout behavior scores. The act of voting, of course, reveals higher propensity to vote.

The geographic location of citizens' residences can also provide valuable information, because campaigns can merge relevant Census and precinct data with the information on citizens in the voter database. Census data—such as average household income, average level of education, average number of children per household, and ethnic distribution—is useful for the development of a host of predictive scores. Campaign data analysts also append the aggregated vote totals cast for each office and issue in past elections in each citizen's precinct to individual voter records in the voter database. Even being mindful of ecological fallacy—that is, inferring someone's individual characteristics based on their membership in a larger group or cluster—this aggregate-level information in fact tends to increase predictive score accuracy.

Campaign data analysts also can append two types of data from consumer databases. First, and most essentially, they seek updated phone numbers. Phone calls are a critical feature of campaigns. While a volunteer knocking on doors will make successful contact with two to four people/hour, a volunteer making phone calls can reach 10–15 people/hour (Nickerson 2006b, 2007a). Using an automated dialer, the total can be even higher. While most official voter files contain phone numbers, they are often out of date and coverage is incomplete. Election officials only request a phone number from voters registering for the first time, and so if someone continues voting in the same jurisdiction over time, it's not uncommon to find phone numbers that are 20 years out of date. Because current phone numbers

[1] The exception to this rule is North Dakota, which does not have a voter registration system. Eligible voters simply show up and prove their eligibility by showing a valid ID, utility bill, or having a neighbor vouch for their residency.

[2] In states that were subject to the Voting Rights Act, the self-identified race of the registrants is included on official voter files, though this may change in light of the Supreme Court's June 25, 2013, ruling in *Shelby County v. Holder* 570 US ___ (2013).

are so important, campaigns find it worthwhile to purchase more accurate contact information available from consumer data firms.

Campaigns can also purchase a wide range of additional information from consumer data vendors relatively inexpensively, such as estimated years of education, home ownership status, and mortgage information. In contrast, information on magazine subscriptions, car purchases, and other consumer tastes are relatively expensive to purchase from vendors, and also tend to be available for very few individuals. Given this limited coverage, this data tends not to be useful in constructing predictive scores for the entire population—and so campaigns generally avoid or limit purchases of this kind of consumer data. The vast majority of these variables literally do nothing to increase the predictive power of models of mass behavior once prior behavior is accounted for (for example, any power of income or education measures to predict voter turnout are subsumed by controlling for prior voter turnout).

While campaigns do purchase some information, the vast majority of the useful information campaigns collect about individuals is provided by the individuals themselves. For example, those who have donated and volunteered in the past are high-value prospects for fundraising and volunteer-recruitment in the future. Moreover, the attributes of these individuals can be used to develop behavior scores to identify others who may be likely to donate or volunteer. Similarly, information about individuals who answered the phone or door in the past can be used to develop behavior scores for others who may be likely to be contactable moving forward. Data collected from online activities can be of particular value as well because such activities require a relatively low threshold for citizens to take action. For the small set of citizens who provide an email address to the campaign to receive campaign emails,[3] all of their activity concerning those emails—for example, sign up, opening emails, clicking links in emails, taking actions like signing petitions—can be tracked and used to predict levels of support for the candidate or focal issue, likelihood of taking action, and in many cases the policy areas of greatest interest (for example, imagine a voter who opens emails about taxes twice as often as any other topic). Thus, a state party or political organization can compile valuable information for developing predictive scores just by maintaining accurate records of its interactions with citizens over time.

In short, many of the claims about the information that campaigns purchase about individuals is overblown; little of the information that is most useful to campaigns is purchased. Official voter files are public records, census and precinct-level information are also freely available, and individual citizens themselves volunteer a wealth of data that can be used to develop scores that predict all citizens' behaviors and preferences. In fact, predictive scores can often allow campaigns to estimate some citizen preferences and behaviors more accurately

[3] In 2012, the Obama campaign had email addresses for 20 million supporters (Haberman 2013) compared with 13 million for the Obama campaign in 2008 and the three million addresses collected by the 2004 Kerry campaign (Vargas 2008).

than direct reports from citizens themselves (Rogers and Aida 2013; Ansolabehere and Hersh 2012). People may not be actively misrepresenting their intentions, but the desire to project a positive image of the self may lead voters to overestimate the degree to which they will participate in a given election. Again, the most important piece of information campaigns purchase tends to be phone numbers—and this is purchased with the intent of performing the old-fashioned task of calling citizens directly. Because the most useful information tends to be collected directly from citizens, one of the most valuable data acquisition activities in which campaigns engage is exchanging their information with that of other allied political organizations (when legal) to increase the breadth and scope of data that will be useful for the development of predictive scores.

An interesting result of the type of data that campaigns acquire directly from citizens is that campaigns are able to predict with greater accuracy which citizens will *support* their candidates and issues better than which citizens will *oppose* their candidates or issues. Information regarding citizens who donate, volunteer, and subscribe to email lists is available to campaigns and can be used to predict which other citizens will be similar. In contrast, citizens who do not perform such behaviors at all, or who perform similar behaviors for opposing campaigns, cannot be directly observed, so discriminating among the citizens who do not actively support a campaign is a much more challenging task. As a result the distribution of support scores typically have two to three times more voters with the highest scores (99 and 100) than the lowest (0 and 1). This imbalance does not imply that the opposition enjoys less passionate support or that the data analysts failed in their predictive task; it is a natural result of being able to observe the activity of only one campaign's supporters in an electoral competition. Similarly, because the foundations of voter databases are official voter files from states, campaigns tend to have much more information on citizens who have voted and are registered than citizens who have never voted and are not registered. Predictive models can still be constructed to predict fruitful geographies or people to target for registration drives, but the data available are much sparser and the models necessarily more coarse. This likely exacerbates the inequality in campaign communication and outreach between those who are already politically engaged and those who are not, and between voters and nonvoters (Rogers and Aida 2013).

How Do Campaigns Analyze Data to Develop Predictive Scores?

The predictive scores campaigns construct can be roughly divided into two types. The first predicts the behavior or attitudes of voters (that is, behavior scores or support scores). These models do not make any causal claim about why these individuals vote or donate or support the candidate; they merely predict the focal trait. As such, causation is not a major concern, and the goal of the analyst is primarily to avoid overfitting the data. The second type of score predicts how voters will respond to campaign outreach (that is, responsiveness scores). These responsiveness scores typically come from exploring heterogeneous reactions to campaign treatments in randomized field experiments. The causal effect of the campaign outreach is established by the experiment and these estimated effects are used as parameters for strategic decisionmaking. However, the moderators predicting strongly positive or weakly positive (or even negative) responsiveness to the treatment are not causal. In other words, the data may have been generated by an experiment, but the enterprise of modeling responsiveness to the treatment remains a matter of finding observed differences across types of subjects that predict large or small treatment effects. For instance, a campaign data analyst may discover that women are more responsive to a treatment than men, but since gender was not randomly manipulated by the campaign it is impossible to know that gender *caused* the differential response to treatment. The campaign data analyst only knows that gender is *correlated* with treatment responsiveness. Thus, even the search for moderators of the treatment effect in an experiment is essentially observational in nature.

Most of the analytic techniques employed by campaign data analysts are taught in standard undergraduate econometrics or statistics classes. Currently, the vast majority of the predictive scores used by campaigns are created by a campaign data analyst (or a team of them) using simple regression techniques: ordinary least squares for continuous outcomes; logistic regression for binary outcomes; and, rarely, tobit for truncated data like dollars donated or hours volunteered. The skills necessary for developing such models are widespread, and the models can easily be customized to specific political environments. For instance, party registration is not predictive of candidate preference for older citizens in many Southern states—because the South was historically solidly Democratic and remained so at the state level well after the civil rights movement transformed the national political environment—but campaign data analysts attuned to contextual facts like this can accommodate them in regression analyses.

There are two major downsides to using regression techniques for constructing campaign models. First, the utility of techniques that uncover correlations is highly dependent on the talent of the particular campaign data analyst employing them. A capable campaign data analyst who is familiar with the properties of the variables available in voter databases can generate highly accurate predictive scores for citizens. However, a slightly less-capable campaign data analyst might generate predictive scores that are only slightly better than the unsophisticated methods employed by earlier campaigns. As an example, consider the task of predicting a person's likelihood of voting in an election. Controlling for the whole set of turnout history available (often more than 50 elections) will typically predict around one-third more variance in individual turnout than the old "of 4" rule of thumb (that is, did the person vote in 0, 1, 2, 3, or 4 of the past elections). However, these variables all tap into a common latent propensity to vote and exhibit considerable collinearity. As a result, the coefficient for several of these variables will be negative and statistically significant. There is no theoretical rationale for why turnout in one election would decrease turnout in a future election, so observing negative coefficients would suggest that the analyst has overfitted the data and should pare back the number of variables used or model the propensity for turnout differently. Experienced analysts also construct relevant variables (for example, past turnout among people in the household) and insert theoretically informed interactions (for example, ethnicity of the voter by ethnicity of the candidate) to improve model fit. The marginal gains from these new variables are rarely as large as the initial gains from using a wide range of past turnout decisions, but that is to be expected—the gains from good predictive models are incremental. Since the people running campaigns rarely have experience or expertise in data analytics, the competence of the campaign data analysts they employ cannot be taken for granted.

The second drawback to using regression techniques in campaign models is that unique regression models typically need to be constructed for different regions, issues, and candidates, so the "modeling by hand" approach to analysis offers few economies of scale. While individual campaign data analysts likely become more efficient with each successive model they develop, constructing models for multiple races around the country requires either a small army of campaign data analysts, or else settling for very general national models that are not adapted for local contexts.

Thus, campaign data analysts have been seeking more systematic methods for selecting a preferred regression. The commercial marketing industry often uses a form of "machine learning" (for example, *k*-means clustering or *k*-nearest neighbor classifiers; see Gan, Ma, and Wu 2007) to divide consumers into categorical types like "blue collar, grilling, SUV owner." However, these statistical methods to group similar individuals or households are less useful for campaign data analysts because strategic cost–benefit decisions in campaign planning are based on individual-specific probabilities for particular outcomes, and knowing that a set of citizens are similar in many dimensions does not assist with targeting if those dimensions are not highly correlated with behaviors like voting, ideology, and propensity to donate. For this reason, *supervised learning* algorithms are typically more appropriate for the task of modeling political data.

Supervised machine learning includes methods such as classification and regression trees (Breiman, Friedman, Stone, and Olshen 1984). In a regression tree approach, the algorithm grows a "forest" by drawing a series of samples from existing data; it divides the sample based on where the parameters best discriminate on the outcome of interest; it then looks at how regressions based on those divisions would predict the rest of the sample and iterates to a preferred fit. The researcher chooses the number of "trees"—that is, how many times the data will be divided. In the particularly popular "random forests" algorithm for implementing a regression tree (Breiman 2001), the algorithm uses only a randomly drawn subset of variables in each tree to decide on the fit rather than the entire set of available variables. The payoff for this approach is that it generates estimates of what parameters are most important: that is, what parameters add the most predictive power when the group of other parameters is unchanged. Aside from its analytical advantages, "random trees" is a popular decision tree ensemble algorithm because it has very few tuning parameters and is available as an **R**-package, so that analysts with little formal education in statistics can develop the models. Bayesian Additive Regression

Trees have similar advantages (Chipman, George, and McCollough 2010; Green and Kern 2012).

Supervised machine learning presents three major advantages for campaign data analytics. First, these classes of estimators are typically nonlinear, so commonly known nonlinear relationships—such as the curvilinear relationship between age and turnout (older cohorts vote at higher rates than younger cohorts but this relationship peaks among group 60–70 years old and then reverses)—are easily accommodated by the algorithms. Second, the approach involves less discretion for the individual campaign data analyst, so the quality of the predictive scores generated is not as heavily dependent on the capabilities and integrity of analysts. People constructing the models still need to input the most diagnostic variables and set up rigorous out-of-sample tests to validate the models, but the algorithms are written in advance and run identically for every citizen in the voter database. Finally, these data-mining algorithms are relatively scalable. Some techniques may be computationally intensive and the variables included may need to be customized, but generally the marginal cost of constructing additional models is lower using these algorithms than having a campaign data analyst construct new models from similar databases by building a series of regressions from the ground up.

The major downside of these regression tree algorithms from the campaign's perspective is that their use is relatively new and not widespread, and it will take experience to see how to trim the regression trees and customize the tuning parameters in a way that satisfies political requirements. Campaign data analysts must also take great care to not overfit their models to their data (Dietterich 1995), in which case the results become less likely to apply outside the model. Typically, there will not be sufficient data from any single jurisdiction to create a unique model, so the data from several jurisdictions will need to be pooled to produce useful predictive scores. Most algorithms can be adapted to accommodate jurisdiction-specific political requirements, but only a small fraction of campaign data analysts today have the necessary skill set. In sum, as campaign data analytics becomes more common, sophisticated, and mature, it will likely move away from judgment-based regressions to regressions based on customized machine learning algorithms like regression trees.

How Are Predictive Scores Used?

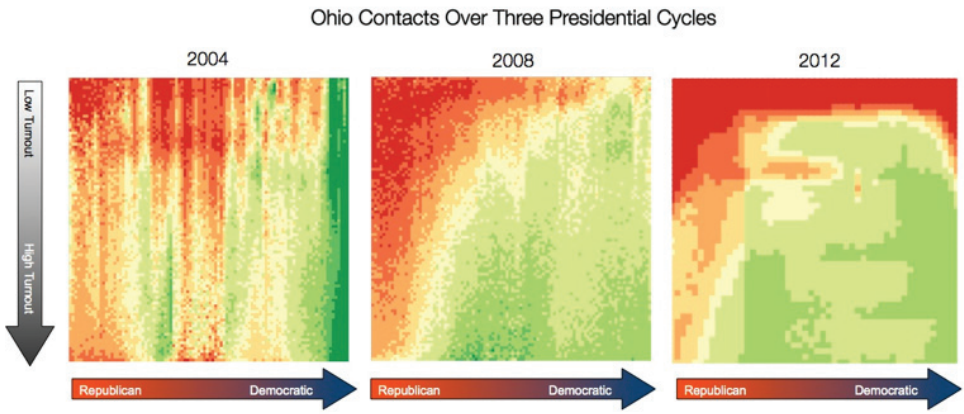
Campaigns use predictive scores to increase the efficiency of efforts to communicate with citizens. For example, professional fundraising phone banks typically charge $4 per completed call (often defined as reaching someone and getting through the entire script), regardless of how much is donated in the end. Suppose a campaign does not use predictive scores and finds that upon completion of the call 60 percent give nothing, 20 percent give $10, 10 percent give $20, and 10 percent give $60. This works out to an average of $10 per completed call. Now assume the campaign sampled a diverse pool of citizens for a wave of initial calls. It can then look through the voter database that includes all citizens it solicited for donations and all the donations it actually generated, along with other variables in the database such as past donation behavior, past volunteer activity, candidate support score, predicted household wealth, and Census-based neighborhood characteristics (Tam Cho and Gimpel 2007). It can then develop a fundraising behavior score that predicts the expected return for a call to a particular citizen. These scores are probabilistic, and of course it would be impossible to only call citizens who would donate $60, but large gains can quickly be realized. For instance, if a fundraising score eliminated half of the calls to citizens who would donate nothing, so that the resulting distribution would be 30 percent donate $0, 35 percent donate $10, 17.5 percent donate $20, and 17.5 percent donate $60, then the expected revenue from each call would increase from $10 to $17.50. Fundraising scores that increase the proportion of big donor prospects relative to small donor prospects would further improve on these efficiency gains.

The same logic can be applied to target expenditures for voter mobilization and persuasive communications. Targeting persuasive communications to citizens who are extremely unlikely to vote is inefficient. Even if the persuasive communication were effective at convincing these citizens to support the campaign's candidate or issue, the usual assumption among practitioners is that changing citizens' candidate or issue preferences does not meaningfully change their likelihood of voting. A similar logic could be applied to citizens who are already extremely likely to support a campaign's candidate or issue. If the support score predicts that a citizen is 98 percent likely to support a campaign's candidate or issue, and assuming the opposing campaign's activities will not meaningfully undermine this citizen's support likelihood, one might decide that persuasive communications would be better targeted to citizens who have a moderate or low likelihood of supporting the campaign's candidate or issue, along with a high likelihood of voting. Relying on turnout scores and support scores to target persuasion efforts in this manner represents an increase in efficiency, just as fundraising scores improve the cost effectiveness of fundraising calls.

The value of using predictive scores for targeting has become widely recognized by campaigns during the past five years. Sophisticated use of these predictive scores allows campaigns to simultaneously broaden the populations targeted while pruning away groups they believe will be cost ineffective.

Catalist, LLC, is a political data vendor that compiles and maintains nationwide registration, demographic, and other political data for progressive, civic, and nonprofit organizations such as labor unions, political candidates, and other advocacy groups. They build predictive scores using this data to help their clients analyze the electorate and target their activities more efficiently. The firm provided an aggregated data visualization for showing how its targeting of populations for its clients evolved over the last three presidential elections in Ohio (Ansolabehere and Hersh 2010). The discussion that follows references analyses of data aggregations that include the activities of independent groups as well as the activities of the Kerry campaign in 2004, the Obama campaign in 2008, and Ohio candidates in 2012 other than Obama. In each election, Catalist had several hundred clients across

Figure 1
Heatmap of Ohio Contacts over Three Presidential Cycles



Source: Derived from Catalist, LLC.
Notes: The x-axis is likelihood of supporting a Democratic candidate over a Republican candidate, ranging from 0 (left) to 100 (right). The y-axis is likelihood of voting ranging, from 100 (low) to 0 (high). Colors (or in grayscale, shade) of each cell indicate how many direct contacts were made by a particular campaign. In the grayscale version of the heatmap, darker means more contacts. In the color version, dark red represents the least contacts and dark green the most contacts. Readers can see the color heatmap in the online version of this paper.

the state of Ohio, for which data on contacts across all elections was aggregated. Catalist categorizes potential Ohio voters along two scales: whether or not they are likely to vote, and whether they are more likely to vote Democratic, Republican, or in-between. Divide each of these measures into a scale with 50 gradations, making a total of 2,500 different cells. You can then create a "heat map" of how often each one of those cells is contacted by campaigns allied with Catalist, including all modes of contact for all purposes across the election cycle, as in Figure 1. The heat maps used in the political campaigns are multicolored, but our print readers will see a grayscale version instead. Because of the centrality of Ohio in the past three presidential elections, the calculations represent tens of millions of voter contacts.

Although Catalist's client base differed across all three cycles, this graphical analysis of contacts for 2004, 2008, and 2012 show the increasing reliance on predictive scores for collective voter targeting efforts (see Figure 1). In 2004, when few clients relied on predictive scores for targeting, Catalist found that most contact was concentrated among people predicted to support Democratic candidates, regardless of their likelihoods of voting. This meant that campaign resources were probably inefficiently allocated, with a substantial share going to Democrats who were extremely unlikely to vote, or to Democrats who were extremely likely to vote and did not require either mobilization or persuasion. In 2008, Catalist clients appeared to have relied more on predictive scores for their targeting. The highest concentrations of direct contacts were observed among citizens who were predicted to support Democratic candidates but who had low likelihoods of voting—that is, those who might be reasonable targets for voter mobilization. They also targeted high-turnout citizens with middling partisanship scores, who might be reasonable targets for "persuasion." The reasonableness of targeting in these ways depends on the likelihood that voters can be moved to turn out, or be persuaded. As mentioned above, a current practice is to develop "responsiveness scores" based on pilot experiments to optimize targeting—particularly for persuasion outreach. As a result, the targeting in 2008 appears much closer to optimal than was observed in 2004. The heat map of contacts for 2012 looks much the same as that of 2008 except with smoother transitions and more consistency across the landscape, suggesting even wider adoption of predictive scores for targeting. One noticeable difference between the 2012 heat map and those of previous cycles is that Catalist clients appear to have avoided communicating with citizens with the lowest turnout probabilities. Catalist's clients may have chosen this strategy for a range of reasons, but regardless of their strategic reasons, apparently Catalist's Ohio clients in 2012 used predictive scores to manifest their strategic plans in ways that they had not in previous cycles.

What Are Predictive Scores Worth?

Campaign organizations have adopted predictive scores, which suggests that they are electorally useful. They use these scores to target nearly every aspect of campaign outreach: door-to-door canvassing; direct mail; phone calls; email; television ad placement; social media outreach (like Facebook and Twitter); and even web page display. Determining exactly how much using these scores affects electoral outcomes is difficult because the counterfactual is unclear. Is the appropriate comparison for assessing the value of campaign analytics to contrast the current uses of predictive scores for targeting with a complete absence of targeting? Or would it be to compare current uses to the basic heuristics that were used for targeting in the relatively recent past? Whatever the specific choice, it is possible to derive bounds as to how much campaign analytics could matter to campaigns.

Persuasive communications is a good place to begin because targeting is so diffuse. There are so many possible targets, including potentially all citizens, and so many strategies, from shoring up support to causing opposition supporters to defect. Thus, persuasive campaign outreach can be directed almost anywhere along the support score spectrum from hard-core supporters to hard-core opponents. Many campaigns use responsiveness scores as part of targeting their persuasive communications (Issenberg 2012a, b, c). Suppose a campaign's persuasive communications has an average treatment effect of 2 percentage points—a number on the high end of persuasion effects observed in high-expense campaigns: that is, if half of citizens who vote already planned to vote for the candidate, 52 percent would support the candidate after the persuasive communication. If a campaign indiscriminately attempted to persuade 8.5 million citizens—about the size of the Florida electorate—it would generate 170,000 votes under this scenario.

Table 1
Hypothetical Example of Persuasion Responsiveness Score's Value
(assuming average effect of campaign contact is 2 percentage points and electorate size is 8.5 million)

Quintile	*Effect multiplier*	*Votes created in quintile*	*Cumulative votes*	*Improvement over no targeting*
Top 20%	3	102,000	102,000	200%
60–80%	2	68,000	170,000	150%
Middle 20%	1	34,000	204,000	100%
20–40%	0	0	204,000	50%
Bottom 20%	−1	−34,000	170,000	20%

Notes: Imagine that a campaign has created a responsiveness score that predicts which citizens would be most responsive to its persuasive communications. Based on the responsiveness score, those in the top quintile are three times more responsive to the persuasive communications than the average citizen, the next quintile is twice as responsive, the middle quintile is no more responsive than average, the second quintile shows no average responsiveness to the persuasive communications, and the bottom quintile actually exhibited backlash to the persuasive communications equal to the overall average treatment effect.

Now imagine that the campaign has created a responsiveness score that predicts which citizens would be most responsive to its persuasive communications. Based on the responsiveness score, those in the top quintile are three times more responsive to the persuasive communications than the average citizen, the next quintile is twice as responsive, the middle quintile is no more responsive than average, the second quintile shows no average responsiveness to the persuasive communications, and the bottom quintile actually exhibited backlash to the persuasive communications equal to the overall average treatment effect. Table 1 illustrates these outcomes.[4] Actual campaign data analysts would construct a continuous responsiveness score, but this example involving quintiles suffices for illustration.

For campaigns with the resources to contact only 20 percent of the electorate, the responsiveness score allows them to create 102,000 votes ($1,700,000 \times 0.02 \times 3 = 102,000$). Without any form of targeting the campaign would generate only 34,000 votes ($1,700,000 \times 0.02 = 34,000$), so using predictive scores increases the number of votes by 200 percent (see Table 1, row 1). A better financed campaign that could contact 40 percent of the electorate and would target the two most promising quintiles of the population. This strategy would yield a total of 170,000 votes, which is a 150 percent increase over having no targeting ($3,400,000 \times 0.02 = 68,000$) (see Table 1, row 2). In this scenario, using predictive scores still improves the campaign's impact, but the gain is less than that of the more resource-constrained campaign. A campaign with the resources to push up against the zero bound where

[4] Backlash is not an uncommon observation among field experiments examining persuasive campaign effects (for example, Arceneaux and Kolodny 2009; Bailey, Hopkins, and Rogers 2013) and among other types of experiments (Nicholson 2012; Hersh and Shaffner 2013).

additional contacts begin to cost the campaign votes would see its efficiency improve by only 50 percent (see Table 1, row 4). This dynamic means that smaller campaigns will benefit most from targeting based on predictive scores, but they are also the campaigns that are least able to afford hiring campaign data analysts and voter databases. Well-financed campaigns benefit from targeting based on predictive scores, but yield smaller relative gains over not using predictive scores for targeting. In this sense, given that small campaigns tend to be less reliant on data analytics, it appears that smaller campaigns may be underinvesting in the development and use of predictive scores.

Again using a fairly generous multiplier regarding responsiveness scores and a baseline 2 percentage point average treatment effect, we can set an upper bound on how the use of such a score might affect campaign outcomes. If there are 8.5 million citizens who will vote in a state (roughly the number of votes cast in the 2012 presidential election in Florida), and a campaign can successfully administer the attempted direct persuasive communications to only half the targeted citizens because of inability to reach all citizens, then a campaign that does not use responsiveness scores would generate 85,000 votes while a campaign that uses responsiveness scores would generate 102,000 votes through direct persuasive communications. While the difference of 17,000 votes is notable, it constitutes only 0.2 percent of the overall vote in this jurisdiction. That said, it would have constituted 23 percent of the 74,309 vote margin of victory for the Obama campaign in Florida in 2012.

Campaigns do not want to mobilize citizens to vote who support their opponent, so one of the most important uses for support scores is to identify which citizens should be targeted during voter mobilization efforts. In an evenly divided electorate, indiscriminately mobilizing citizens would net zero votes—because as many opponents would be mobilized as supporters. In this setting, a naïve comparison of data-based campaigning to absolutely no targeting is not appropriate. Instead, consider a comparison with the following relatively basic targeting strategy that is still employed today in electoral settings that do not have access to predictive scores. Imagine that a campaign attempts to identify individual citizens who support their candidate or issue by directly contacting them in person or over the phone. Imagine that this campaign can successfully reach half of the population and accurately identify their candidate/issue preference. For the remaining half of the population for whom the campaign has not identified a preference, the campaign proceeds to sweep through neighborhoods where more than half of the population supports the campaign's candidate, on the assumption that this approach will lead to a net gain in votes. The only people not targeted in these sweeps are those individuals concretely identified as supporters of the opponent. We can therefore express the expected yield in votes from this targeting strategy as

$$0.5\beta N_j(\%Support_j) \text{ if } \%Support_j < 0.5$$

$$\beta N_j(\%Support) - 0.5\beta N_j(\%Oppose) \text{ if } \%Support_j > 0.5,$$

where β, is the mobilization effect from the campaign, $\%Support_j$ is the level of support for the candidate (a number between 0 and 1) in precinct j, and N_j is the number of registered voters in precinct j.

The first line points out that in precincts where support for the candidate is less than 50 percent, the only effect of this plan will be the direct contacts with supportive voters. However, by assumption the campaign only has the ability to identify half of these people. The second line points out that in areas where support for the candidate is more than 50 percent, the strategy will have two effects. The first is the benefit from mobilizing supporters in the precinct. Unfortunately, the sweep also mobilizes opponents in the proportion to which they are present ($\%Oppose$). However, the campaign managed to identify half of the people supporting the opposition and can choose to avoid these individuals, so the counterproductive mobilization can be cut in half.

We can now contrast this targeting strategy to an imagined predicted-support-score strategy. It would obviously be an unfair comparison to argue that the predicted-support-score strategy worked without error, so we assume that it includes both false positives (misidentifying opponents as supporters) and false negatives (misidentifying supporters as opponents). One can think of these errors as reflecting the political diversity of a given neighborhood. In precincts where the vote is split 50/50, the false positive and false negative error rates are both 15 percent, because these would be the precincts where it is most difficult to infer political beliefs. However, in this hypothetical example the error rate tapers linearly as the precinct becomes more informative of resident beliefs, so that if a precinct unanimously supports one candidate or another, the error rate would obviously be zero. The relationships below presents the formula used in this hypothetical model:

$$\beta N_j[\%Support_j(0.85) - \%Oppose_j(0.15)] \text{ if } \%Support_j = 0.50$$

$$\beta N_j\left[\%Support_j\left(1 - 0.15 \times \frac{\%Support_j}{0.5}\right) - \%Oppose_j \times 0.15\frac{\%Oppose_j}{0.5}\right]$$

$$\text{if } \%Support_j < 0.50$$

$$\beta N_j\left[\%Support_j\left(1 - 0.15 \times \left(1 - \frac{\%Support_j}{0.5}\right)\right) - \%Oppose_j \times 0.15 \times \left(\frac{1 - \%Oppose_j}{0.5}\right)\right]$$

$$\text{if } \%Support_j > 0.50.$$

The equations make clear one underappreciated aspect of predictive modeling; modeling can only increase the efficiency of mobilization efforts. If the outreach from the campaign is not effective (that is, $\beta = 0$), then no votes are generated. Big data analytics may receive media attention, but its effectiveness is entirely reliant on the strength of more traditional aspects of the campaign. If a campaign does not have effective outreach to voters, then predictive analytics cannot solve that problem.

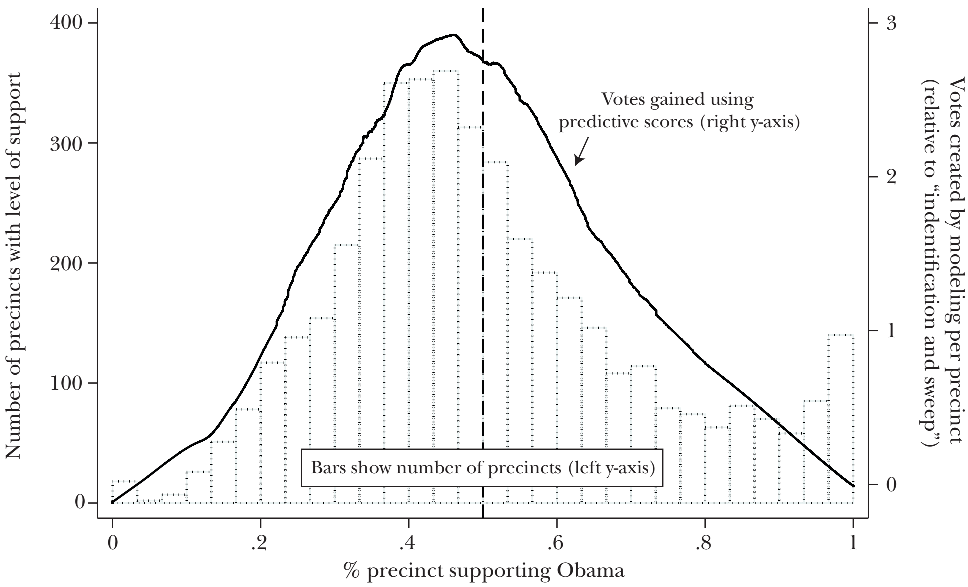
Comparing the traditional strategy of "identification and sweep" to the predictive model, two advantages of the predictive model become clear. First, predictive analytics allows the campaign to target likely supporters in otherwise unfriendly territory. Before accurate prediction was possible, campaigns would leave votes on the table by ignoring supporters living in opponent strongholds. Given the expense of actually identifying individual voter's preferences and the relatively low yield in terms of identifying supporters, avoiding these areas was not optimal tactically, but it was understandable. Second, precinct sweeps are inefficient because in evenly divided precincts many nonsupporters are also mobilized and thereby decrease the overall effectiveness of mobilization drives. Predictive scores (to the extent they are accurate) can reduce this inefficiency. As a result, conditional on precinct size, the biggest difference between the traditional "identification and sweep" tactic and modeled scores is found in the most evenly divided precincts.

Figure 2 shows the results of a thought experiment if these two tactics had been used in Florida across all 4,354 precincts during the 2012 election. The x-axis depicts the percent of votes cast in favor of President Obama in each precinct, and the left-hand y-axis shows in how many precincts President Obama received that share of the vote. Thus, President Obama received between 0 and 3 percent of the vote in about 20 precincts (the left-most bar) and received between 97 and 100 percent of the vote in 140 precincts (the right-most bar). Now imagine as a hypothetical example that the Obama campaign knows the distribution of its support across precincts before the election and is considering two possible strategies to increase its vote: the old-style "identification and sweep" combination of direct contact and precinct targeting, or the method using prediction scores. The solid line, measured on the right y-axis, shows the difference in the number of votes generated from these two approaches. The biggest difference between the two strategies takes place in the middle of the distribution where precincts are most evenly split.[5] The reason for this is clear when the tails are considered. In areas where support for Obama was low, there were not many Obama supporters to mobilize. In the areas where support for Obama was high, there were many supporters to mobilize, but both targeting strategies would target these citizens and neither would mistakenly mobilize those who support the opposing campaign's candidate. It is in areas where the precinct-level data is not predictive of which candidate the citizens support where predictive scores at the individual-level yield the greatest value—even given the inevitably higher number of false positives and false negatives in these precincts.

Using these assumptions, we can gain a rough sense of the impact of the Obama 2012 mobilization effort in Florida using the predictive scores for

[5] If the number of registered voters were held constant across precincts, then the point of maximum difference would be at 0.5. However, the precincts where Obama received 42–45 percent of the vote are larger than precincts with an even split, so there are more votes to be harvested just to the left of the 50/50 mark.

Figure 2
Difference between Predictive Scores and Older Campaign Targeting Heuristics



Notes: Figure 2 represents a thought experiment: In Florida during the 2012 presidential election campaign, how would use of predictive scores for targeting compare to a strategy of "identification and sweep." (See text for details.) The x-axis shows the percent of the two-party vote share for Barack Obama in a precinct in the 2012 general election. The height of dotted bars, read off the left y-axis, report the number of precincts with a given level of support for Obama. The height of the solid line, read off the right y-axis, reports the hypothesized difference between the use of predictive scores for targeting and the use of "identification and sweep." β is assumed to be 0.01. The distribution of precinct data comes from all 4,354 precincts in the 2012 presidential election in Florida.

targeting (which was the strategy the campaign reportedly employed) compared to a precinct-based targeting strategy. Assuming the campaign had a 1 percentage point effect on turnout among the half of the citizens that it targeted for mobilization and successfully contacted, we estimate that it would have generated 8,525 more votes in Florida targeting based on predictive scores relative to targeting based on precinct. This vote total would have been decisive in the 2000 election between George W. Bush and Al Gore, and still constitutes 11 percent of the 74,309 vote margin of victory Barack Obama enjoyed in that state in 2012. Combined with the persuasion analysis above, this thumbnail sketch makes an argument that the 2012 Obama re-election would have been closer in key states had it used the older and coarser targeting technologies, rather than the predictive scores produced by its campaign data analysts.

Conclusion: Some Thoughts on Coordination

Sophisticated campaigns develop and use voter databases that contain a range of detailed information on individual citizens. As a result, campaign data analysts occupy an increasingly important role in politics. They develop predictive models that produce individual-level scores that predict citizens' likelihoods of performing certain political behaviors, supporting candidates and issues, and responding to targeted interventions. The use of these scores has increased dramatically during the last few election cycles. Simulations suggest that these advances could yield sizable and electorally meaningful gains to campaigns that harness them.

Since predictive scores make campaigns more effective and efficient by increasing the cost effectiveness of communicating with citizens, a broad range of organizations do and will employ the technologies. To the extent that predictive scores are useful and reveal true unobserved characteristics about citizens, it means that multiple organizations will produce predictive scores that recommend targeting the same sets of citizens. For example, some citizens might find themselves contacted many times, while other citizens—like those with low turnout behavior scores in 2012—might be ignored by nearly every campaign. The marginal effect of the fifth or sixth contact from a campaign will be less than the marginal effect of the first contact from a campaign. Thus, concentrating attention on the same set of citizens due to widespread adoption of predictive scores may offset some of the gains reaped from developing predictive scores in the first place. In this way, developing and using predictive scores creates a coordination game in which allied organizations would prefer to partition the electorate and not to duplicate efforts.

Coordination could theoretically happen between partisan organizations, like state parties, candidate campaigns, and coordinated campaigns, and across nonpartisan activities, like civil rights groups, labor unions, and environmental groups. However, partisan and nonpartisan organizations are not allowed to coordinate their electoral activities. Since it is nearly impossible to observe whom campaigns target for direct communications—that is, direct mail, knock on doors, and making phone calls—this coordination game has incomplete information, which means that inefficiencies from overlapping contacts are inevitable.

Even when coordination is allowed by law, coalitions may have conflicting incentives. There is enough regional variation in ideology that it is possible for local candidates to appeal to citizens who oppose the national candidate. For instance, local Republicans mobilizing citizens in liberal districts could have hurt Mitt Romney, and local Democrats mobilizing citizens in conservative districts could have hurt Barack Obama in 2012. The same dynamic plays out among nonpartisan groups as well. While labor union members and environmentalists agree on many policies and values, it is likely that some members do not hold that same views on both labor and environmental issues. In states like West Virginia, where the local coal industry is considered "dirty" by environmentalists, the groups could be working at cross-purposes, both with regards to messaging and targeting. Thus, mobilizing a set of citizens for a labor-related ballot initiative might result in less support for an environmentally friendly candidate. This tension is endemic to the very nature of the federal system of representation and coalition politics. The tension has always been present, but now that groups can share very detailed targeting plans and support scores, the tension can and will bubble to the surface more often than in the past.

The improved capability to target individual voters offers campaigns an opportunity to concentrate their resources where they will be most effective. This power, however, has not radically transformed the nature of campaign work. One could argue that the growing impact of data analytics in campaigns has amplified the importance of traditional campaign work. Message polling (that is, polls designed to gauge voter reactions to different campaign messages) no longer solely dictates targeting, but the increased demand for information during the campaign has increased the amount of polling used to generate snapshots of the electorate. Professional phone interviews are still used for message development and tracking, but they are also essential for developing predictive scores of candidate support and measuring changes in voter preferences in randomized experiments. Similarly, better targeting has made grassroots campaign tactics more efficient and therefore more cost competitive with mass communication forms of outreach. Volunteers still need to persuade skeptical neighbors, but they are now better able to focus on persuadable neighbors and use messages more likely to resonate. This leads to higher-quality interactions and (potentially) a more pleasant volunteer experience. So while savvy campaigns will harness the power of predictive scores, the scores will only help the campaigns that were already effective.

References

Ansolabehere, Stephen, and Eitan Hersh. 2010. "The Quality of Voter Registration Records: A State-by-State Analysis." Working paper, Institute for Quantitative Social Science, Harvard University.

Ansolabehere, Stephen, and Eitan Hersh. 2012. "Validation: What Big Data Reveal about Survey Misreporting and the Real Electorate." *Political Analysis* 20(4): 437–59.

Arceneaux, Kevin, and Robin Kolodny. 2009. "Educating the Least Informed: Group Endorsements in a Grassroots Campaign." *American Journal of Political Science* 53(4): 755–70.

Arceneaux, Kevin, and David W. Nickerson. 2009. "Who is Mobilized to Vote? A Re-Analysis of 11 Randomized Field Experiments." *American Journal of Political Science* 53(1): 1–16.

Arceneaux, Kevin, and David W. Nickerson. 2010. "Comparing Negative and Positive Campaign Messages: Evidence from Two Field Experiments." *American Politics Research* 38(1): 54–83.

Bailey, Michael A., Daniel J. Hopkins, and Todd Rogers. 2013. "Unresponsive and Unpersuaded: The Unintended Consequences of Voter Persuasion Efforts." Available at SSRN: http://ssrn.com/abstract=2307631.

BigData-Startups. 2013. "Big Data during the Obama Campaign." http://www.bigdata-startups.com/BigData-startup/big-data-obama-campaign/. Accessed February 18, 2014.

Breiman, Leo. 2001. "Random Forests." *Machine Learning* 45(1): 5–32.

Breiman, Leo, Jerome Friedman, Charles J. Stone, and R. A. Olshen. 1984. *Classification and Regression Trees.* Boca Raton, FL: CRC Press.

Bryan, Christopher J., Gregory M. Walton, Todd Rogers, and Carol S. Dweck. 2011. "Motivating Voter Turnout by Invoking the Self." *PNAS* 108(31): 12653–156.

Burt, Gabriel. 2013. "My Time at 'Obama for America.'" *Tech in Motion,* August 27, 2013. http://www.techinmotionevents.com/blog/post/2013/08/27/my-time-at-obama-for-america.

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

Dietterich, Tom. 1995. "Overfitting and Undercomputing in Machine Learning." *ACM Computing Surveys* 27(3): 326–27.

Duhigg, Charles. 2012. "Campaigns Mine Personal Lives to Get Out Vote." *New York Times,* October 13, page A1.

Gallagher, Sean. 2012. "Built to Win: Deep inside Obama's Campaign Tech." Arstechnica.com, November 14, 2012. http://arstechnica.com/information-technology/2012/11/built-to-win-deep-inside-obamas-campaign-tech/.

Gan, Guojun, Chaoqun Ma, and Jianhong Wu. 2007. *Data Clustering: Theory, Algorithms, and Applications.* Alexandria, VA: SIAM, Society for Industrial and Applied Mathematics.

Gerber, Alan S., and Donald P. Green. 2000. "The Effects of Canvassing, Direct Mail, and Telephone Contact on Voter Turnout: A Field Experiment." *American Political Science Review* 94(3): 653–63.

Gerber, Alan S., and Todd Rogers. 2009. "Descriptive Social Norms and Motivation to Vote: Everybody's Voting and So Should You." *Journal of Politics* 71(1): 1–14.

Green, Donald, and Alan S. Gerber. 2008. *Get Out the Vote! How to Increase Voter Turnout.* Washington, DC: Brookings Institution Press.

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.

Haberman, Maggie. 2013. "Obama '12 Vets to Share Some, Not All, Data." *Politico,* November 20, 2013. http://www.politico.com/story/2013/11/barack-obama-2012-campaign-data-100133.html.

Hersh, Eitan D., and Brian F. Schaffner. 2013. "Targeted Campaign Appeals and the Value of Ambiguity." *Journal of Politics* 75(2): 520–34.

Hillygus, D. Sunshine, and Todd G. Shields. 2008. *The Persuadable Voter: Wedge Issues in Presidential Campaigns.* Princeton, NJ: Princeton University Press.

Issenberg, Sasha. 2010. "Nudge the Vote: How Behavioral Science is Remaking Politics." *New York Times Sunday Magazine*, October 31, page MM28.

Issenberg, Sasha. 2012a. *The Victory Lab: The Secret Science of Winning Campaigns.* New York, NY: Crown Publishers.

Issenberg, Sasha. 2012b. "The Death of the Hunch." *Slate,* May 22. http://www.slate.com/articles/news_and_politics/victory_lab/2012/05/obama_campaign_ads_how_the_analyst_institute_is_helping_him_hone_his_message_.html.

Issenberg, Sasha. 2012c. "How President Obama's Campaign Used Big Data to Rally Individual Voters, Part 2." *MIT Technology Review,* December 17. http://www.technologyreview.com/featuredstory/508851/how-obama-wrangled-data-to-win-his-second-term/.

Kuhn, David Paul. 2004. "Vacation Politics." CBSNews.com, December 5, 2007. http://www.cbsnews.com/2100-250_162-607234.html.

Madrigal, Alexis C. 2012. "When the Nerds Go Marching In." TheAtlantic.com, November 16. http://www.theatlantic.com/technology/archive/2012/11/when-the-nerds-go-marching-in/265325.

Malchow, Hal. 2003. *The New Political Targeting.* Washington, DC: Campaigns and Elections Magazine.

McAuliffe, Terry, with Steve Kettman. 2007. *What A Party!: My Life Among Democrats: Presidents, Candidates, Donors, Activists, Alligators, and Other Wild Animals.* New York, NY: St. Martin's Press.

Nicholson, Steven P. 2012. "Polarizing Cues." *American Journal of Political Science* 56(1): 52–66.

Nickerson, David W. 2005. "Partisan Mobilization Using Volunteer Phone Banks and Door Hangers." *Annals of the American Academy of Political and Social Science* 601:10–27.

Nickerson, David W. 2006a. "Hunting the Elusive Young Voter." *Journal of Political Marketing* 5(3): 47–69.

Nickerson, David W. 2006b. "Volunteer Phone Calls Can Increase Turnout: Evidence from Eight Field Experiments." *American Politics Research* 34(3): 271–92.

Nickerson, David W. 2007a. "Quality is Job One: Professional and Volunteer Mobilization Calls." *American Journal of Political Science* 51(2): 269–82.

Nickerson, David W. 2007b. "Don't Talk to Strangers: Experimental Evidence of the Need for Targeting." Paper presented at the Annual Meeting of the Midwest Political Science Association, Chicago, IL, April 12–15.

Nickerson, David W., Ryan F. Friedrichs, and David C. King. 2006. "Partisan Mobilization Campaigns in the Field: Results from a Statewide Turnout Experiment in Michigan." *Political Research Quarterly* 59(1): 85–97.

Nickerson, David, and Todd Rogers. 2010. "Do You Have a Voting Plan?: Implementation

Intentions, Voter Turnout, and Organic Plan Making." *Psychological Science* 21(2): 194–99.

Rogers, Todd, and Masa Aida. 2013. "Vote Self-Prediction Hardly Predicts Who Will Vote, And Is (Misleadingly) Unbiased." *American Politics Research*, published online before print, September 5.

Rogers, Todd, and David W. Nickerson. 2013. "Can Inaccurate Beliefs about Incumbents Be Changed? And Can Reframing Change Votes?" HKS Faculty Research Working Paper Series, RWP13-018.

Scherer, Michael. 2012. "Inside the Secret World of Quants and Data Crunchers Who Helped Obama Win." *Time*, November 19, p. 56–60.

Tam Cho, Wendy K., and James G. Gimpel. 2007. "Prospecting for (Campaign) Gold." *American Journal of Political Science* 51(2): 255–68.

Vargas, Jose Antonio. 2008. "Obama Raised Half a Billion Online." *Washingtonpost.com,* November 20. http://voices.washingtonpost.com/44/2008/11/20/obama_raised_half_a_billion_on.html.

Journal of Economic Perspectives—Volume 28, Number 2—Spring 2014—Pages 75–98

Privacy and Data-Based Research

Ori Heffetz and Katrina Ligett

On August 9, 2006, the "Technology" section of the *New York Times* contained a news item titled "A Face Is Exposed for AOL Searcher No. 4417749," in which reporters Michael Barbaro and Tom Zeller (2006) tell a story about big data and privacy:

> Buried in a list of 20 million Web search queries collected by AOL and recently released on the Internet is user No. 4417749. The number was assigned by the company to protect the searcher's anonymity, but it was not much of a shield. No. 4417749 conducted hundreds of searches over a three-month period on topics ranging from "numb fingers" to "60 single men" to "dog that urinates on everything." And search by search, click by click, the identity of AOL user No. 4417749 became easier to discern. There are queries for "landscapers in Lilburn, Ga," several people with the last name Arnold and "homes sold in shadow lake subdivision gwinnett county georgia." It did not take much investigating to follow that data trail to Thelma Arnold, a 62-year-old widow who lives in Lilburn, Ga. . . . Ms. Arnold, who agreed to discuss her searches with a reporter, said she was shocked to hear that AOL had saved and published three months' worth of them. "My goodness, it's my whole personal life," she said. "I had no idea somebody was looking over my shoulder." . . ."We all have a right to privacy," she said. "Nobody should have found this all out."

■ *Ori Heffetz is Assistant Professor of Economics, Samuel Curtis Johnson Graduate School of Management, Cornell University, Ithaca, New York. Katrina Ligett is Assistant Professor of Computer Science and Economics, California Institute of Technology, Pasadena, California. Their email addresses are oh33@cornell.edu and katrina@caltech.edu.*

http://dx.doi.org/10.1257/jep.28.2.75 doi=10.1257/jep.28.2.75

Empirical economists are increasingly users, and even producers, of large datasets with potentially sensitive information. Some researchers have for decades handled such data (for example, certain kinds of Census data), and routinely think and write about privacy. Many others, however, are not accustomed to think about privacy, perhaps because their research traditionally relies on already-publicly-available data, or because they gather their data through relatively small, "mostly harmless" surveys and experiments. This ignorant bliss may not last long; detailed data of unprecedented quantity and accessibility are now ubiquitous. Common examples include a private database from an Internet company, data from a field experiment on massive groups of unsuspecting subjects, and confidential administrative records in digital form from a government agency. The AOL story above is from 2006; our ability to track, store, and analyze data has since then dramatically improved. While big data become difficult to avoid, getting privacy right is far from easy—even for data scientists.

This paper aims to encourage data-based researchers to think more about issues such as privacy and anonymity. Many of us routinely promise anonymity to the subjects who participate in our studies, either directly through informed consent procedures, or indirectly through our correspondence with Institutional Review Boards. But what is the informational content of such promises? Given that our goal is, ultimately, to publish the results of our research—formally, to publish functions of the data—under what circumstances, and to what extent, can we guarantee that individuals' privacy will not be breached and their anonymity will not be compromised?

These questions may be particularly relevant in a big data context, where there may be a risk of more harm due to both the often-sensitive content and the vastly larger numbers of people affected. As we discuss below, it is also in a big data context that privacy guarantees of the sort we consider may be most effective.

Our paper proceeds in three steps. First, we retell the stories of several privacy debacles that often serve as motivating examples in work on privacy. The first three stories concern intentional releases of de-identified data for research purposes. The fourth story illustrates how individuals' privacy could be breached even when the data themselves are not released, but only a seemingly innocuous function of personal data is visible to outsiders. None of our stories involves *security* horrors such as stolen data, broken locks and passwords, or compromised secure connections. Rather, in all of them information was released that had been *thought* to have been anonymized, but, as was soon pointed out, was rather revealing.

Second, we shift gears and discuss *differential privacy*, a rigorous, portable privacy notion introduced roughly a decade ago by computer scientists aiming to enable the release of information while providing *provable* privacy guarantees. At the heart of this concept is the idea that the addition or removal of a single individual from a dataset should have nearly no effect on any publicly released functions of the data, but achieving this goal requires introducing randomness into the released outcome. We discuss simple applications, highlighting a privacy-accuracy tension: randomness leads to more privacy, but less accuracy.

Third, we offer lessons and reflections, discuss some limitations, and briefly mention additional applications. We conclude with reflections on current promises of "anonymity" to study participants—promises that, given common practices in empirical research, are not guaranteed to be kept. We invite researchers to consider either backing such promises with meaningful privacy-preserving techniques, or qualifying them. While we are not aware of major privacy debacles in economics research to date, the stakes are only getting higher.

Intuition May Not Be Enough: Cautionary Tales

Well-intentioned government or private entities in possession of a sensitive database may wish to make an anonymized version of the data public—for example, to facilitate research. We retell and discuss a few cautionary tales that illustrate how intuition-based attempts at anonymization may fail, sometimes spectacularly.[1]

When Anonymization Failed

The first story is from the mid 1990s, when William Weld, then Governor of Massachusetts, approved the release of certain medical records of state employees to researchers, assuring the public that individual anonymity would be protected by eliminating obvious identifiers from the data (Greely 2007). A few days after Weld's announcement, Latanya Sweeney—then a graduate student at MIT—re-identified Weld's personal records (including diagnoses and prescriptions) in the database; she then had his records delivered to his office.

While the medical data—officially, the Massachusetts "Group Insurance Commission" (GIC) data—had been "de-identified" by removing fields containing patients' name, address, and social security number (SSN) prior to the data release, the nearly 100 remaining fields included ZIP code, birth date, and sex. As Ohm (2010) tells the story, Sweeney

> . . . knew that Governor Weld resided in Cambridge, Massachusetts, a city of 54,000 residents and seven ZIP codes. For twenty dollars, she purchased the complete voter rolls from the city of Cambridge—a database containing, among other things, the name, address, ZIP code, birth date, and sex of every voter. By combining this data with the GIC records, Sweeney found Governor Weld with ease. Only six people in Cambridge shared his birth date; only three were men, and of the three, only he lived in his ZIP code.

Barth-Jones (2012) revisits and critiques this story. Perhaps in response, Sweeney, Abu, and Winn (2013) use a similar method to re-identify individuals in the publicly

[1] As hinted above, these stories are well known in the computer science community that studies privacy. The first three were revisited and discussed by Ohm (2010), a legal scholar, who provides further references and links to primary sources.

available Personal Genome Project database. Sweeney's "How Unique Are You?" interactive website invites the visitor to "Enter your ZIP code, date of birth, and gender to see how unique you are (and therefore how easy it is to identify you from these values)." Her methodology is explained on the website (http://aboutmyinfo.net, accessed on August 9, 2013):

> Birthdate . . . gender, and 5-digit postal code (ZIP) uniquely identifies most people in the United States. Surprised? . . . 365 days in a year × 100 years × 2 genders = 73,000 unique combinations, and because most postal codes have fewer people, the surprise fades. . . . [T]here are more than 32,000 5-digit ZIP codes in the United States; so 73,000 × 32,000 is more than 2 billion possible combinations but there are only 310 million people in the United States.

The next story, involving Ms. Arnold above, is from roughly a decade later. In 2006, AOL Research released detailed Internet search records of 650,000 users covering a three-month period, amounting to 20 million search queries.[2] The stated purpose of the release was expressed by then AOL Research head Abdur Chowdhury:

> AOL is embarking on a new direction for its business—making its content and products freely available to all consumers. To support those goals, AOL is also embracing the vision of an open research community, which is creating opportunities for researchers in academia and industry alike. . . . with the goal of facilitating closer collaboration between AOL and anyone with a desire to work on interesting problems.[3]

Prior to the data release, the search logs were de-identified, for example by removing usernames and IP addresses, using instead unique identifiers (such as "4417749") to link all of a single user's queries. This de-identification, however, quickly proved far from sufficient for the intended anonymization, as illustrated by the *New York Times* article on Ms. Arnold. Within days of the release, AOL apologized, removed the data website as well as a few employees, and silenced its research division. Of course, to this day, the data are widely available through a simple web search; once published, you cannot take it back.

The third story is also from 2006. About two months after the AOL debacle, Netflix announced a competition—the Netflix Prize—for improving the company's algorithm that predicts user ratings of films, using only past user ratings. To allow competitors to train their algorithms, Netflix released a database with 100 million ratings of 17,770 films by about 500,000 subscribers covering a six-year period.

[2] As Ohm (2010) notes, different numbers appear in different accounts. The 650,000 figure above was described as 500,000 in the original post, and the 20 million figure in the original post has later been reported by some as 36 million.

[3] Posting of Abdur Chowdhury, cabdur@aol.com, to SIGIR-IRList, irlist-editor@acm.org, http://sifaka.cs.uiuc.edu/xshen/aol/20060803_SIG-IRListEmail.txt, as cited in Ohm (2010, accessed on August 9, 2013).

Each record contained a movie title, a rating date, and a five-point rating. As in the Massachusetts Group Insurance Commission and AOL cases, records were de-identified prior to the release, replacing user names with unique identifiers.

The illusion of protecting users' anonymity was, again, short-lived. Two weeks after the data release, Narayanan and Shmatikov (2008; first version posted in 2006) demonstrated that "an adversary who knows a little bit about some subscriber can easily identify her record if it is present in the dataset, or, at the very least, identify a small set of records which include the subscriber's record." How little is "a little bit"? In many cases, a user could be identified knowing as little as that user's approximate dates and ratings of two or three movies. In their demonstration, Narayanan and Shmatikov used ratings from the Internet Movie Database (IMDB), which are publicly available and are linked to the raters' identities, and showed how a handful of a user's IMDB ratings, even when they yield imprecise information, could uniquely identify that user in the Netflix database.

Whereas IMDB's public ratings may reveal only those movies that individuals are willing to tell the world that they have watched, Netflix ratings may reveal *all* of the movies an individual has rated, including those the rater may prefer to keep private—for example, films that may reflect a person's sexual, social, political, or religious preferences. Moreover, to be re-identified, one does not have to be on IMDB: as Ohm (2010) advises his readers, "the next time your dinner party host asks you to list your six favorite obscure movies, unless you want everybody at the table to know every movie you have ever rated on Netflix, say nothing at all."

De-identification and Beyond

De-identified data were defined by Sweeney (1997) as data in which "all explicit identifiers, such as SSN (Social Security number), name, address, and telephone number, are removed, generalized, or replaced with a made-up alternative." Her definition seems to describe accurately the released Massachusetts health insurance, AOL, and Netflix data in the stories above. Some more recent definitions (like those under federal health records privacy regulations) are stricter and would not consider the Massachusetts data released by Weld as de-identified, but these definitions still focus on removing only specific kinds of information (Greely 2007). Indeed, more than 15 years after Sweeney's powerful demonstration, her definition of de-identified data *still* describes, more or less accurately, commonplace practices among many researchers. For example, prior to publicly posting their data online (as required by some academic journals), economists often de-identify their data by merely withholding explicit identifiers such as subject names. However, as in the stories above, the stated aim of such de-identification—and what is often promised to subjects, directly or via an Institutional Review Board—is *anonymization*. In Sweeney's (1997) definition, *anonymous data* "cannot be manipulated or linked to identify an individual." Clearly, de-identification does not guarantee anonymization.

Sweeney's re-identification of people in the Massachusetts Group Insurance Commission data used birthday and five-digit ZIP code, neither of which are typically included in datasets publicly posted by economists. But it is not difficult to imagine re-identification of specific subjects based on combinations of demographics such as study major, age/class, gender, and race, which are often not considered "identifiable private information" and are routinely included in posted data.[4] Re-identification is still easier with knowledge regarding, for example, the day and time in which a classmate or a roommate participated in a specific study session. (Sweeney, 2013, applies this idea outside the lab: she uses newspaper stories that contain the word "hospitalized" to re-identify individual patients in a publicly available health dataset in Washington state.) But re-identification is possible even without such special knowledge, and it may be straightforward when targeting certain individuals who have a characteristic that is uncommon in a specific setting, such as minorities or women in certain fields and occupations.

This discussion highlights a weakness of de-identification: if one assumes no restrictions on outside information (also referred to below as auxiliary information), then, short of removing *all* data fields prior to a release, some individuals may be uniquely identified by the remaining fields. One potential response to this weakness is an approach called *k-anonymity*, which combines the assumption that there are *some* restrictions on outside information with the removal (or partial removal) of *some* fields. Specifically, assuming that outside information could only cover certain fields in the database, one could suppress these fields or, when possible, generalize them (for example, replace date of birth with year of birth) so that any combination of the values reported in these fields would correspond to at least k individuals in the data (Sweeney 2002). This approach has several weaknesses, and in many applications it implies either an unreasonably weak privacy guarantee or a massive suppression of data: notice that the amount of information that can be released is expected to shrink as k grows and as restrictions on outside information are weakened. Narayanan and Shmatikov (2008), for example, discuss the issues with such an approach in the Netflix context.

An alternative approach is to make it harder for an attacker to leverage outside information. For example, prior to making the query logs publicly available, AOL could have replaced not only user identities but also the search keywords themselves with uniquely identifying random strings. Similarly, Netflix could have replaced movie names with unique identifiers. Such an approach, known as "token-based hashing," would preserve many features of the data, hence maintaining usefulness of

[4] For example, according to Cornell's Office of Research Integrity and Assurance (at http://www.irb.cornell.edu, accessed on August 13, 2013):

> Identifiable private information is defined as: name; address; elements of dates related to an individual (e.g., birth date); email address; numbers: telephone, fax, social security, medical record, health beneficiary/health insurance, certificate or license numbers, vehicle, account numbers (e.g., credit card), device identification numbers, serial numbers, any unique identifying numbers, characteristics, or codes (e.g., Global Positioning System (GPS) readings); Web URLs; Internet Protocol (IP) addresses; biometric identifiers (e.g., voice, fingerprints); full face photographs or comparable images.

the database for some (though clearly not all) research purposes. But the preserved features of the underlying data make this type of scheme vulnerable as well.

Indeed, shortly after the disaster at AOL Research, a group at Yahoo! Research (Kumar, Novak, Pang, and Tomkins 2007) showed that an attacker with access to a "reference" query log (for example, early logs released by Excite or Altavista) could use it to extract statistical properties of tokenized words in the database, and "invert the hash function"—that is, break the coding scheme—based on co-occurrences of tokens within searches. Along similar lines, Narayanan and Shmatikov (2008) speculate that in the Netflix case, such an approach "does not appear to make de-anonymization impossible, but merely harder."

Privacy Risk without Data Release

Our fourth story, of privacy compromised on Facebook by Korolova (2011), "illustrates how a real-world system designed with an intention to protect privacy but without rigorous privacy guarantees can leak private information . . . Furthermore, it shows that user privacy may be breached not only as a result of data publishing using improper anonymization techniques, but also as a result of internal data-mining of that data."

Facebook's advertising system allows advertisers to specify characteristics of individuals to whom an ad should be shown. At the time of Korolova's (2011) attack, it was possible to specify those characteristics (for example, gender, age, location, workplace, alma mater) so finely that they would correspond to a unique Facebook user. Then, two versions of the ad campaign could be run—for example, one with those same characteristics plus "Interested in women"; the other with those characteristics plus "Interested in men." Even if this user's interests were not visible to her friends, if she had entered them in her profile, they would be used for ad targeting. Thus, if the advertiser received a report that, for example, the "Interested in women" version of her ad had been displayed, the advertiser could infer the targeted individual's private interests. Other attacks were possible too. "Using the microtargeting capability, one can estimate the frequency of a particular person's Facebook usage, determine whether they have logged in to the site on a particular day, or infer the times of day during which a user tends to browse Facebook."

Korolova (2011) quotes failed promises by Facebook executives, such as that Facebook doesn't "share your personal information with services you don't want" and doesn't "give advertisers access to your personal information." She notes: "We communicated our findings to Facebook on July 13, 2010, and received a very prompt response. On July 20, 2010, Facebook launched a change to their advertising system that made the kind of attacks we describe much more difficult to implement in practice, even though, as we discuss, they remain possible in principle."

This Facebook story helps demonstrate that if one seeks to use functions of data—be it via research findings, policy decisions, or commercial services and products—the privacy of the individuals comprising the data may be at risk without an approach providing (provable) privacy guarantees.

Differential Privacy

A common theme in the examples above has been the crucial role played by *auxiliary information*, that is, knowledge from sources outside the dataset under consideration. In the examples above, attackers consulted various outside sources not foreseen by the database owners, including public records such as voter rolls, complementary databases such as IMDB, or, simply, personal familiarity with an individual in the database. To identify individuals, the attackers then carried out a variant of a so-called "linkage attack": they matched fields that overlap across the auxiliary data and the attacked database.

More generally, one may invite trouble when making specific assumptions regarding what information a potential attacker might have and how the attacker might use it. If such assumptions are ever violated—even in the future, as new technology and information become available—privacy may be compromised. One approach to addressing the auxiliary-information concern would be to seek to provide privacy guarantees free from such assumptions. The approach we discuss here, *differential privacy*, seeks to do just that. It emerged from work in computer science theory by Dinur and Nissim (2003), Dwork and Nissim (2004), and Dwork, McSherry, Nissim, and Smith (2006). Our discussion and examples draw on a number of surveys, including Dwork (2006), Dwork and Smith (2010), Dwork (2011b, a), and Dwork, McSherry, Nissim, and Smith (2011). These surveys additionally present historical aspects of the development of the differential privacy definition, more examples, and a much broader range of applications than we discuss here. Our working paper, Heffetz and Ligett (2013), contains slightly more technical detail than presented here as well as more references to recent work on differential privacy. We also recommend a recent popular article on differential privacy research by Klarreich (2012).

The Differential Privacy Definition

To fix ideas, consider the released outcome of some function of a database: for example, the released number of Facebook users to whom an ad was displayed, or some published table of statistics in an empirical research paper, or even a released version of the entire database. Consider a potential participant in the database: for example, someone who considers joining Facebook, or someone who considers participating in a research study. Compare two possible scenarios: in one, this person joins and is added to the database; in the other, the person does not join and hence is not in the database.

Informally, differential privacy seeks to guarantee to the potential participant that, irrespective of the decision whether to participate, *almost* the same things can be learned from the released outcome—regardless of outside information, of the data already in the database, or of the participant's own personal data. Differential privacy hence gives participants (and nonparticipants) in the database a form of plausible deniability: they could always deny that their data took specific values or even that they participated (or did not participate), and an observer would have almost no evidence either way.

Here is an often-used example: one could conduct a differentially-private analysis that revealed a correlation between smoking and cancer, so long as that correlation depended only negligibly on the participation of any one individual in the study. Revealing (that is, publishing) this correlation might allow observers to draw inferences about an individual smoker, and that person might then feel that his or her privacy has been harmed. But since essentially the same conclusions would have been drawn regardless of whether that smoker participated in the study, the *differential* privacy of that person has been respected.

In a more formal sense, consider pairs of databases that are identical except that one of the databases has one additional row (or record) over the other. We refer to such a pair as *neighboring* databases, and think of each row as corresponding to one individual. Thus, two neighboring databases differ by only the participation of one individual. Now consider some computation that is carried out on such databases, and consider the space of possible outcomes of the computation. A differentially-private computation (or function, or mechanism) selects its output using a degree of randomness, such that the probability of any given outcome is similar under any two neighboring databases.

How similar? A common differential privacy definition, ϵ-*differential privacy* (Dwork, McSherry, Nissim, and Smith 2006), requires that the probability of any given outcome under any two neighboring databases cannot differ by more than a multiplicative constant, e^{ϵ}, where e is Euler's number and the parameter ϵ is a positive number that quantifies the amount of privacy.[5] The smaller ϵ is, the stronger is the privacy guarantee, but the less useful is the computation's output: in the limiting case of $\epsilon = 0$, we would replace the word "similar" above with "identical" since in that limiting case, e^{ϵ} would equal 1, requiring that the differentially-private mechanism be indistinguishable on any two input databases. In other words, maximum differential privacy means useless published output. More generally, the definition makes precise an intuitive tradeoff between privacy and usefulness.

The output of a differentially-private mechanism is readily publishable. It could, for example, be a single statistic (or a collection of statistics) to which a sufficient amount of random noise was added so that the inclusion of an additional record in, or exclusion of an existing record from, the database would have

[5] Here is a formal definition:

A randomized function K provides ϵ-differential privacy if for every $\mathcal{S} \in \mathrm{Range}(K)$ and for all neighboring databases D and D',

$$\mathrm{Prob}[K(D) = \mathcal{S}] \leq e^{\epsilon} \cdot \mathrm{Prob}[K(D') = \mathcal{S}]$$

for $\epsilon \geq 0$ and where the probability space in each case is over the randomness of K.

Note that in particular, for any neighboring pair (D, D'), the definition must hold with the larger quantity (that is, $\max\{\mathrm{Prob}[K(D) = \mathcal{S}],\ \mathrm{Prob}[K(D') = \mathcal{S}]\}$) on the left, constraining it to be larger by at most a multiplicative e^{ϵ}. There are other variants on this definition, which we do not emphasize here. A common generalization of differential privacy allows an *additive* δ difference in the probabilities, in addition to the mulitiplicative difference e^{ϵ} (for example, Dwork, Kenthapadi, McSherry, Mironov, and Naor 2006). Such generalization provides a weaker privacy guarantee, but may allow for more accurate outcomes.

almost no effect on the distribution of the statistic (or statistics). Or it could be an entire *synthetic* database—a database consisting of artificial records, created with a degree of randomness from the original records in a way that preserves certain statistical properties of the original database but does not give away the inclusion of any individual record. The following subsections will have more to say about ways of using randomness and how much randomness is necessary. As discussed above, notice again that when we write here "would have almost no effect" and "does not give away" we imply that ϵ is small. How small should it be? The definition of differential privacy does not prescribe an answer to this normative question, a point we return to below.

Observations Regarding the Definition

The concept of differential privacy readily extends to provide a privacy guarantee to a group of individuals of a certain size: an ϵ-differentially-private mechanism is $k\epsilon$-differentially private from the point of view of a group of k individuals (or one individual whose data comprise k rows in the database). Intuitively, the inclusion in or exclusion from a database of a *group* of rows could have larger cumulative effect on outcomes of computations, weakening the privacy guarantee. In the smoking-and-cancer example above, it is more difficult to guarantee that adding an entire group of people to the study—say, all the residents of a specific city—would have almost no effect on outcomes.

The differential privacy definition also immediately yields an elegant composition property: running ℓ ϵ-differentially-private mechanisms—for example, publishing ℓ statistics based on a database—gives a guarantee of $\ell\epsilon$-differential privacy.[6] Equivalently, one may split a fixed total "privacy budget" of ϵ across a set of desired computations.

This composition property is particularly important in the context of potential real-world applications—including academic research and public- and private-sector implementations—where individuals may participate in more than one database, and where on each database typically more than one analysis is conducted. Differential privacy hence provides a tool for understanding the cumulative privacy harm incurred by an individual whose data appear in multiple databases, potentially used by different entities for different purposes and at different points in time. One could discuss assessments of individuals' cumulative, lifelong privacy loss, and use them as an input into the discussion of how small ϵ should be in each specific computation. Moreover, some socially desired cap on such cumulative privacy loss could be thought of as an individual's lifetime privacy budget. That budget is then to be carefully allocated, and prudently spent, across computations over one's lifetime to guarantee a desired amount of lifetime privacy.

Finally, it can be shown that differential privacy guarantees hold up under post-processing of their outputs: if one conducts an ϵ-differentially-private

[6] More generally, running any ℓ differentially-private mechanisms with guarantees $\epsilon_1, \ldots, \epsilon_\ell$ gives $\left(\sum_{i=1}^{\ell} \epsilon_i\right)$-differential privacy.

computation, one is then free to perform any subsequent computation on the output of that computation, and the result will still be ϵ-differentially private. In other words, once one has produced differentially-private statistics on a dataset, those statistics can be made public for all eternity, without concern that at some later date a clever hacker will find some new privacy-revealing weakness.

From a Bayesian point of view, differential privacy can be given the following interpretation: an observer with access to the output of a differentially-private function should draw almost the same conclusions whether or not one individual's data are included in the analyzed database, regardless of the observer's prior. This interpretation highlights that differential privacy is a property of the function (the mapping from databases into outcomes), not of the output (a particular outcome). Kasiviswanathan and Smith (2008) credit Cynthia Dwork and Frank McSherry with the first formulation of this interpretation, which can be formalized and proven equivalent to differential privacy.

The Bayesian "observer" may of course refer to anyone with access to the output of the function, including malicious attackers, (legitimate) advertisers on Facebook, or the readers of a research paper that reports some statistic. Notice that this Bayesian interpretation does not rule out performing analyses and reporting outcomes that vastly alter the observer's posterior view of the world, so long as the outcomes are not very sensitive to the presence or absence of any one individual in the original database. Our example above, regarding a differentially-private analysis that revealed a correlation between smoking and cancer, illustrates this point.

From Definition to Application: Noise and Sensitivity

With the concept of differential privacy in hand, consider the computation (and subsequent release) of the mean income of individuals in a database. While the mean might seem like a fairly innocuous statistic, all statistics reveal *something* about the data, and in certain worst-case situations, the mean might be quite revealing. For example, if the mean salary in a certain economics department prior to hiring a new faculty member is known to an observer (for instance, due to a previous release), then releasing the new mean after the hire reveals the new hire's salary. This is a variant of the so-called "differencing attack."

A simple technique for guaranteeing differential privacy is to add randomly generated noise to the true mean prior to its release. How much noise? Since differential privacy is a worst-case guarantee over all possible pairs of neighboring databases and over all possible outcomes, if the distribution of incomes is not a priori bounded, the noise would have to be unboundedly large (to guarantee that even the addition of an extreme outlier to the database would have little effect on the differentially-private statistic). With a limit on the range of incomes, however, one could add a limited amount of noise to the true mean in order to guarantee differential privacy.

More formally, when a function that we wish to compute on a database returns a real number, we say that the *sensitivity* of that function, denoted Δf, is the largest possible difference (in absolute value) between the two outputs one might get when applying that function to two neighboring databases.[7] The definition makes it clear that sensitivity is a property of the function, given a universe of possible databases, and is independent of the actual input database. Intuitively, this maximum difference between the values that the function could take on any two neighboring databases must be hidden in order to preserve differential privacy. We next focus on a technique that hides this maximum difference by adding noise, in the context of a concrete example.

A Single-Statistic Example: Mean Salary

To illustrate some of the delicate issues involved in actually carrying out a differentially-private computation, consider the release of mean salary among (all or some of) the faculty in an economics department. For concreteness, consider the following scenario: each faculty member is asked to voluntarily and confidentially agree to have their individual salary included in some database; statistics from the database are to be released in a differentially-private manner.

Notice that the details of this scenario, including details of the differentially-private mechanism to be used, can be publicly announced. What one aims to hide is only the confidential participation decision by any individual faculty. Our example will illustrate that this individual decision is easier to hide the larger the database is, the fewer statistics are to be published, and the less sensitive these statistics are given the considered universe of possible databases.

Dwork, McSherry, Nissim, and Smith (2006) show that one way to get an ϵ-differentially-private release of a statistic is to add "Laplace noise" to the (true) statistic prior to its release: that is, the noise is drawn from a Laplace distribution with mean equal to zero and standard deviation $= \sqrt{2}\,\Delta f/\epsilon$, where Δf is the sensitivity of the statistic—that is, the maximum difference in the statistic between any two neighboring databases—and ϵ is the parameter that quantifies the level of privacy we choose to guarantee.[8]

To apply this technique one therefore needs, first, to choose a value for ϵ (the smaller it is, the stronger is the privacy guarantee), and second, to calculate Δf. Remember that because these quantities do not depend on the underlying data, they are not in themselves private.

Choosing a value for ϵ: Recall that ϵ quantifies the maximum multiplicative difference possible between a differentially-private computation's outcome probabilities across two neighboring databases. In choosing a value for ϵ, one therefore chooses the maximum such difference that one is willing to allow under the differential privacy guarantee. But what should this maximum difference be? For the

[7] Formally, the sensitivity of a function f is $\Delta f = \max_{D,\,D'} |f(D) - f(D')|$, for (D, D') neighboring databases.

[8] With scale parameter $b = \Delta f/\epsilon$, the probability density function of this distribution is $\frac{1}{2b}\, e^{-\frac{|x|}{b}}$ and its standard deviation is $\sqrt{2}\, b$. This distribution is a natural choice because its exponential form satisfies the multiplicative e^{ϵ} constraint in the differential privacy definition.

most part, the differential privacy literature is silent on this question. Developing the reasoning and intuition necessary for determining a socially desired value may take time. Concrete proposals may eventually emerge from a combination of philosophical and ethical inquiry, and social, political, and legislative processes, and could depend on context; further research is clearly needed.[9] For illustrative purposes only, in our mean salary example we will consider the values $\epsilon = 0.1$ and $\epsilon = 1$.

Calculating Δf*:* As mentioned above, to yield practical results our technique requires Δf to be bounded. Our example involves salary rather than total income, because salary is bounded from below (in the worst case, at zero). One still needs an upper bound, which cannot be naively calculated from the data, but should be a property of the known universe of possible salaries. For simplicity, we assume that it is known to be some $\bar{y}$. With these bounds, and with mean salary as our function of interest, the absolute value of the difference between the function applied to two neighboring databases will be less than or equal to the highest possible salary divided by the number (denoted by n) of individuals in the larger database: $|f(D) - f(D')| \leq \bar{y}/n$, for any two neighboring databases (D, D').

Since the number of faculty members participating in the database is not publicly known, the universe of possible databases includes the case $n = 1$, and therefore $\Delta f = \bar{y}$.[10] With such high sensitivity, a naive application of the Laplace noise technique yields a uselessly uninformative outcome at any n: the noise added to the true mean has standard deviation $\sqrt{2}\,\bar{y}/\epsilon$, which, even with $\epsilon = 1$, is larger than the upper bound on salaries.

An easy modification of the technique, however, yields noise that shrinks with n. The idea is to think of the mean as the function sum/n, that is, as a function of two statistics—the sum of salaries, and the sample size n—to be calculated and released in a differentially-private manner. One then divides the privacy budget ϵ between the two statistics: $\epsilon_{sum} + \epsilon_n = \epsilon$ (recall that the composition property allows such a division). The sensitivity of the sum of salaries is $\bar{y}$ because the maximum difference between the sum of salaries across two databases that differ only in the inclusion versus exclusion of one record is the upper bound on one additional salary. The sensitivity of n is 1, because by the definition of neighboring

[9] As Dwork et al. (2011) note in a defense of differential privacy:

> Yes, this research is incomplete. Yes, theorems of the following form seem frighteningly restrictive:
>
> > If an individual participates in 10,000 adversarially chosen databases, and if we wish to ensure that her cumulative privacy loss will, with probability at least $1 - e^{-32}$, be bounded by e^1, then it is sufficient that each of these databases will be $\epsilon = 1/801$-differentially private.
>
> But how else can we find a starting point for understanding how to relax our worst-case adversary protection? How else can we measure the effect of doing so? And what other technology permits one to prove such a claim?

[10] For simplicity (and conservativeness), we define mean salary in a database with 0 individuals to be at the lower bound 0.

databases, the difference between the number of records across any two neighboring databases is 1. The noise added to the two statistics would therefore have standard deviations $\sqrt{2}\,\bar{y}/\epsilon_{sum}$ and $\sqrt{2}/\epsilon_n$, respectively. Because the two statistics increase with n, the noise-to-true-statistic ratio of each vanishes asymptotically. With $\epsilon = 1$ and a favorable setting—a large department with high rate of voluntary participation in the database, and with mean salary not much below $\bar{y}$—the differentially-private release may convey some usable information about the true mean; but generally, the promise of the approach is more apparent on bigger data.

For illustration, consider $\epsilon = 1$, mean salary $\bar{y}$ (this is the unrealistically favorable case of all salaries in the department equal, at the upper limit), and $n = 30$. Then the standard deviation on the noise added by the Laplace technique would be $(\sqrt{2}/0.5)/30 = 9.4$ percent of each of the two (true) statistics, assuming for simplicity we divide the privacy budget equally between the two statistics.

For comparison, consider mean salary among the American Economic Association (AEA) membership in 2012, reported at 18,061 members (Rousseau 2013). Pick a tenfold stronger privacy guarantee, that is, $\epsilon = 0.1$, and assume a more realistic relation between the upper bound and the true mean, say, mean salary $= \bar{y}/10$. Assuming that all members volunteer to participate in the database, the much larger n means that in spite of these significantly more conservative conditions, the standard deviation on the noise added by the Laplace technique would be a much more tolerable 1.6 percent of the true sum of salaries and 0.16 percent of the true n (that is, a standard deviation of 28 members), if the privacy budget is again divided equally—rather than optimally—between the two statistics. Of course, things look still better with still bigger data and cleverer techniques.

Mean Salary Revisited: When the Database Size is Known

Dwork (2011a) suggests that "[s]ometimes, for example, in the census, an individual's participation is known, so hiding presence or absence makes no sense; instead we wish to hide the values in an individual's row." Our examples above could be modified to match such settings. Instead of a scenario where each faculty member is asked to voluntarily join a database, consider a different scenario where some administrative database with everyone's salaries is already known to exist. As above, statistics from the database are to be released in a differentially-private manner. Under this modified scenario, the databases from our examples above are now known to include all the faculty in an economics department and all AEA members.

In such settings, where participation is publicly known, it may make sense to modify our above definition of neighboring databases, from pairs "that are identical except that one of the databases has one additional row," to pairs of known size n, that differ in the content of exactly one row. In this form, differential privacy guarantees participants that if their true salary y were replaced with some fake salary $y' \in [0, \bar{y}]$, the probability of any given outcome would not change by much. With this modification, differentially-private release of mean salary requires only the sum of salaries to be computed and released in a differentially-private manner.

Historically, this alternate definition (with databases of fixed and publicly known n) was used in the first papers that sparked the differential privacy literature, and it is still used in much of the work on differential privacy and statistics—a body of work that has grown quickly over the past few years. Work in this area has repeatedly established the feasibility of achieving common statistical goals while maintaining differential privacy. Differentially-private versions have been developed for large classes of estimators—including those used routinely by empirical economists—often with little effective cost in terms of accuracy of the released results.[11]

Multiple Statistics

Of course, researchers wish to publish more than one statistic per database. In our example above, the privacy budget ϵ was divided between two statistics, *sum* and n, and each was then independently computed in a differential-privacy-preserving way. An alternative approach is to compute the two (or more) statistics jointly, which in some cases may significantly reduce the amount of added noise, as demonstrated by the case of histograms (Dwork, McSherry, Nissim, and Smith 2006).

Consider the release of a frequency histogram of salaries in some database. Treating each bin as a separate statistic (for example, "the count of rows with salary \$0–10,000" is one statistic) would require dividing the privacy budget ϵ between the bins. The sensitivity (that is, Δf) of each such bin statistic is 1. It turns out that a generalized sensitivity concept applied jointly to the entire histogram is also 1, since adding an individual to a database always adds 1 to the count of exactly one of the bins and 0 to all others. In this example, calculating all the bins of the histogram jointly reduces the added noise because it saves the need to first divide the privacy budget between the statistics—a division whose cost in added noise increases with the number of bins. More generally, consider the maximum possible effect on a statistic of adding one individual to the database; if such a worst-case effect cannot occur on each of a group of statistics at the same time, considering them jointly may improve results.

One of the main focuses of research in differential privacy in recent years has been to develop algorithms that can handle very large numbers of queries jointly with far less noise than simple noise addition would permit. This large literature, which begins with Blum, Ligett, and Roth (2013) and continues with Hardt and Rothblum (2010) and Hardt, Ligett, and McSherry (2012), develops techniques for generating "synthetic data"—a set of valid database rows—that approximate the correct answers to

[11] Here we provide a few examples; see Heffetz and Ligett (2013) for a fuller reference list. Dwork and Lei (2009) demonstrate differentially-private algorithms for interquartile distance, median, and linear regression. Lei (2011) and Nekipelov and Yakovlev (2011) study differentially-private M-estimators. Smith (2008, 2011) finds that for almost any estimator that is asymptotically normal on independent and identically distributed samples from the underlying distribution (including linear regression, logistic regression, and parametric maximum likelihood estimators, under regularity conditions), there are differentially-private versions with asymptotically no additional perturbation. Along with these and other theoretical papers, a number of papers empirically investigate the performance of differentially-private estimators; useful starting points include Vu and Slavkovic (2009), Chaudhuri, Monteleoni and Sarwate (2011), and Abowd, Schneider and Vilhuber (2013).

all of a large, fixed set of queries. These techniques go far beyond just perturbing the data. Using ideas from geometry and computational learning theory, they generate synthetic data consisting of artificial records that cannot be connected with a single or small number of records in the original data. These approaches have started to show practicality, in the form of simple implementations that achieve good accuracy when tested on common statistical tasks using standard benchmark data (Hardt, Ligett and McSherry 2012), but much remains to be done.[12]

From Intuitions to Provable Guarantees

What insights can the differential privacy literature offer regarding the cautionary tales above? What tools could it provide for researchers working with data? We offer some thoughts, and highlight how different approaches respond differently to the inherent, unavoidable tradeoff between privacy and accuracy. We then discuss some of the limitations, as well as additional applications, of differential privacy.

Lessons and Reflections

In the Massachusetts Group Insurance Commission case—and, more generally, regarding the "anonymization" of complex datasets—lessons from differential privacy suggest considering two alternatives. First, one could release a differentially-private, synthetic (that is, artificial) version of the original database, after removing or coarsening complex fields such as text (which, without coarsening, would have made the data too high-dimensional for a synthetic version to be feasible in practice). The synthetic data would only be useful for a predetermined (though potentially quite large) set of statistics.[13] Second, one could withhold the full data but provide a differentially-private interface to allow researchers (or possibly the general public) to issue queries against the database.

Both approaches—providing a sanitized database, and providing sanitized answers to individual queries—face the inescapable tradeoff between privacy and usefulness (or accuracy). To achieve privacy, they limit usefulness in different ways: while the first approach limits in advance the type of queries (and hence of analysis)

[12] Another growing literature considers large sets of queries of a particular type, and aims to get a better understanding of the privacy–accuracy tradeoffs for a specific combined task. Beginning with Barak et al. (2007), one application that has received substantial attention is contingency tables, which are computed from sets of *k-way marginal* queries; see Heffetz and Ligett (2013) for references to more recent work.

[13] Kinney et al. (2011) provide evidence of the promise of synthetic data, describing the generation of an initial version of the SynLBD, a synthetic version of the US Census Bureau's Longitudinal Business Database (https://www.census.gov/ces/dataproducts/synlbd/). Their synthetic database is designed to preserve aggregate means and correlations from the underlying, confidential data. While their algorithm for generating synthetic data is not explicitly designed to preserve any particular level of differential privacy, they present an interesting illustrative assessment—inspired by differential privacy—of privacy risk. See Machanavajjhala et al. (2008) for an earlier exploration of the challenges of generating privacy-preserving synthetic data from other Census datasets.

possible, the second maintains flexibility but might more severely limit the overall *number* of queries, since the system has to manage a privacy budget dynamically (and hence potentially less efficiently) to answer arbitrary queries as they arrive and would eventually run out of its ϵ privacy budget and then would have to refuse new queries. This idea of an overall limit—a privacy budget that places a quantifiable constraint on any approach—is a useful metaphor that highlights one of the costs of preserving privacy: it imposes fundamental limits on how much information can be revealed about the data.

In the case of the AOL debacle, the data to be released were so high-dimensional (the space of rows being all possible search histories) that they clearly could not be handled with differential privacy without some initial dimension reduction. This point in itself is worth observing—free text and other high-dimensional data (for example, genetic information) are potentially extraordinarily revealing, and deserve careful attention. Korolova, Kenthapadi, Mishra, and Ntoulas (2009), in response to AOL's data release, propose releasing an alternate data structure called a *query click graph*, and demonstrate on real search log data that a differentially-private query click graph can be used to perform some research tasks that one might typically run on search logs.[14] As the authors note, it remains to be seen how broadly useful such sanitized data are, but such findings "offer a glimmer of hope" on reconciling research usability with privacy concerns.

Regarding the Netflix challenge, the *manner* in which it was carried out—releasing a large, very high-dimensional dataset—is difficult to implement in a differentially-private way. However, the *goals* of the challenge—namely, producing recommendations from collective user behavior—could be achievable while guaranteeing differential privacy. To explore this possibility, McSherry and Mironov (2009) evaluate several of the algorithmic approaches used in the challenge, showing that they could have been implemented in a differentially-private manner (via privacy-preserving queries issued against the database) without significant effect on their accuracy.

The Facebook goal—giving advertisers a count of the number of times their ad was shown—at first sounds as if it might be well-suited to differential privacy: one could simply add an appropriate level of noise to the true count. However, charging advertisers based on noisy counts may be considered objectionable, and regardless, privacy would then degrade as the number of ad campaigns increased (or, alternatively, Facebook would have to discontinue the service once they ran out of a certain ϵ budget to which they had committed). Even if we assume that advertisers do not share the statistics Facebook reports to them (and so perhaps each advertiser can be apportioned a separate privacy budget rather than sharing a single budget among

[14] The differentially-private query click graph the authors propose to publish is a noisy version of a "graph where the vertices correspond to both queries and URLs and there is an edge from a query to a URL with weight equal to the number of users who click on that URL given they posed the query. Each query node is labeled by the number of times this query was posed in the log. Similarly, there is an edge from one query to another query with weight equal to the number of users who posed one query and reformulated to another."

them all), large advertisers likely run so many campaigns that the noise necessary in order to ensure any reasonable level of privacy would swamp any signal in the data. Korolova (2011) suggests that an approach like differential privacy would provide the most robust starting point for privately addressing Facebook's goal, and discusses these and other challenges that leave the targeted-ads application an intriguing open problem.

More generally, what tools and other thoughts could differential privacy potentially offer to researchers who work with data?

While no standardized implementations yet exist, and while conventions (for example regarding setting ϵ) have not yet been established, a rich set of theoretical results already provides the foundations for a useful toolbox for the data-based researcher.

If one would like to publish a single statistic (or a small set of statistics), differentially-private estimator versions might already exist. As discussed above, the accuracy cost imposed by the added noise may be negligible when the sample size n is sufficiently large.

Regardless of whether the statistic of interest has received attention in the differential privacy literature, the study of differential privacy suggests that it may be helpful to understand the *sensitivity* of the statistic to changes in one person's information—how much can varying one entry in the database affect the statistic? Such understanding not only helps assess how much noise one could add to achieve differential privacy in the simplest manner; it is also helpful for getting an intuitive understanding of how and why a statistic might be revealing. There are also differentially-private techniques that can provide good accuracy even on high-sensitivity statistics, so long as the statistics are "well-behaved" on the data of interest (Nissim, Raskhodnikova, and Smith 2007; Dwork and Lei 2009). Finally, if one wishes to publish a large set of statistics or produce sanitized data, as we discussed, general purpose techniques for doing so already exist, but it is possible that a researcher's particular properties of interest would be even better served by a specialized differentially-private mechanism.

The centrality of the notion of sensitivity to the ongoing research on differential privacy highlights an old truth from a new perspective: it underscores the importance of thinking about the robustness of the statistics we report. If reporting a statistic while preserving privacy requires introducing an unacceptable level of randomness, this may indicate that one's dataset is too small for one's desired levels of privacy and accuracy, but it may also suggest that worst-case scenarios exist under which the statistic is simply not robust—that is, it may be quite sensitive to potential individual outliers.

Finally, the concept of differential privacy offers one way to quantify the often loosely used notions of privacy and anonymity. Researchers may find such quantification helpful in thinking about whether study participants should be given a different, more qualified, promise of privacy/anonymity than is typically given—especially in settings where implementing a specific guarantee (not necessarily the one offered by differential privacy) is not practical.

Limitations

Like any other rigorous approach, the differential privacy approach makes some assumptions that may be questioned. For example, it assumes that an individual's private data are conveniently represented as a row in a database (an assumption violated by, for example, social network data), and it implicitly assumes that a particular definition—involving a bound on the ratio of outcome probabilities—captures what we mean by privacy.

Strong privacy guarantees necessarily obscure information. The intentional introduction of randomness into published outcomes may require adjustments to specific implementations of scientific replication. More generally, for some applications the very idea of deliberately introducing randomness is problematic: preventable mistakes such as allocating the wrong resources to the wrong groups or making the wrong policy decisions could have grave consequences.

As hinted above, a potential limitation of differentially-private mechanisms producing synthetic data is that they require the data analyst to specify the query set in advance. In many research settings, one may not know in advance exactly which statistics one wishes to compute or what properties of a dataset must be preserved in order for the data to be useful. There is a natural tension between an analyst's desire to "look at the data" before deciding what to do with them and a privacy researcher's desire that all computations that touch the original data be made formal and privacy-preserving.

As a practical response to this limitation, rather than attempting to define the query set a priori, one could consider using some of the privacy budget for *interactive queries* where the analyst poses queries one at a time and receives privacy-preserving answers, and could then base the choice of future queries on the answers previously received. The analyst thus establishes via this sequence of interactive queries what properties of the original database to preserve in the sanitized version, and can then use the rest of the privacy budget to produce sanitized data.

More generally, with the growth of big data, the "look at the data" approach is destined to change: in practical terms, "looking" at enormous datasets means running analyses on them. As soon as "looking at the data" has a technical meaning, one can try to enable it in a privacy-preserving manner.

Finally, for particular applications, differentially-private mechanisms may not yet have been developed, or the existing technology may not enable a satisfying privacy–accuracy tradeoff. Such limitations may merely suggest that more research is needed. Even when a satisfying privacy–accuracy tradeoff is formally proved impossible, in many cases such impossibility results are not specific to differential privacy, but rather reflect that certain tasks are inherently revealing and hence may be fundamentally incompatible with privacy.

Differential Privacy and Mechanism Design

The last few years have seen a growth of interest in a number of topics at the intersection of differential privacy and economics, in particular, privacy and mechanism design; see Pai and Roth (2013) for a survey. Some of the key questions under consideration include how one might incorporate privacy considerations into utility functions and how one might model the value of privacy. Work in this area includes Ghosh and Roth (2011), Nissim, Orlandi, and Smorodinsky (2012), Fleischer and Lyu (2012), Roth and Schoenebeck (2012), Ligett and Roth (2012), Xiao (2013), Chen et al. (2013), and Ghosh and Ligett (2013).

From a mechanism design point of view, the differential privacy guarantee—that a participant's inclusion or removal from the database would have almost no effect on the outcome—could be viewed as a valuable guarantee even in the absence of privacy concerns. In particular, consider settings where participants in a database can misrepresent their individual data, and have preferences over the possible outcomes of a function to be computed from the data. A differentially-private computation implies that such participants have only limited incentive to lie, because lying would have only a limited effect on the outcome. McSherry and Talwar (2007) were the first to observe that differential privacy implies asymptotic (or approximate) "strategyproofness" (or truthfulness). Of course, under differential privacy, not only do individuals have almost no incentive to lie; they also have almost no incentive to tell the truth (Nissim, Smorodinsky and Tennenholtz 2012; Xiao, 2013); however, a small psychological cost of lying could strictly incentivize truth-telling.

This implication of approximate truthfulness may be of particular interest to researchers who wish to gather survey data in settings where participation is voluntary and the accuracy of responses cannot be easily verified. More generally, the asymptotic strategyproofness implied by differential privacy inherits some of the latter's useful additional properties. For example, because of the way differential privacy extends to groups of k individuals, this strategyproofness extends to the case of k colluding individuals (a collusion resistance that deteriorates with the coalition size k). The strategyproofness also holds under repeated application of the mechanism (again, with a deterioration as the number of repetitions rises). Finally, this asymptotic truthfulness has inspired further work on privacy-preserving mechanism design (Huang and Kannan 2012; Kearns, Pai, Roth, and Ullman 2014) and has enabled differential privacy to be used as a tool in the design of truly strategyproof mechanisms (for example, Nissim, Smorodinsky, and Tennenholtz 2012).

Concluding Thoughts

Privacy concerns in the face of unprecedented access to big data are nothing new. More than 35 years ago, Dalenius (1977) was discussing "the proliferation of computerized information system[s]" and "the present era of public concern about 'invasion of privacy.'" But as big data get bigger, so do the concerns. Greely (2007) discusses genomic databases, concluding:

> The size, the cost, the breadth, the desired broad researcher access, and the likely high public profile of genomic databases will make these issues especially

> important to them. Dealing with these issues will be both intellectually and politically difficult, time-consuming, inconvenient, and possibly expensive. But it is not a solution to say that "anonymity" means only "not terribly easy to identify," . . . or that "informed consent" is satisfied by largely ignorant blanket permission.

Replacing "genomic databases" with "big data" in general, our overall conclusion may be similar.

The stories in the first part of this paper demonstrate that relying on intuition when attempting to protect subject privacy may not be enough. Moreover, privacy failures may occur even when the raw data are never publicly released and only some seemingly innocuous *function* of the data, such as a statistic, is published.

The differential privacy literature provides a framework for thinking more precisely about privacy–accuracy tradeoffs. With computer scientists using phrases such as "the amount of privacy loss" and "the privacy budget," the time seems ripe for more economists to join the conversation. Is privacy a term in the utility function that can in principle be compared against the utility from access to accurate data? Should individuals be entitled to privacy—or to a certain lifelong privacy budget—as a basic right, or as a property right? Should a certain privacy budget be allocated across interested users of publicly owned data, like Census data, and if so, how? If a budget were allocated to individuals, should fungible, transferable ϵ be allowed to be sold in markets from private individuals to potential data users, and if so, what would its price be?

When big data means large n, an increasing number of common computations can be achieved in a differentially-private manner with little cost to precision. It is not inconceivable that within a few years, many of the computations that have been—and those that are yet to be—proven achievable in theory will be applied in practice. Dwork and Smith (2010) write that they "would like to see a library of differentially-private versions of the algorithms in R and SAS." In a similar spirit, we would be happy to have a differentially-private option in estimation commands in STATA. But ready-to-use, commercial-grade applications will not be developed without sufficient demand from potential users. We hope that the incorporation of privacy considerations into the vocabulary of empirical researchers will help raise demand, and stimulate further discussion and research—including, we hope, regarding additional approaches to privacy.

Until such applications are available, it might be wise to pause and reconsider researchers' promises and, more generally, obligations to subjects. When researchers (and Institutional Review Boards!) are confident that the data pose only negligible privacy risks—as in the case of some innocuous small surveys and lab experiments—it may be preferable to replace promises of complete anonymity with promises for "not terribly easy" identification or, indeed, with no promises at all. In particular, researchers could explicitly inform subjects that a determined attacker may be able to identify them in posted data, or even learn things about them merely by looking at the empirical results of a research paper. We caution against taking the naive alternate route of simply refraining from making seemingly harmless data publicly available; freedom of information, access to data, transparency, and scientific replication are all dear to us.[15] Of course, the tradeoffs, and in particular the question of what privacy risks are negligible and what data are harmless, should be considered and discussed; a useful question to ask ourselves may resemble the old "*New York Times* test": Would our subjects mind if their data were identified and published in the *New York Times*?

■ *For useful comments on an early draft, we thank Dan Benjamin, Avrim Blum, Kamalika Chaudhuri, Hank Greely, Aleksandra Korolova, Frank McSherry, Ilya Mironov, Denis Nekipelov, Kobbi Nissim, Ted O'Donoghue, Grant Schoenebeck, Moses Shayo, Adam Smith, Latanya Sweeney, Kunal Talwar, and Jonathan Ullman. We are also grateful to the editors—David Autor, Chang-Tai Hsieh, Ulrike Malmendier, and Timothy Taylor—for their encouragement and comments. Ligett's work was supported in part by an NSF CAREER award (CNS-1254169), the US–Israel Binational Science Foundation (grant 2012348), the Charles Lee Powell Foundation, a Google Faculty Research Award, and a Microsoft Faculty Fellowship.*

References

Abowd, John M., Matthew J. Schneider, and Lars Vilhuber. 2013. "Differential Privacy Applications to Bayesian and Linear Mixed Model Estimation." *Journal of Privacy and Confidentiality* 5(1).

Barak, Boaz, Kamalika Chaudhuri, Cynthia Dwork, Satyen Kale, Frank McSherry, and Kunal Talwar. 2007. "Privacy, Accuracy, and Consistency Too: A Holistic Solution to Contingency Table Release." In *PODS '07: Proceedings of the Twenty-Sixth ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems,* pp. 273–82. ACM Digital Library.

Barbaro, Michael, and Tom Zeller. 2006. "A Face Is Exposed for AOL Searcher No. 4417749." *New York Times*, August 9. http://www.nytimes.com/2006/08/09/technology/09aol.html.

Barth-Jones, Daniel C. 2012. "The 'Re-Identification' of Governor William Weld's Medical Information: A Critical Re-Examination of Health Data Identification Risks and Privacy Protections, Then and Now." Available at SSRN: http://ssrn.com/abstract=2076397.

Blum, Avrim, Katrina Ligett, and Aaron Roth. 2013. "A Learning Theory Approach to Noninteractive Database Privacy." *Journal of the ACM* 60(2): Article 12.

Chaudhuri, Kamalika, Claire Monteleoni, and Anand D. Sarwate. 2011. "Differentially Private Empirical Risk Minimization." *Journal of Machine Learning Research* 12(March): 1069–1109.

Chen, Yiling, Stephen Chong, Ian A. Kash, Tal Moran, and Salil P. Vadhan. 2013. "Truthful Mechanisms for Agents that Value Privacy." In *EC '13: Proceedings of Fourteenth ACM Conference on Electronic Commerce*, pp. 215–32. ACM Digital Library.

[15] Flood, Katz, Ong, and Smith (2013) provide a comprehensive discussion of such a transparency–confidentiality tradeoff in a context that is very different from ours, yet of great interest to economists—that of financial supervision and regulation.

Dalenius, Tore. 1977. "Towards a Methodology for Statistical Disclosure Control." *Statistisk tidskrift,* 15: 429–44.

Dinur, Irit, and Kobbi Nissim. 2003. "Revealing Information While Preserving Privacy." In *PODS '03: Proceedings of the Twenty-second ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems,* pp. 202–210. ACM Digital Library.

Dwork, Cynthia. 2006. "Differential Privacy." ICALP '06: Proceedings of the 33rd *International Conference on Automata, Languages and Programming,* pp. 1–12. ACM Digital Library.

Dwork, Cynthia. 2011a. "A Firm Foundation for Private Data Analysis." *Communications of the ACM* 54(1): 86–95.

Dwork, Cynthia. 2011b. "The Promise of Differential Privacy: A Tutorial on Algorithmic Techniques." In *FOCS '11: Proceedings of the 52nd Annual IEEE Symposium on Foundations of Computer Science*, pp. 1–2. ACM Digital Library.

Dwork, Cynthia, and Jing Lei. 2009. "Differential Privacy and Robust Statistics." In *STOC '09: Procceedings of the Forty-first Annual ACM Symposium on Theory of Computing*, pp. 371–80. ACM Digital Library.

Dwork, Cynthia, and Kobbi Nissim. 2004. "Privacy-Preserving Datamining on Vertically Partitioned Databases." In *Advances in Cryptology—CRYPTO* 2004, *24th Annual International Cryptology Conference,* pp. 528–44.

Dwork, Cynthia, Krishnaram Kenthapadi, Frank McSherry, Ilya Mironov, and Moni Naor. 2006. "Our Data, Ourselves: Privacy via Distributed Noise Generation." In *EUROCRYTO '06: Proceedings of the 24th Annual International Conference on the Theory and Applications of Cryptographic Techniques*, pp. 486–503. ACM Digital Library.

Dwork, Cynthia, Frank McSherry, Kobbi Nissim, and Adam Smith. 2006. "Calibrating Noise to Sensitivity in Private Data Analysis." In *TTC '06: Proceedings of the Third Conference on Theory of Cryptography,* pp. 265–84. ACM Digital Library.

Dwork, Cynthia, Frank McSherry, Kobbi Nissim, and Adam Smith. 2011. "Differential Privacy: A Primer for the Perplexed." Joint UNECE/Eurostat Work Session on Statistical Data Confidentiality, WP. 26.

Dwork, Cynthia, and Adam Smith. 2010. "Differential Privacy for Statistics: What We Know and What We Want to Learn." *Journal of Privacy and Confidentiality* 1(2): 135–54.

Fleischer, Lisa, and Yu-Han Lyu. 2012. "Approximately Optimal Auctions for Selling Privacy When Costs Are Correlated with Data." In *EC '12: Proceedings of the 13th ACM Conference on Electronic Commerce,* pp. 568–85. ACM Digital Library.

Flood, Mark, Jonathan Katz, Stephen J. Ong, and Adam Smith. 2013. "Cryptography and the Economics of Supervisory Information: Balancing Transparency and Confidentiality." Federal Reserve Bank of Cleveland Working Paper 1312.

Ghosh, Arpita, and Katrina Ligett. 2013. "Privacy and Coordination: Computing on Databases with Endogenous Participation." In *EC '13: Proceedings of the Fourteenth ACM Conference on Electronic Commerce*, pp. 543–60. ACM Digital Library.

Ghosh, Arpita, and Aaron Roth. 2011. "Selling Privacy at Auction." *EC '11: Proceedings of the 12th ACM Conference on Electronic Commerce*, pp. 199–208. ACM Digital Library.

Greely, Henry T. 2007. "The Uneasy Ethical and Legal Underpinnings of Large-Scale Genomic Biobanks." *Annual Review of Genomics and Human Genetics* 8: 343–64.

Hardt, Moritz, Katrina Ligett, and Frank McSherry. 2012. "A Simple and Practical Algorithm for Differentially Private Data Release." In *Advances in Neural Information Processing Systems* 25: 2348–56.

Hardt, Moritz, and Guy N. Rothblum. 2010. "A Multiplicative Weights Mechanism for Privacy-Preserving Data Analysis." In *FOCS '10: Proceedings of the 51st Annual IEEE Symposium on Foundations of Computer Science*, pp. 61–70. ACM Digital Library.

Heffetz, Ori, and Katrina Ligett. 2013. "Privacy and Data-Based Research." NBER Working Paper 19433.

Huang, Zhiyi, and Sampath Kannan. 2012. "The Exponential Mechanism for Social Welfare: Private, Truthful, and Nearly Optimal." In *FOCS '12: Proceedings of the 53rd Annual IEEE Symposium on Foundations of Computer Science*, pp. 140–49. ACM Digital Library.

Kasiviswanathan, Shiva Prasad, and Adam Smith. 2008. "A Note on Differential Privacy: Defining Resistance to Arbitrary Side Information." Unpublished paper.

Kearns, Michael, Mallesh Pai, Aaron Roth, and Jon Ullman. 2014. "Mechanism Design in Large Games: Incentives and Privacy." In *ITCS '14: Proceedings of the 5th Conference on Innovations in Theoretical Computer Science*, pp. 403–410. ACM Digital Library.

Kinney, Satkartar K., Jerome P. Reiter, Arnold P. Reznek, Javier Miranda, Ron S. Jarmin, and John M. Abowd. 2011. "Towards Unrestricted Public Use Business Microdata: The Synthetic Longitudinal Business Database." *International Statistical Review* 79(3): 362–84.

Klarreich, Erica. 2012. "Privacy by the Numbers: A New Approach to Safeguarding Data." *Quanta Magazine,* December 10.

Korolova, Aleksandra. 2011. "Privacy Violations Using Microtargeted Ads: A Case Study." *Journal of Privacy and Confidentiality* 3(1).

Korolova, Aleksandra, Krishnaram Kenthapadi, Nina Mishra, and Alexandros Ntoulas. 2009. "Releasing Search Queries and Clicks Privately." In *WWW '09: Proceedings of the 18th International Conference on the World Wide Web,* pp. 171–180. ACM Digital Library.

Kumar, Ravi, Jasmine Novak, Bo Pang, and Andrew Tomkins. 2007. "On Anonymizing Query Logs via Token-based Hashing." In *WWW '07: Proceedings of the 16th International Conference on the World Wide Web*, pp. 629–38. ACM Digital Library.

Lei, Jing. 2011. "Differentially Private M-Estimators." *Advances in Neural Information Processing Systems* 24: 361–69.

Ligett, Katrina, and Aaron Roth. 2012. "Take it or Leave it: Running a Survey When Privacy Comes at a Cost." *Internet and Network Economics* (WINE '12: Proceedings of the 8th international conference on Internet and Network Economics), pp. 378–391.

Machanavajjhala, Ashwin, Daniel Kifer, John M. Abowd, Johannes Gehrke, and Lars Vilhuber. 2008. "Privacy: Theory Meets Practice on the Map." In *ICDE '08: Proceedings of the 24th IEEE International Conference on Data Engineering,* pp. 277–86. ACM Digital Library.

McSherry, Frank, and Ilya Mironov. 2009. "Differentially Private Recommender Systems: Building Privacy into the Net." In *KDD' 09: Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining,* pp. 627–36. ACM Digital Library.

McSherry, Frank, and Kunal Talwar. 2007. "Mechanism Design via Differential Privacy." In *FOCS '07: Proceedings of the 48th Annual IEEE Symposium on Foundations of Computer Science*, pp. 94–103. ACM Digital Library.

Narayanan, Arvind, and Vitaly Shmatikov. 2008. "Robust De-anonymization of Large Sparse Datasets." In *SP '08: 2008 IEEE Symposium on Security and Privacy*, pp. 111–25. ACM Digital Library.

Nekipelov, Denis, and Evegeny Yakovlev. 2011. "Private Extremum Estimation." Unpublished paper.

Nissim, Kobbi, Claudio Orlandi, and Rann Smorodinsky. 2012. "Privacy-aware Mechanism Design." In *EC '12: Proceedings of the 13th ACM Conference on Electronic Commerce,* pp. 774–89. ACM Digital Library.

Nissim, Kobbi, Sofya Raskhodnikova, and Adam Smith. 2007. "Smooth Sensitivity and Sampling in Private Data Analysis." In *STOC '07: Proceedings of the Thirty-Ninth Annual ACM Symposium on Theory of Computing*, pp. 75–84.

Nissim, Kobbi, Rann Smorodinsky, and Moshe Tennenholtz. 2012. "Approximately Optimal Mechanism Design via Differential Privacy." In *ITCS '12: Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*, pp. 203–213. ACM Digital Library.

Ohm, Paul. 2010. "Broken Promises of Privacy: Responding to the Surprising Failure of Anonymization." *UCLA Law Review* 57(6): 1701–77.

Pai, Mallesh, and Aaron Roth. 2013. "Privacy and Mechanism Design." *Sigecom Exchanges* 12(1).

Roth, Aaron, and Grant Schoenebeck. 2012. "Conducting Truthful Surveys, Cheaply." In *EC '12: Proceedings of the 13th ACM Conference on Electronic Commerce*, pp. 826–43. ACM Digital Library.

Rousseau, Peter L. 2013. "Report of the Secretary." *American Economic Review* 103(3): 669–72.

Smith, Adam. 2008. "Efficient, Differentially Private Point Estimators." arXiv:0809.4794.

Smith, Adam. 2011. "Privacy-Preserving Statistical Estimation with Optimal Convergence Rates." In *STOC '11: Proceedings of the Forty-third Annual ACM Symposium on Theory of Computing*, pp. 813–22. ACM Digital Library.

Sweeney, Latanya. 1997. "Weaving Technology and Policy Together to Maintain Confidentiality." *Journal of Law, Medicine & Ethics* 25(2–3): 98–110.

Sweeney, Latanya. 2002. "Achieving *k*-anonymity Privacy Protection Using Generalization and Suppression." *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems* 10(5): 571–88.

Sweeney, Latanya. 2013. "Matching Known Patients to Health Records in Washington State Data." http://thedatamap.org/1089-1.pdf.

Sweeney, Latanya, Akua Abu, and Julia Winn. 2013. "Identifying Participants in the Personal Genome Project by Name." http://dataprivacylab.org/projects/pgp/1021-1.pdf.

Vu, Duy, and Aleksandra Slavkovic. 2009. "Differential Privacy for Clinical Trial Data: Preliminary Evaluations." In *ICDMW '09: Proceedings of the 2009 IEEE International Conference on Data Mining Workshops*, pp. 138–43. ACM Digital Library.

Xiao, David. 2013. "Is Privacy Compatible with Truthfulness?" In *ITCS '13: Proceedings of the 4th Conference on Innovations in Theoretical Computer Science,* pp. 67–86. ACM Digital Library.

Journal of Economic Perspectives—Volume 28, Number 2—Spring 2014—Pages 99–118

Slicing Up Global Value Chains†

Marcel P. Timmer, Abdul Azeez Erumban, Bart Los, Robert Stehrer, and Gaaitzen J. de Vries

In an already classic study of the iPod, Dedrick, Kramer, and Linden (2010) discuss how the iPod is assembled in China from several hundred components and parts that are sourced from around the world. This production network is led by Apple, a US-based company, which is estimated to capture between one-third and one-half of an iPod's retail price. Asian firms like Toshiba from Japan and Samsung from South Korea capture another major part as profits from manufacturing high-value components, such as the hard-disk drive, display, and memory. In contrast, assembling and testing activities by Chinese workers is estimated to capture no more than 2 percent. Other studies of tablets, mobile telephones, and laptops suggest a similar pattern of specialization; advanced nations deliver capital and high-skilled labor, capturing most of the value, while emerging countries contribute low-skilled activities that add little value: in another vivid example, Ali-Yrkkö, Rouvinen, Seppälä, and Ylä-Anttila (2011) discuss the Nokia N95 smartphone.

Such case studies are mainly conducted for high-end electronics and for one point in time, which raises obvious questions about the extent to which they represent broader patterns. How pervasive is the process of international production fragmentation for a

■ *Marcel P. Timmer is Professor and Director of the Groningen Growth and Development Centre (GGDC), University of Groningen, Groningen, Netherlands. Abdul Azeez Erumban is Senior Economist at The Conference Board, Brussels, Belgium. Bart Los is Associate Professor of Economics, University of Groningen, Groningen, Netherlands. Robert Stehrer is Staff Economist and Deputy Scientific Director at the Vienna Institute for International Economic Studies (wiiw), Vienna, Austria. Gaaitzen J. de Vries is Assistant Professor of Economics, University of Groningen, Groningen, Netherlands. Their email addresses are m.p.timmer@rug.nl, eabdul75@gmail.com, b.los@rug.nl, stehrer@wiiw.ac.at, and g.j.de.vries@rug.nl.*

† To access the Appendix, visit
http://dx.doi.org/10.1257/jep.28.2.99

doi=10.1257/jep.28.2.99

wider set of products? How does the factor content of these production chains change over time when fragmentation deepens? And how do specialization patterns differ between high-income and emerging economies that participate in these chains? In this paper, we provide a macroeconomic and longitudinal analogy of the iPod exercise, using many countries and many manufacturing products. We "slice up the global value chain" (to borrow the term from Krugman 1995) using a decomposition technique that has recently become feasible due to the development of the World Input-Output Database (Timmer et al. 2014). We trace the value added by all labor and capital that is directly and indirectly needed for the production of final manufacturing goods. The production systems of these goods are highly prone to international fragmentation as many stages can be undertaken in any country with little variation in quality.

We seek to establish a series of facts concerning the global fragmentation of production that can serve as a starting point for future analysis. After a short overview of our data and methods, we discuss four major trends. First, international fragmentation, as measured by the foreign value-added content of production, has rapidly increased since the early 1990s when it made its appearance on a global scale (Feenstra 1998). Second, in most global value chains there is a strong shift towards value being added by capital and high-skilled labor, and away from less-skilled labor. This suggests a pervasive process of technological change that is biased towards the use of skilled labor and capital. Third, within global value chains, advanced nations increasingly specialize in activities carried out by high-skilled workers. The direction of this change follows the intuitive notion of comparative advantage driven by relative factor endowments across countries, but the pace at which it occurs has not been established before. Fourth, emerging economies surprisingly specialize in capital-intensive activities; the capital share in their value added is rising, while the share of low-skilled labor in their value added is declining.

International Fragmentation and Factors of Production: Method and Data

Before laying out some patterns as to how the international fragmentation of production is occurring, it is useful to offer some background on terminology, methods, and data.

Concepts and Definitions

We wish to study the production fragmentation of final products. A final product is consumed, in contrast to intermediate products that continue on in the production process. Consumption is broadly defined to include private and public consumption, as well as investment. A global value chain of a final product is defined as the value added of all activities that are directly and indirectly needed to produce it. This global value chain is identified by the country-industry where the last stage of production takes place before delivery to the final user: for example, the global value chain of electronics from Chinese electrical equipment manufacturing, or of cars from German transport equipment manufacturing. However, it is important to note that the fact that a product is "completed" in a particular country does not necessarily mean that domestic firms are governing the value chain: for example, Apple governs the production network of iPods, although they are completed in China. For more on governance in global value chain production, a useful starting point is Gereffi (1999).

The fragmentation of production processes can take many forms, sometimes characterized as "snakes" and "spiders" (Baldwin and Venables 2013). Snakes involve a sequence in which intermediate goods are sent from country A to B, and incorporated into intermediate goods sent from B to C, and so on until they reach the final stage of production. Spiders involve multiple parts coming together from a number of destinations to a single location for assembly of a new component or final product. Most production processes are complex mixtures of the two. To stick with commonly used terms, we refer to all fragmented production processes as "chains," despite the snake-like connotation of this term.

In this paper we will focus on the global value chains of final manufacturing products, which we refer to as "manufactures." Of course, these do not only contain activities in the manufacturing sector, but also in other sectors such as agriculture, utilities, and business services that provide inputs at any stage of the production process of manufactures. These indirect contributions are sizeable and will be explicitly accounted for through the modeling of input-output linkages across industries. The value added in manufactures chains accounts for about 23 percent of global GDP in 1995. Similar analysis of global production of final services is possible in principle, but the current data is not detailed enough to do so.

The World Input-Output Database

To measure value added in global value chains, we need to track the flow of products across industries and countries. The World Input-Output Database, which is freely available at http://www.wiod.org, has been specifically constructed for this type of analyses (Timmer et al. 2014; Dietzenbacher et al. 2013). It provides world input-output tables for each year since 1995 covering 40 countries, including all 27 countries of the European Union (as of January 1, 2007) and 13 other major economies: Australia, Brazil, Canada, China, India, Indonesia, Japan, Mexico, Russia, South Korea, Taiwan, Turkey, and the United States. These 40 countries represent more than 85 percent of world GDP. In addition, a model for the remaining noncovered part of the world economy is provided such that the value-added decomposition of final output is complete. It contains data for 35 industries covering the overall economy, including agriculture, mining, construction, utilities, 14 manufacturing industries, and 17 services industries. The tables have been constructed by combining national input-output tables with bilateral international trade data, following the conventions of the System of National Accounts.[1]

[1] An online appendix available with this paper at http://e-jep.org offers more detail on the construction of this data.

One also needs detailed value-added accounts that provide information on labor and capital used in production. Three types of workers are identified on the basis of educational attainment levels as defined in the International Standard Classification of Education (ISCED). "Low skilled" (ISCED categories 0, 1, and 2) roughly corresponds to less than secondary schooling. "Medium skilled" (3 and 4) means secondary schooling and above, including certain professional qualifications, but below college degree. "High skilled" (5 and 6) includes those with a college degree and above. For most advanced countries, this data is constructed by extending the EU KLEMS database (O'Mahony and Timmer 2009). For other countries, additional data has been collected according to the same principles. Workers include the self-employed and family workers, and an imputation for their income is made. Capital income is derived as a residual and defined as gross value added minus labor income. It represents remuneration for capital in the broadest sense, including physical capital (such as machinery and buildings), land (including mineral resources), intangible capital (such as patents and trademarks), and financial capital.

Decomposing Global Value Chains

Our aim is to decompose the value of a final product into the value added by all labor and capital employed in its global value chain. We begin by modeling the world economy as an input-output model in the tradition of Leontief (1936) and trace the amount of factor inputs needed to produce a certain amount of final output. Leontief's seminal insight is rather straightforward and intuitive: to produce output one needs labor, capital, and intermediate inputs. These intermediates need to be produced themselves, involving again production factors and intermediates, and so on until all intermediates are accounted for. He provided a mathematical model that allows one to trace the inputs needed in all the stages of production. For an introduction to input-output analysis, Miller and Blair (2009) provide a useful starting point.[2] As an end result, the value of any particular final product is decomposed into the value added by all labor and capital that was needed in any stage of production. In this way, one can provide a consistent accounting system of all value added and all global value chains in the world, as illustrated by Figure 1.

The final column in Figure 1 provides the value added by workers and capital employed in a particular industry and country. A row shows the distribution of this value added across all global value chains in which the industry participates. The global value chains are represented by the columns. There is one column for each final good or service produced in each country. The cells in the column show the origin of all value added needed for the production of the final good. The

[2] A formal description of the method can be found in the appendix with the papers at the JEP website: http://e-jep.org. Our approach is related to Johnson and Noguera (2012a) and Koopman, Wang, and Wei (2014). Rather than using Leontief's insight to analyze the value-added content of trade flows, we focus on the value-added content of final demand. This is more in the spirit of work by Dietzenbacher and Romero (2007) and Antràs, Chor, Fally, and Hillberry (2012), who compute the average number of "transactions" a dollar of a given product will go through before being sold for final use.

Figure 1
An Accounting Framework for Global Value Chains

<table>
<tr><th colspan="3" rowspan="3"></th><th colspan="7">Final products of a global value chain, identified by country and industry of completion</th><th rowspan="3">Value added</th></tr>
<tr><th colspan="3">Country 1</th><th>…</th><th colspan="3">Country M</th></tr>
<tr><th>Industry 1</th><th>…</th><th>Industry N</th><th>…</th><th>Industry 1</th><th>…</th><th>Industry N</th></tr>
<tr><td rowspan="7">Value added from country-industries participating in global value chains</td><td rowspan="3">Country 1</td><td>Industry 1</td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td></tr>
<tr><td>…</td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td></tr>
<tr><td>Industry N</td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td></tr>
<tr><td>…</td><td>…</td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td></tr>
<tr><td rowspan="3">Country M</td><td>Industry 1</td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td></tr>
<tr><td>…</td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td></tr>
<tr><td>Industry N</td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td></tr>
<tr><td colspan="3">Total final output value</td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td>World GDP</td></tr>
</table>

Note: Cell values represent the value added generated in the country-industry given in the row, within the global value chain corresponding to the country-industry of completion given by the column.

sum across all participating industries makes up the gross output value of the final product, given in the bottom row. Note that these industries are domestic as well as foreign. As all final products are being consumed somewhere in the world, output values will equal expenditure. Thus both the columns and the rows add up to world GDP as global final expenditure must be equal to global value added by national accounting convention.

In Table 1, we provide a real world example of the results of such decomposition for the final output of the transport equipment manufacturing industry in Germany—in short, German cars.[3] By summing over all value that is added by labor and capital employed in German industries, the domestic value-added content of the product can be calculated. This includes value added in the car industry itself, but also in other German industries that deliver along the production chain, including services industries. Between 1995 and 2008, the domestic value-added content dropped from 79 to 66 percent. On the flip side, the foreign value-added share increased as intermediates were increasingly imported, generating income for labor and capital employed outside Germany. The foreign value-added share is an indicator of the international fragmentation of production and will be used later on.

[3] In this example, as well as in the remainder of the paper, we will analyze the value of final products at basic prices, which is the ex-factory gate price before delivery to the final consumer. This means that retail trade margins and net taxes are not included. Retail margins can be sizable, and the World Input-Output Database provides data to analyze these margins as well, but this is outside the scope of the present paper as retailing is an activity that is still mainly domestic by nature.

Table 1
Slicing Up the Global Value Chain of German Cars
(*percent of final output value*)

	1995	*2008*
German value added	**79%**	**66%**
High-skilled labor	16%	17%
Medium-skilled labor	34%	25%
Low-skilled labor	7%	4%
Capital	21%	20%
Foreign value added	**21%**	**34%**
High-skilled labor	3%	6%
Medium-skilled labor	6%	9%
Low-skilled labor	4%	4%
Capital	8%	15%
Total final output	**100%**	**100%**

Source: Authors' calculations based on World Input-Output Database, November 2013 Release.
Note: The table gives a breakdown of the value added to final output from German transport equipment manufacturing (ISIC rev. 3 industries 34 and 35).

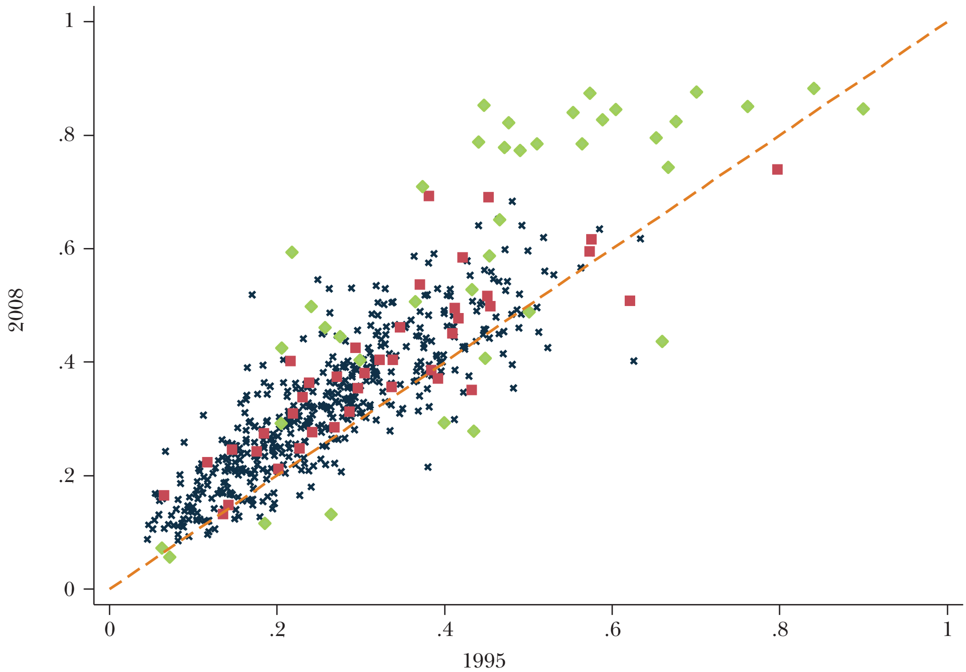
The factor content of the global value chain of German cars changed as well. To see this, one can sum over value added by all labor, irrespective of its location, and similarly for capital. We find that the value added by capital increased from 29 to 35 percent, while the share of labor dropped from 71 to 65 percent. The drop in labor was almost exclusively for less-skilled workers in Germany. The share for high-skilled workers both within and outside Germany increased.

The patterns of shifting location and factor content of activities in the global value chain of German cars are representative for many other chains of manufactures, as we will see in the remainder of this paper. Throughout we will focus on the period from 1995 to 2008 because our data starts in 1995, and 2008 marks the end of a period as the global financial crisis struck. The findings do not depend on the particular choice of beginning or ending year as all of the trends we discuss in this paper are gradual and monotonic, unless noted otherwise.

Trend 1: International Fragmentation of Production is Expanding

With plummeting costs of communication and coordination, it has become increasingly profitable to split the production process, with each stage at its lowest-cost location. Knowledge about the extent and development of international production fragmentation remains sketchy however. Some empirical papers have studied cross-border fragmentation based on foreign investment flow data of firms

Figure 2
Foreign Value-Added Shares in 560 Global Value Chains, 1995 and 2008



Source: Authors' calculations based on World Input-Output Database, November 2013 Release.
Notes: Each dot represents the share of foreign value added in output of a manufactures global value chain in 1995 and 2008. Shares are plotted for 560 global value chains, identified by 14 manufacturing industries of completion in 40 countries. Squares indicate global value chains of electrical equipment (ISIC rev. 3 industries 30–33), and diamonds indicate petroleum refining (ISIC 23). The dashed line is the 45-degree line.

and their affiliates: for example, see Fukao, Ishido, and Ito (2003) and Ando and Kimura (2005) for Japanese firms; Hanson, Mataloni, and Slaughter (2005) for US firms; and Marin (2011) for German and Austrian multinationals. Macroeconomic evidence has been presented by Hummels, Ishii, and Yi (2001) and Johnson and Noguera (2012a, b), who found increasing vertical specialization in trade for most countries (see the contribution of Johnson in this symposium for an overview). Here we provide complementary analysis that provides direct evidence of fragmentation focusing on the value chains of final products.

In Figure 2, we plot foreign value-added shares in 1995 on the horizontal axis and 2008 on the vertical axis, together with a 45-degree line. Products are identified by the country and industry of completion, so we have data for 560 final products from 14 manufacturing industries in 40 countries for each year. For 85 percent of the product chains, the foreign value-added share has increased, indicating the pervasiveness of international fragmentation. The (unweighted) average foreign share rose from 28 to 34 percent.

The extent of fragmentation varies greatly across products. Petroleum products are represented by diamonds in the figure. They have very high foreign value-added shares because most countries do not have access to domestic oil feedstock, reflected in a cluster of diamond-shaped points in the upper part of Figure 2. Value chains for electrical equipment, typically regarded as the paragon of international production fragmentation, are shown by square points. For these products, foreign value-added shares are indeed above average and increased from 33 to 40 percent. In contrast, manufactured foodstuffs have relatively low shares, as most of the intermediates are sourced from local agriculture. But even for these products, foreign shares have increased over time.

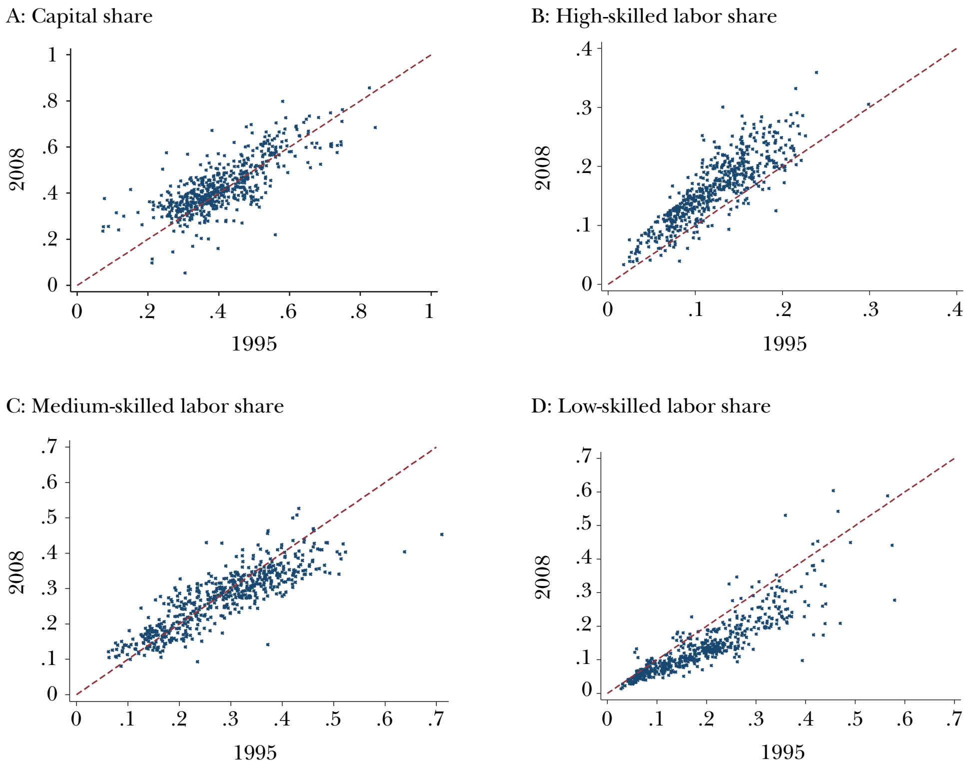
The global financial crisis created a dip in fragmentation in 2008 and 2009, but Los, Timmer, and de Vries (forthcoming) show that the trend picked up again in 2010. Contrary to the anecdotes of multinationals re-shoring production, they found no serious signs of a major reversal yet. However, they do find a major change in the geographical nature of fragmentation. In the 1990s, fragmentation mainly took place within regional blocks: North America (NAFTA), the European Union, and Asia. But in the 2000s, global value chains have started to become truly global with the advance of emerging economies as major suppliers of intermediates. Whether this trend towards global fragmentation of value chains will continue in the future will depend on a host of determinants, including developments in wages and productivity, costs of transportation and trading, coordination costs, risk considerations, and the strength of linkages between various activities. For example, Baldwin and Venables (2013) argue that certain high-value-added tasks may well remain clustered in space because of strong localized complementarities, leading to possibly large discontinuities in the fragmentation process. Furthermore, offshored activities that are currently low-skilled-labor intensive might be re-shored if technological progress makes mechanized production in capital-abundant countries cheaper. It remains to be seen how the different forces will play out in the future.

Trend 2: More Value Added from High-Skilled Labor and Capital

The opening up of China, India, and other emerging economies provided an enduring increase in the global supply of low-skilled labor. How has this affected the factor income distribution in global value chains? This is driven by the relative prices of various types of labor and capital, as well as possibilities for factor substitution, both within and across countries. We first provide evidence on factor content changes at the global level, followed by specialization trends in high-income and emerging economies separately. After that we offer some discussion of a framework that might be suitable when thinking about these trends.

Changes in factor income shares in global value chains have been plotted in Figure 3. The value of final manufacturing goods is decomposed into value added by four factors: capital and low-, medium-, and high-skilled labor. (In our approach, value added and income of factors are equivalent, so these terms will be used

Figure 3
Factor Shares in Value Added of 560 Global Value Chains of Manufactures, 1995 and 2008



Source: Authors' calculations based on World Input-Output Database, November 2013 Release.
Notes: Factor shares in value added of 560 global value chains, identified by 14 manufacturing industries of completion in 40 countries, in 1995 (x-axis) and in 2008 (y-axis). The dashed line is the 45-degree line.

interchangeably.) For each factor we show on the horizontal axis the income share in 1995 and on the vertical axis the share in 2008. Points above the 45-degree line indicate global value chains in which the factor has increased its share. As before, we have in total 560 value chains: 14 manufacturing product groups with 40 possible countries of completion. In 64 percent of the chains, the share of value added by capital has increased. The average increase was about 1 percentage point, with a large variance: in some chains the capital share increased by more than 20 percentage points. It was particularly strong in those production chains where the final output was high, such as transport equipment and machinery with China, Germany, and the United States as countries of completion, in which capital shares increased by 8 percentage points or more. The increase in income shares for high-skilled workers was particularly pervasive and positive, happening in 92 percent of the chains. The

Table 2
Factor Shares in Global Value Chains of All Manufactures

Value added	*1995*	*2008*	*2008 minus 1995*
Total (billion US$)	**$6,586**	**$8,684**	**$2,098**
By:			
capital (%)	40.9%	47.4%	6.5%
high-skilled labor (%)	13.8%	15.4%	1.5%
medium-skilled labor (%)	28.7%	24.4%	−4.2%
low-skilled labor (%)	16.6%	12.8%	−3.8%

Source: Authors' calculations based on World Input-Output Database, November 2013 Release.
Notes: The table presents shares of production factors in total value added based on all global value chains of manufactures. Shares add up to 100 percent. Value added is at basic prices (hence excluding net taxes, trade, and transport margins on output). It is converted to US dollars with official exchange rates and deflated to 1995 prices with the US Consumer Price Index. Figures shown may not add due to rounding.

unweighted average was about 4 percentage points, with a much lower variance than for capital. A notable outlier is the US electrical equipment industry where the share increased by 12 percentage points. On the flip side, the income shares for medium- and low-skilled labor dropped in many value chains. The medium-skilled share declined in 56 percent of the cases, with an average of 1 percentage point. The decline has been particularly severe in major chains, like those of machinery and of transport equipment with Germany and the United States as countries of completion (6 to 8 percentage points decline). The clearest trend is found for low-skilled shares, which declined in 91 percent of the cases. The average decline was 5 percentage points with occasional declines of more than 10 percentage points, in particular in European food chains—for example, with France, Italy, and Spain as countries of completion. As we will see later, declines in low-skilled shares are not only found in chains ending in high-income countries, but also in many chains that have a low-income economy as country-of-completion.

What are the macroeconomic effects? In the analysis above, each product chain was considered irrespective of its size. But bigger chains play a larger role in the global economy than smaller ones. Chains of products like food, transport equipment, and machinery typically have larger final output, as well as chains ending in bigger economies. To account for this, we take final output of all manufactures together (by summing over 560 manufactures chains) and provide a similar decomposition of value added. In effect, the factor shares are now weighted by the final output of their chain. The results are given in Table 2. Global expenditure on manufactures increased by almost one-third, from $6,586 billion in 1995 to $8,684 billion in 2008 (in constant 1995 prices). We find that the shares of value added by capital and high-skilled workers increased at this aggregate level. This confirms that the patterns found above are not driven by developments in small chains only, but are economically significant. The share of value added by capital increased by more than 6 percentage points as the upward shift was most pronounced in bigger chains. The share of high-skilled workers increased as well, but not as fast, with 1.5 percentage points. The shares of low- and medium-skilled workers declined both by about 4 percentage points.

Thus, we find a bifurcation in the factor content of global value chains with increasing capital and high-skilled labor income on the one hand, and declining shares for medium- and particularly for low-skilled labor on the other. Together capital and high-skilled labor captured 55 percent of manufactures value in 1995, increasing to 63 percent in 2008. This increase is especially marked at the end of the 1990s and again from 2003 to 2006. The latter period coincides with a step up in the global presence of China after its accession to the World Trade Organization in 2001. This finding is consistent with the model of Rodrik (1997). He argues that the opening up of international capital markets increased the opportunities for quick relocation of capital. In his argument, this led to a decline in the bargaining power of labor around the world, limiting the share of labor in value added vis-à-vis capital.

Trend 3: Enhanced Specialization in High-Skilled Labor in High-Income Countries

What happened to the location of value added in global value chains? And did specialization patterns vary between regions? To this end, we group Australia, Canada, Japan, South Korea, Taiwan, the United States, and the 15 pre-2004 members of the European Union in one group and place all other countries in the world in another group. Roughly speaking, this can be viewed as a comparison of the high-income countries of the world and other countries that play an active role in international trade (Hanson 2012). The share of high-income countries in total value added generated in all manufactures chains declined from 74 percent in 1995 to 56 percent in 2008. The share of high-income East Asia declined from 21 to 11 percent. Shares in North America and high-income Europe declined by around 4 percentage points each. In contrast, emerging regions have rapidly increased shares by 18 percentage points. China is responsible for half of this increase, from 4 to 13 percent, accelerating in the period after it joined the World Trade Organization in 2001. Shares also rapidly increased in other emerging economies, including Brazil, Russia, India, and Mexico.[4]

[4] This is shown in Appendix Table 1, available online with this article at http://e-jep.org. Given sizable flows of foreign investment, part of the value added in emerging regions will accrue as income to multinational firms headquartered in advanced regions. However, analyzing capital income on a national rather than a domestic basis is notoriously difficult. To establish the full link from production value-added to factor incomes and finally to personal income distributions, one would additionally need data on the actual ownership of firms (Lipsey 2010).

Table 3
Factor Shares in Global Value Chains of Manufactures, by Region

Value added in value chains of manufactures	*1995*	*2008*	*2008 minus 1995*
In high-income countries (billion US$)	**$4,863**	**$4,864**	**$1**
By:			
capital (%)	35.9%	38.7%	2.9%
high-skilled labor (%)	16.8%	21.8%	5.0%
medium-skilled labor (%)	33.3%	30.3%	−3.0%
low-skilled labor (%)	14.0%	9.1%	−4.9%
In other countries (billion US$)	**$1,723**	**$3,820**	**$2,097**
By:			
capital (%)	55.2%	58.4%	3.2%
high-skilled labor (%)	5.4%	7.1%	1.7%
medium-skilled labor (%)	15.6%	17.0%	1.4%
low-skilled labor (%)	23.8%	17.5%	−6.3%
Worldwide (billion US$)	**$6,586**	**$8,684**	**$2,098**

Source: Authors' calculations based on World Input-Output Database, November 2013 Release.
Notes: Shares of production factors in total value added in a region, based on all global value chains of manufactures. Value added by a region is sum of value added by labor and capital on the domestic territory. High-income countries include Australia, Canada, and the United States; Japan, South Korea, and Taiwan; and all 15 countries that joined the European Union before 2004. Value added and expenditure is at basic prices (hence excluding net taxes, trade, and transport margins on output). It is converted to US dollars with official exchange rates and deflated to 1995 prices with the US Consumer Price Index. Figures may not add due to rounding.

Concomitant with this change in location of production, specialization patterns changed as well. In the traditional Heckscher–Ohlin model of trade, countries will focus on producing goods intensive in those factors that are relatively abundant. As a production chain fragments across countries, one might expect that the standard Heckscher–Ohlin predictions will still hold: the rise of China and other emerging economies accelerates the erosion of mature economies' comparative advantage in labor-intensive production tasks, while simultaneously offering new opportunities for offshoring (Hanson 2012). Thus, advanced countries will focus more on activities that require high-skilled labor and capital, and other countries will specialize in less-skilled activities.

To test these predictions, we provide more information on the factor content of global value chain production in the two regions in Table 3. The upper panel shows that in the high-income countries the share of capital increased from 36 to 39 percent, while the share of labor declined correspondingly. But the major income shift is observed across labor categories. The value added by high-skilled workers increased by 5 percentage points, while the combined share of medium- and

Table 4
Changes in Factor Shares over 1995–2008 in Global Value Chains of Manufactures, by Country
(in percentage points)

	Capital	*Low-skilled labor*	*Medium-skilled labor*	*High-skilled labor*
United States	3.9	−1.9	−5.9	4.0
Japan	4.5	−5.4	−2.1	3.1
Germany	6.8	−2.8	−7.4	3.4
France	0.2	−8.7	0.1	8.4
United Kingdom	−3.4	−8.0	1.2	10.2
Italy	−1.1	−14.8	10.4	5.5
Spain	0.1	−12.9	4.7	8.1
Canada	1.8	−2.0	−4.6	4.8
Australia	6.0	−8.4	−0.9	3.3
South Korea	9.3	−11.6	−5.6	8.0
Netherlands	5.5	−7.3	−7.1	8.9
Total all high-income	**2.9**	**−4.9**	**−3.0**	**5.0**
China	9.3	−9.3	−2.1	2.0
Russian Federation	1.1	−1.6	−2.4	2.8
Brazil	−6.7	−4.8	7.5	4.0
India	4.5	−5.9	−1.7	3.1
Mexico	6.4	−4.2	−0.5	−1.7
Turkey	−12.7	4.5	5.2	3.1
Indonesia	5.3	−8.1	1.3	1.6
World minus all high-income	**3.2**	**−6.3**	**1.4**	**1.7**
World	**6.5**	**−3.8**	**−4.2**	**1.5**

Source: Authors' calculations based on World Input-Output Database, November 2013 Release.
Notes: See Table 3. In this table, the percentage point changes in factor shares are given for each country. Changes in four factors for each country add up to zero by definition, but here they may not due to rounding. Countries are ranked by GDP.

low-skilled workers declined by 8 percentage points. The direction of this change is in line with the Heckscher–Ohlin intuition, but the magnitude of the changes differs across countries. In Table 4, we provide similar decompositions for individual countries. Looking first at the high-income group of countries, capital income shares increased in most countries, except in the United Kingdom and Italy, with the largest increases found in Germany and South Korea (7 and 9 percentage points). The value-added share by high-skilled workers increased in all countries in this group, ranging from around 3 percentage points in Australia, Germany, and Japan and 4 in the United States to more than 8 in France, the Netherlands, South Korea, and the United Kingdom. Income shares of other labor declined all around in the high-income countries. In Canada, Germany, and the United States, medium-skilled labor shares declined faster than low-skilled shares. In other countries like France, the United Kingdom, Italy, and Spain, as well as in South Korea and Australia, low-skilled workers' income shares suffered most, sometimes by more than 10 percentage points.

Declining incomes and jobs for less-skilled workers have stirred major policy concerns, mostly framed in terms of "manufacturing decline," and have prompted various initiatives for "re-industrialization" in a number of former industrial strongholds. Setting aside the merits of such proposals, it is important to note that with fragmented production, sectors like "manufacturing" are becoming the wrong way to evaluate economic performance and to frame public policies. Competitiveness is no longer solely determined by domestic clusters of manufacturing firms but relies increasingly also on the successful integration of other tasks in the chain, both domestic and foreign ones. To illustrate, the production of final manufactures involves not only jobs in the manufacturing sector but also jobs outside manufacturing that are indirectly related through the delivery of intermediate goods and services. In fact, in 2008 the latter made up almost half of all jobs related to manufactures production. Specialization in global value chains might therefore lead to declining jobs in traditional manufacturing but might also generate new jobs outside manufacturing. Indeed, in almost all high-income countries, the number of services jobs related to manufacturing production increased during this period, with the notable exceptions of the United Kingdom and the United States. In Germany and Italy, this increase was even faster than the decline in manufacturing jobs such that the net effect was positive (Timmer, Los, Stehrer, and de Vries 2013). Trade, labor, and industrial policies would do well to take into account the increased vertical integration of production within and across countries (Baldwin and Evenett 2012).

Trend 4: Enhanced Specialization in Capital in Emerging Economies

What happened to specialization patterns in the rest of the world? Based on the standard Heckscher–Ohlin predictions, one might expect the value-added share of less-skilled workers to increase in this region. This did not happen, as shown in the lower part of Table 3. The share of low-skilled workers declined by 6 percentage points from 24 percent in 1995 to 18 percent in 2008. The share of medium-skilled workers increased, but only by one percentage point. This is not to say that the number of workers employed in global value chains in manufacturing declined. On the contrary: 42 million jobs in China were added, 20 million in India, 6 million in Brazil, and 2 million in Mexico. (These figures are spelled out in Appendix Table 2, available online with this article at http://e-jep.org.) But in these countries as a whole, wages remained relatively low, and global value chain production mainly benefited capital. In 1995, the value-added share of capital in emerging economies was already high at 55 percent, compared to 36 percent in the high-income region. This is perhaps not surprising because these countries are abundant in labor, but it actually increased even further by 3 percentage points in the period up to 2008. The capital share in China increased by almost 10 percentage points. Capital shares in other major emerging economies like India, Indonesia, and Mexico also increased, by around 5 percentage points, as shown in Table 4. These developments fit a modern variant of the classical story of surplus labor by Lewis (1954). With capital being globally mobile, it will relocate to locations with high rental-wage ratios. As long as there is a reservoir of unskilled labor that can be employed at wages well below their marginal productivity, rental-wage ratios will remain high. Thus, the income share of capital will increase in early stages of development, rather than decline.

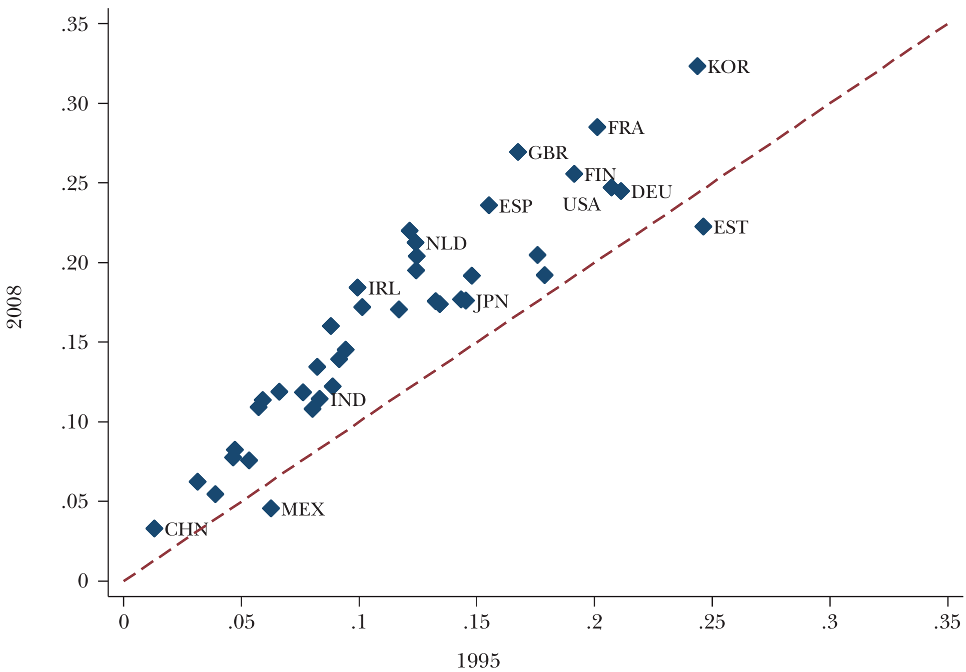
Table 4 also shows that the value-added share of high-skilled workers increased in almost all emerging economies. This echoes the changes that took place in Mexico when it entered into production chains with the United States in the 1980s (Feenstra and Hanson 1996). In their seminal model of offshoring (Feenstra and Hanson 1997), they related this to the establishment of so-called "maquiladoras" by US firms located across the US–Mexican border. Suppose that the good originally produced by the United States can be divided into two tasks. One task is relatively low-skilled intensive, like assembly of components, and the other task is high-skilled intensive, like producing high-tech components. As the relatively low-skilled task is offshored to Mexico, production in the United States will become more high skilled, further specializing in its abundant factor. Average skill intensity in Mexico increased after fragmentation in the 1980s. However, this is only one possible outcome, which will depend on the skill intensity of the offshored task compared to the existing skill intensity of production in the country (Arndt and Kierzkovski 2001; Feenstra 2010). It could also be that the average skill intensity of production would actually decrease rather than go up, as illustrated by more recent trends in Mexico; see Table 4. In fact, many outcomes are theoretically possible and to fully understand the complex patterns at work we need to further refine our thinking about the production process. In the final section, we sketch the main elements of such a framework.

Tasks, Substitution, Complements, and Technological Change

Production processes in manufacturing have increasingly fragmented across national borders, and the change in their factor content was clearly biased towards high-skilled labor and capital. This pattern was not only found for activities carried out in high-income countries, but also in emerging economies. In particular, the widespread increase in the value-added share of high-skilled workers was remarkable. In Figure 4, we plot for each of the 40 countries in the World Input-Output Database the share of value added by high-skilled workers for 1995 on the horizontal axis and for 2008 on the vertical axis. All observations, except two (Mexico and Estonia), are above the dotted 45-degree line, indicating a global shift towards use of relatively more high-skilled workers in global value chains in all of these countries.

What might account for this pattern? In traditional models of production, factor shares are determined by the interplay of relative prices of factors, their elasticities of substitution, and the nature of technical change. For example, opening up Asian economies led to a shock in the global supply of unskilled workers. Whether this change will lead to an increase in its factor share will depend on the elasticity of

Figure 4
Shares of High-Skilled Labor in Value Added of All Global Value Chains of Manufactures, by Country



Source: Authors' calculations based on World Input-Output Database, November 2013 Release.
Notes: Shares of high-skilled workers in a country's value added, based on all global value chains of manufactures, in 1995 (x-axis) and in 2008 (y-axis). The dashed line is the 45-degree line. Indicated are China (CHN), India (IND), Mexico (MEX), Ireland (IRL), Japan (JPN), the Netherlands (NLD), Spain (ESP), the United Kingdom (GBR), Finland (FIN), France (FRA), Germany (DEU), South Korea (KOR), Estonia (EST), and the United States (USA).

substitution between unskilled workers in Asia and elsewhere, but also on the substitution possibilities between unskilled and skilled workers, as well as between unskilled workers and capital. Another important element is the rapid advance in the information and communication technology industry, driving down the relative price of information technology capital (Jorgenson 2001). Again, the effects on the share of capital income will crucially depend on the substitution possibilities between information technology capital on the one hand, and various types of labor on the other.

Substitution possibilities are hard to model and measure. Archetype models of growth and international trade rely on production functions where elasticities of substitution are rather restricted.[5] In these models, the production process is conceived of as a mapping from factor inputs to output, as if taking place in one

[5] The most often-used production functions are the so-called Cobb–Douglas and the constant elasticity of substitution (CES) functions. In the Cobb–Douglas function, elasticities are always one. Hence factor

stage. With fragmentation, however, it can be more useful to model the generation of output as a result of a set of "tasks" which are to be completed by various combinations of production factors. So rather than a direct mapping from labor and capital inputs to output, factors map into tasks, which subsequently map into output. This framework allows for a richer modeling of complementarities and substitution possibilities between various factors of production, both domestic and foreign.

An example of this is found in recent models of labor demand discussed in Acemoglu and Autor (2011). They outline a general framework that revolves around differences in comparative advantages of factors in carrying out tasks: all workers can carry out all tasks, but some are relatively better at carrying out certain tasks (hence are said to have a comparative advantage in this task). Substitution of skills across tasks is possible, such that there is an endogenous mapping from workers to tasks depending solely on labor supplies and the comparative advantages of the various skill types. The framework also allows for capital as an input, by modeling it as another source competing with labor for the supplying of certain tasks. For example, new information technology capital might be much better in handling routine administrative tasks than skilled white-collar labor. According to the "routinization hypothesis" put forward by Autor, Levy, and Murnane (2003), information technology capital complements highly educated workers engaged in abstract tasks, substitutes for moderately educated workers performing routine tasks, and has little effect on less-skilled workers performing manual tasks and tasks that require personal interactions, such as in many services. The latter tasks are less important in manufacturing global value chains, which is consistent with our observation that income shares for both low- and medium-skilled workers in manufactures are declining (Foster-McGregor, Stehrer, and de Vries 2013).

The increasing importance of intangible capital provides another potential explanation for the increasing value-added shares of capital and high-skilled workers. Recent investment in advanced countries is increasingly directed towards intangibles such as intellectual capital (including software and databases, research and development, and design), brand names, and organizational firm-specific capital (Corrado, Haskel, Jona-Lasinio, and Iommi 2012). To the extent that the build-up of intangibles requires high-skilled labor, this will increase demand for the latter. In an extended Heckscher–Ohlin framework, Haskel, Lawrence, Leamer, and Slaughter (2012) assume that skilled workers are more productive in tasks involving working with intangible capital and show how this might explain the evolution of relative wages in the United States. Moreover, intangibles like patents or trademarks are different from traditional capital assets as they typically have a large fixed-cost component. This often gives rise to imperfect product markets and possibilities for mark-ups. When firms operating in such an environment enlarge

cost shares cannot change over time. In the CES function, elasticities are also constant over time, but might vary from one. However, in cases of more than two factor inputs, they are difficult to define.

their scale of operations, capital is likely to gain more relative to labor. In a dynamic model of growth, increased openness and trade might reinforce higher levels of investment in intangibles as it expands the incentives for their creation: the larger the market in which the new invention will be used, the higher the potential for profits accruing to the investor.

Concluding Remarks

International production fragmentation is underway. The patterns of specialization found in case studies like the iPod have a macroeconomic equivalent. Our findings fit a broad story in which firms in mature economies relocate their unskilled-labor-intensive production activities to lower-wage countries, while keeping strategic and high-value-added functions concentrated at home where the skilled workers and intangible capital they need are available. But this shift of activities was decidedly non-neutral: capital shares in value added increased in both high-income and emerging economies. Further, we found declining value-added shares of low-skilled workers in emerging economies, contradicting traditional notions of comparative advantage. Squaring these facts will be an interesting challenge for further research. One possible explanation is a shift in manufacturing technologies that could have led to a worldwide decline in the demand for unskilled workers. This question can only be investigated from a global value chain perspective as analyses focusing on industries cannot distinguish between offshoring and technological change.

More generally, the impact of trade and cross-border investments on the distribution of income across and within countries have been extensively debated (for an overview, see Harrison, McLaren, and McMillan 2011). In essence, international fragmentation expands the opportunities of countries to specialize according to comparative advantage and hence to gain from trade. As such, it is on average welfare improving, but not necessarily for all workers and owners of capital. We believe that the trade-offs involved can be better understood by conceptualizing the production process as a set of tasks to be performed by combinations of factor inputs. For example, Costinot, Vogel, and Wang (2012) develop a model in which heterogeneous workers sort themselves into various stages of the production process. They find that the consequences of opening up to trade on wage inequality may be very different from standard models, depending on the position of the workers in the chain. In particular, they find that in the less-advanced country all workers move to upstream stages of production, decreasing wage inequality at the bottom of the skill distribution but at the same time increasing it at the top.

Many outcomes are theoretically possible, and it becomes ultimately an empirical issue as to which patterns prevail. The development of world input-output tables is a first step in this investigation. Future statistical frameworks, based on further integration of micro- and macro-statistics will allow for increasingly richer explanations of the drivers and consequences of international production fragmentation.

■ *We thank Richard Baldwin, Lucian Cernat, Carol Corrado, Hubert Escaith, Robert Feenstra, Susan Houseman, Robert Johnson, Dale Jorgenson, Bob Koopman, Bart Van Ark, Zhi Wang, Adrian Wood, and the editors of this journal for helpful comments and suggestions. Reitze Gouma provided superb research assistance. Financial support for this research was provided by the European Commission, Research Directorate General as part of the 7th Framework Program, Theme 8: Socio-Economic Sciences and Humanities, Grant Agreement no: 225 281.*

References

Acemoglu, Daron, and David H. Autor. 2011. "Skills, Tasks and Technologies: Implications for Employment and Earnings." In *Handbook of Labor Economics,* Volume 4B edited by David Card and Orley Ashenfelter, 1043–1171. Amsterdam: Elsevier.

Ali-Yrkkö, Jyrki, Petri Rouvinen, Timo Seppälä, and Pekka Ylä-Anttila. 2011. "Who Captures Value in Global Supply Chains? Case NOKIA N95 Smartphone." *Journal of Industry, Competition, and Trade* 11(3): 263–78.

Ando, Mitsuyo, and Fukunari Kimura. 2005. "The Formation of International Production and Distribution Networks in East Asia." In *International Trade in East Asia, NBER–East Asia Seminar on Economics,* Vol. 14, edited by Takatoshi Ito and Andrew K. Rose, 177–216. University of Chicago Press.

Antràs, Pol, Davin Chor, Thibault Fally, and Russell Hillberry. 2012. "Measuring the Upstreamness of Production and Trade Flows." *American Economic Review* 102(3): 412–16.

Arndt, Sven W., and Henryk Kierzkowski, eds. 2001. *Fragmentation: New Production Patterns in the World Economy*. Oxford University Press.

Autor, David H., Frank Levy, and Richard J. Murnane. 2003. "The Skill Content of Recent Technological Change: An Empirical Exploration." *Quarterly Journal of Economics* 118(4): 1279–1333.

Baldwin, Richard E., and Simon J. Evenett. 2012. "Value Creation and Trade in 21st Century Manufacturing: What Policies for UK Manufacturing?" In *The UK in a Global World,* edited by David Greenaway, 71–128. London: Centre for Economic Policy Research.

Baldwin, Richard, and Anthony J. Venables. 2013. "Spiders and Snakes: Offshoring and Agglomeration in the Global Economy." *Journal of International Economics* 90(2): 245–54.

Corrado, Carol, Jonathan Haskel, Cecilia Jona-Lasinio, and Massimiliano Iommi. 2012. "Intangible Capital and Growth in Advanced Economies: Measurement and Comparative Results." CEPR Discussion Papers 9061.

Costinot, Arnaud, Jonathan Vogel, and Su Wang. 2012. "Global Supply Chains and Wage Inequality." *American Economic Review* 102(3): 396–401.

Dedrick, Jason, Kenneth L. Kraemer, and Greg Linden. 2010. "Who Profits From Innovation in Global Value Chains? A Study of the iPod and Notebook PCs." *Industrial and Corporate Change* 19(1): 81–116.

Dietzenbacher, Erik, Bart Los, Robert Stehrer, Marcel P. Timmer, and Gaaitzen de Vries. 2013. "The Construction of World Input–Output Tables in the WIOD Project." *Economic Systems Research* 25(1): 71–98.

Dietzenbacher, Erik, and Isidoro Romero. 2007. "Production Chains in an Interregional Framework: Identification by Means of Average Propagation Lengths." *International Regional Science Review* 30(4): 362–83.

Feenstra, Robert C. 1998. "Integration of Trade and Disintegration of Production in the Global Economy." *Journal of Economic Perspectives* 12(4): 31–50.

Feenstra, Robert C. 2010. *Offshoring in the Global Economy: Microeconomic Structure and Macroeconomic Implications.* Cambridge: MIT Press.

Feenstra, Robert C., and Gordon H. Hanson. 1996. "Foreign Investment, Outsourcing, and Relative Wages." In *The Political Economy of Trade*

Policy: Papers in Honor of Jagdish Bhagwati, edited by Robert C. Feenstra, Gene M. Grossman, and Douglas A. Irwin, 89–127. Cambridge: MIT Press.

Feenstra, Robert C., and Gordon H. Hanson. 1997. "Foreign Direct Investment and Relative Wages: Evidence from Mexico's Maquiladoras." *Journal of International Economics* 42(3–4): 371–93.

Foster-McGregor, Neil, Robert Stehrer, and Gaaitzen J. de Vries. 2013. "Offshoring and the Skill Structure of Labour Demand." *Review of World Economics* 149(4): 631–62.

Fukao, Kyoji, Hikari Ishido, and Keiko Ito. 2003. "Vertical Intra-Industry Trade and Foreign Direct Investment in East Asia." *Journal of the Japanese and International Economies* 17(4): 468–506.

Gereffi, Gary. 1999, "International Trade and Industrial Upgrading in the Apparel Commodity Chain." *Journal of International Economics* 48(1): 37–70.

Hanson, Gordon H. 2012. "The Rise of Middle Kingdoms: Emerging Economies in Global Trade." *Journal of Economic Perspectives* 26(2): 41–64.

Hanson, Gordon H., Raymond J. Mataloni, and Matthew J. Slaughter. 2005. "Vertical Production Networks in Multinational Firms." *Review of Economics and Statistics* 87(4): 664–78.

Harrison, Ann, John McLaren, and Margaret McMillan. 2011. "Recent Perspectives on Trade and Inequality." *Annual Review of Economics* 3, pp. 261–89.

Haskel, Jonathan, Robert Z. Lawrence, Edward E. Leamer, and Matthew J. Slaughter. 2012. "Globalization and U.S. Wages: Modifying Classic Theory to Explain Recent Facts." *Journal of Economic Perspectives* 26(2): 119–40.

Hummels, David, Jun Ishii, and Kei-Mu Yi. 2001. "The Nature and Growth of Vertical Specialization in World Trade." *Journal of International Economics* 54(1): 75–96.

Johnson, Robert C., and Guillermo Noguera. 2012a. "Accounting for Intermediates: Production Sharing and Trade in Value Added." *Journal of International Economics* 86(2): 224–36.

Johnson, Robert C., and Guillermo Noguera. 2012b. "Fragmentation and Trade in Value Added over Four Decades." NBER Working Paper 18186.

Jorgenson, Dale W. 2001. "Information Technology and the U.S. Economy." *American Economic Review* 91(1): 1–32.

Koopman, Robert, Zhi Wang, and Shang-Jin Wei. 2014. "Tracing Value-Added and Double Counting in Gross Exports." *American Economic Review* 104(2): 459–94.

Krugman, Paul R. 1995. "Growing World Trade: Causes and Consequences." *Brookings Papers on Economic Activity* 26(1): 327–77.

Leontief, Wassily W. 1936. "Quantitative Input-Output Relations in the Economic System of the United States." *Review of Economics and Statistics* 18(3): 105–25.

Lewis, W. Arthur. 1954. "Economic Development with Unlimited Supplies of Labour." *The Manchester School* 22(2): 139–91.

Lipsey, Robert E. 2010. "Measuring the Location of Production in a World of Intangible Productive Assets, FDI and Intrafirm Trade." *Review of Income and Wealth* 56(s1): S99–S110.

Los, Bart, Marcel P. Timmer, and Gaaitzen J. de Vries. Forthcoming. "How Global Are Global Value Chains? A New Approach to Measure International Fragmentation." *Journal of Regional Science.*

Marin, Dalia. 2011. "The Opening Up of Eastern Europe at 20: Jobs, Skills, and 'Reverse Maquiladoras.'" In *International Handbook on the Economics of Integration,* edited by Miroslav N. Jovanovic. Vol. 2, Chap. 13, pp. 293–323. Edward Elgar.

Miller, Ronald E., and Peter D. Blair. 2009. "Input-Output Analysis: Foundations and Extensions." Cambridge University Press.

O'Mahony, Mary, and Marcel P. Timmer. 2009. "Output, Input and Productivity Measures at the Industry Level: The EU KLEMS Database." *Economic Journal* 119(538): F374–F403.

Rodrik, Dani. 1997. "Has Globalization Gone Too Far?" Washington, DC: Institute for International Economics.

Timmer, Marcel P., Erik Dietzenbacher, Bart Los, Robert Stehrer, and Gaaitzen J. de Vries. 2014. "The World Input-Output Database (WIOD): Contents, Concepts and Applications." *GGDC Research Memorandum* 144, Groningen Growth and Development Centre.

Timmer, Marcel P., Bart Los, Robert Stehrer, and Gaaitzen J. de Vries. 2013. "Fragmentation, Incomes and Jobs. An Analysis of European Competitiveness." *Economic Policy* 28(76): 613–61.

Journal of Economic Perspectives—Volume 28, Number 2—Spring 2014—Pages 119–142

Five Facts about Value-Added Exports and Implications for Macroeconomics and Trade Research

Robert C. Johnson

International trade data record the gross value of goods as they cross borders. This poses a challenge for researchers who want to connect canonical international trade and macroeconomic models, which are typically cast in value-added terms, with the data. The most common approach has been to treat gross trade data *as if* it is comparable to data on value added. In the past, this assumption was tolerable. Vertical specialization in trade—that is, the use of imports to produce exports—was limited in most countries (Hummels, Ishii, and Yi 2001). In other words, gross exports contained very nearly 100 percent domestic value added.

In recent decades, the emergence of global supply chains has changed matters. As inputs pass through these chains, they cross borders many times. This means that gross trade data include substantial double-counting, so gross exports overstate the amount of domestic value-added in exports. When supply chains span multiple countries, it also means that bilateral gross exports do not tell us where the value added embodied in those exports is ultimately consumed. As a result, gross trade is an increasingly misleading guide to how value added is exchanged between countries.

This realization has prompted concerns that gross trade data distort perceptions about the nature of international integration and the role of particular countries in international markets, which in turn leads to tensions in the world trade system. Lamy (2011), for example, emphasizes that a clearer view of how countries are linked together via global supply chains breaks down mercantilist ("us" versus "them") views of trade. Prompted by these real world concerns, along with the

■ *Robert C. Johnson is Assistant Professor of Economics, Dartmouth College, Hanover, New Hampshire. He is also a Faculty Research Fellow, National Bureau of Economic Research, Cambridge, Massachusetts. His email address is robert.c.johnson@dartmouth.edu.*

http://dx.doi.org/10.1257/jep.28.2.119 doi=10.1257/jep.28.2.119

desire to measure trade in a manner consistent with the value-added thinking that underlies popular models, there has been a recent push toward developing new value-added measures of trade.

To measure trade in value added, we need to follow goods through the global supply chain from input producers to final consumers, allocating the value added in final goods to producers at each stage. Of course, this is easier said than done. While national input-output accounts describe domestic supply chains, they stop at the border. To overcome this problem, recent work has combined national input-output tables with bilateral trade data to construct input-output tables with global scope. These global input-output tables describe input shipments across both sectors and countries, and hence enable us to trace the value added embodied in final goods back to its source. With this new data, we can measure the hidden trade in value added underlying gross trade.

In this paper, I highlight one measure of trade in value added—"value-added exports." Value-added exports measure the amount of domestic value added embodied in final expenditure in each destination (Johnson and Noguera 2012a). Just as gross exports break down gross output sold across destinations, value-added exports break down GDP sold across destinations. This value-added export concept is the appropriate measure of exports in international models that are written in value-added terms.

After describing how value-added exports are computed, I summarize five key facts about differences between gross and value-added exports. First, these differences are large and growing over time, currently around 25 percent. Second, manufacturing trade looks more important (relative to services) in gross than value-added terms. Third, these differences are heterogenous across countries, with value-added exports ranging from 50 percent (Taiwan) to 90 percent (Russia) of gross exports. Fourth, the differences between gross and value-added exports are heterogeneous across bilateral partners, with even more variation across partners than across individual countries. Fifth, these differences are changing unevenly over time across countries and partners, with fast-growing emerging markets and pairs of countries that adopt bilateral trade agreements seeing larger declines in value-added relative to gross exports.

Taking these five facts into account points researchers toward better quantitative answers to important macroeconomic and trade questions. To illustrate this point, I discuss how value-added exports can be applied in analysis of some widely discussed questions. In macroeconomics, value-added exports help quantify the strength of demand spillovers, the consequences of relative price movements for competitiveness, and the size of relative price changes needed to close trade imbalances. In trade, value-added exports can be applied in analysis of the impact of frictions on trade, the role of endowments and comparative advantage in trade, and trade policy.

Background: Computing Value-Added Exports

A basic fact of national income accounting is that expenditure on final goods equals the amount of value added generated during the production process.

Therefore, final expenditure directly tells us how much value added is consumed in each country. But the national accounts do not tell us where that value added comes from.

For specific goods (like iPods or notebook PCs), we can try to decompose the value added embodied in them across countries by breaking them apart and examining their constituent parts (Linden, Kraemer, and Dedrick 2009; Debrick, Kraemer, and Linden 2010). This deconstructive approach is conceptually straightforward, but complicated in practice. One reason is that the production process has many layers. It is not enough to break the iPod into component parts (for example, the screen, disk drive, plastic shell); one needs to also break down those components into subcomponents (metal, plastic, and so on). Pushing further, even the subcomponents need to be further broken down until one knows where the value added in the metal, plastic, and other basic inputs originates.

The goal of this process is to be able to make statements like "one third of the $299 value of an iPod sold in the United States is Japanese value added."[1] Though this value added is produced in Japan, it is consumed in the United States. As a matter of definition, we then say that Japan exports roughly $100 of value added to the United States as part of the iPod production process.

Implementing this approach on a good-by-good basis and then aggregating up to produce aggregate value-added export data is nigh impossible. Nonetheless, the basic logic of this good-by-good calculation can be adapted to track value added in the aggregate. To see this, it is useful to think of the process in two steps.

The first step is to measure how much output from each source country is needed to produce the final goods that are consumed in a given destination. For example, how much Japanese gross output (disk drives, metal, and everything else) is needed to produce final goods (iPods) consumed in the United States? In this step, we need to know not only how many Japanese disk drives are used, but also how much Japanese metal and plastic are used in production of those disk drives.

The second step is to measure how much local value added is generated in production of that gross output. That is, how much Japanese value added is generated in assembling the disk drives, plus how much Japanese value added is embodied in the metal used?

To implement this two-step approach economy-wide, we need to describe the sector-level production process in a manner analogous to how we described the production process for individual goods. That is, we need to measure the value of final goods purchased from each source country and measure input use and value-added contributions along the production chain. To do this, we turn to a

[1] This estimate is based on Linden, Kraemer, and Dedrick (2009). It is the value of iPod components from Japanese-headquartered companies (for example, the hard drive from Toshiba) divided by the sales price of the iPod. However, this estimate does not actually measure true value added by Japanese suppliers. First, it does not identify where Toshiba produces the hard drive, which determines the country in which value added is recorded. Second, it only captures the last layer of the production process. For example, it does not identify whether Toshiba uses imported inputs to produce the hard drive. To my knowledge, no product case study has yet been able to address these problems.

global input-output framework. On the input side, global input-output tables record the sectors and countries from which inputs are sourced to produce output in a given country and sector. On the output side, they record the destinations to which final goods from each sector are shipped. Combining these, we can take final goods shipments and trace backwards using input requirements to allocate the value added in those final goods to their source.

For example, suppose we see final goods being shipped from US manufacturers to Canadian final consumers. Then, if we know the sector and country origin of inputs used in US manufacturing (for example, inputs from Japan used in US manufacturing), this is analogous to knowing the iPod's components. Building on this, if we also know where those input suppliers get their own inputs (for example, inputs from China used by Japan), this is analogous to knowing the breakdown of components into subcomponents. Further, the input-output accounts record how much value added is generated in producing output in each country and sector, which enables us to convert gross production at each stage to value added. In this way, applying input-output accounting principles, we approximate the iPod accounting exercise for the economy as a whole.[2]

The main challenge in implementing this approach lies in assembling the data needed to form the global input-output framework. In an ideal world, national statistical authorities would coordinate to produce these input-output accounts. As a second best, various researchers and organizations have assembled synthetic input-output tables from existing national accounts and trade data. The basic procedure uses bilateral trade data to split sector-level multilateral final and intermediate goods imports, which are reported in official input-output tables, across source countries. The result is a global input-output table, which describes bilateral final and intermediate input use. Table 1 lists several public use datasets that contain national input-output tables, global or regional input-output tables, or value-added trade data.

Not surprisingly, different research teams have used varying data sources and assumptions in constructing these global input-output tables. I will not dwell here on the many different choices that have been made by various researchers. Rather, I want to highlight that there is tremendous agreement across alternative data sets about how value-added exports compare to gross exports. The core facts that I discuss below are robust across alternative datasets.

[2] To sketch the underlying math, suppose we observe a global input-output matrix, denoted A, which is a square matrix of input use requirements with dimensions equal to the number of countries times the number of sectors. The columns of this matrix describe input requirements for producing gross output in each country and sector, with elements equal to the value of inputs purchased from a particular source country and sector as a share of gross sector-level output in the destination. The "Leontief inverse" of the global input-output matrix, given by $(I - A)^{-1}$, tells us how much output from each source country and sector is needed to produce any vector of final goods, where final goods are identified by sector and country source. To convert these gross output requirements into value added, multiply by value added to output ratios in the source country and sector, which can be obtained by taking one minus the column sums of A. See Johnson and Noguera (2012a) for details.

Table 1
Public Datasets for Research on Value-Added Exports

Name of dataset	*Key features*	*Selected research using this data*
Global Trade Analyis Project Database	Input-output tables for over 100 countries for various benchmark years, mostly after 2000. https://www.gtap.agecon.purdue.edu	Trefler and Zhu (2010), Daudin, Rifflart, and Schweisguth (2011), Johnson and Noguera (2012a), and Koopman, Wang, and Wei (2014)
World Input-Output Database	Global tables covering OECD countries and major emerging markets from 1995–2011. http://www.wiod.org	Baldwin and Lopez-Gonzales (2013), Costinot and Rodríguez-Clare (2013), Timmer, Los, Stehrer, and de Vries (2013)
IDE-JETRO Asian Input-Output Tables	Regional tables covering 8 East Asian countries at five-year intervals between 1985 and 2000. http://www.ide.go.jp	Various chapters in Hiratsuka and Uchida (2010), IDE-JETRO and WTO (2011), Puzzello (2012)
WTO-OECD TiVA Database (Trade in Value Added)	Value-added exports and other measures of global supply chain activity for 57 countries in 1995, 2000, 2005, 2008 and 2009. http://stats.oecd.org	De Backer and Miroudot (2013)
OECD Input-Output Tables	Input-output tables for OECD countries and major emerging markets, available various years from 1970–2005. http://www.oecd.org/trade/input-outputtables.htm	Hummels, Ishii, and Yi (2001), Johnson and Noguera (2012b, 2014)

Five Facts about Value-Added Exports

This section reviews five high-level facts about how value-added exports compare to gross exports for the world as a whole, across sectors, across countries, and across bilateral trade partners.

Fact 1: World value-added exports are equal to about 70–75 percent of gross exports today, down from about 85 percent in the 1970s and 1980s.

Recent estimates suggest that value-added exports are equal to 70–75 percent of the value of gross exports. Using the World Input-Output Database, the ratio of value-added to gross exports was about 0.71 in 2008. Johnson and Noguera (2014) put it at about 0.76 in the same year. Johnson and Noguera (2012a) report that the median ratio of value-added to gross exports across 94 countries was 0.73 in 2004. Despite differences in underlying data and methods, these estimates lie within a comfortingly narrow range.

The ratio of value-added to gross trade has declined over time, down from around 85 percent in the early 1970s (Johnson and Noguera 2014). This decline implies that there is more double counting in gross trade data now than in the past. This increased double counting is symptomatic of the growing importance of global supply chains in mediating trade, as goods cross borders more than once when supply chains span multiple countries.[3]

One important feature of the data is that the decline in value-added relative to gross exports occurs almost entirely after 1990 (Johnson and Noguera 2014). This decline coincides with rapid changes in the world economy: trade liberalization in emerging markets, the expansion of the European Union, the adoption of major regional trade agreements, and the information technology revolution. Writ large, these events lowered international trade costs, induced substitution of foreign for domestic input suppliers, and drove down the value-added content of trade.

Fact 2: Manufacturing trade is relatively smaller, and services trade relatively larger, when measured in value-added terms.

For the world as a whole, manufacturing accounts for nearly 70 percent, and services account for 20 percent, of total gross exports. In contrast, manufacturing and services both account for about 40 percent of total value-added exports. This reallocation of trade shares is illustrated in Figure 1.

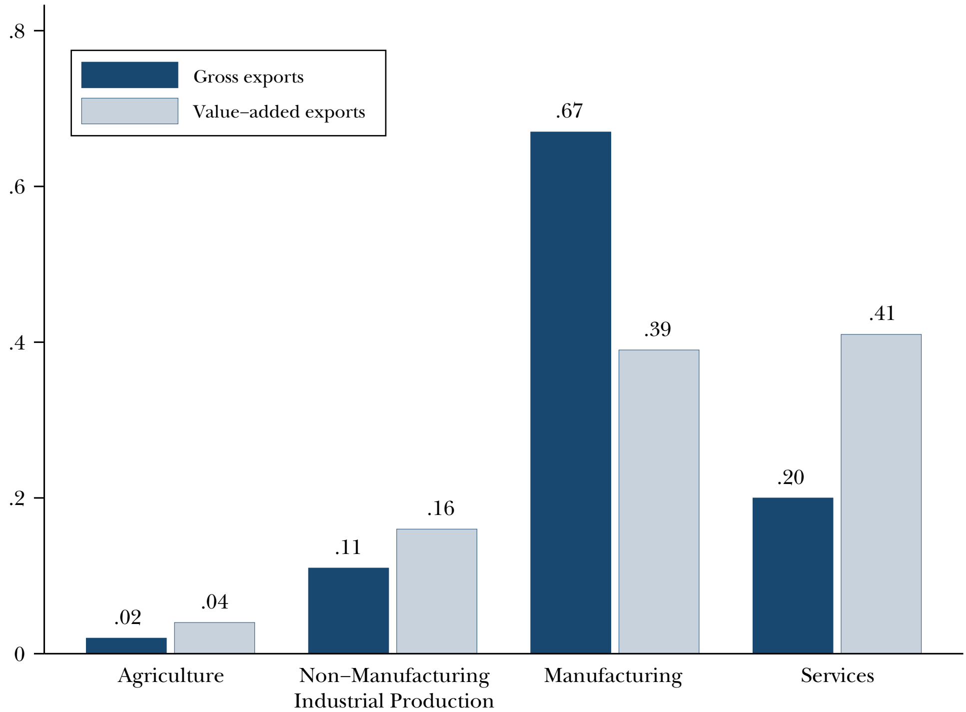
Put differently, the ratio of value-added to gross exports is lower for manufacturing than services trade. There are two reasons for this. First, gross manufacturing exports include value added from the services sector, because manufacturing firms buy services as inputs. The value-added export data strip this services value added out of manufacturing exports and reassign it to the services sector. Second, manufacturing features a higher degree of vertical specialization than services (that is, a higher import content of exports), which pushes down the ratio of value-added to gross exports in manufacturing relative to other sectors.

Fact 3: Across countries, value-added exports range from 50 to 90 percent of the value of gross exports.

There is wide variation across countries in the ratio of value-added to gross exports. Table 2 reports the ratio of value-added to gross exports in 2008 for the top 20 exporting countries, computed using the World Input-Output Database. Among these countries, the range is roughly 0.5 to 0.9. Using a broader 89 country sample from the Global Trade Analysis Project database, in Johnson and Noguera (2012a), we document a 10th–90th percentile spread of about 0.6 to 0.85 in 2004.

[3] Fally (2012) shows that the inverse of the world value-added to gross export ratio can be interpreted as a weighted average count of the number of border crossings associated with producing $1 of final goods, where the weights reflect the value added by each country.

Figure 1
Sector Shares in Total World Value-Added and Gross Exports



Sources: World Input-Output Database (WIOD) and author's calculations.
Notes: Data are for 2008. Agriculture includes Forestry, Hunting, and Fishing. Non-Manufacturing Industrial Production includes Mining and Quarrying, Electricity/Gas/Water Supply, and Construction. Manufacturing is the remainder of Industrial Production.

Among other determinants, the ratio of value-added to gross exports is strongly negatively correlated with the share of manufacturing in total exports (Johnson and Noguera 2012a). This observation relates back to Fact 2. The ratio of value-added to gross exports is lowest in manufacturing, so the composition of trade matters.

Fact 4: Gaps between bilateral value-added and gross exports are large and heterogeneous across trade partners.

Table 3 reports the ratio of value-added to gross exports for the top four exporting countries for alternative destination countries and composite regions. Though regional aggregation obscures many interesting bilateral details, it serves to highlight some key aspects of the data.

First, there is as much variation across bilateral partners as there is across source countries. For Germany, the ratio of value-added to gross exports ranges

Table 2

The Ratio of Value-Added to Gross Exports for the Top 20 Exporting Countries

	WIOD 2008	*WIOD Change 1995–2008*	*Johnson–Noguera Change 1970–2008*
Germany	0.69	−0.10	−0.16
United States	0.78	−0.05	−0.14
China	0.75	−0.09	−0.20
Japan	0.80	−0.12	−0.09
United Kingdom	0.78	−0.01	−0.04
France	0.71	−0.08	−0.13
Italy	0.73	−0.07	−0.12
Netherlands	0.62	−0.06	−0.11
Canada	0.76	0.02	−0.11
South Korea	0.58	−0.18	−0.18
Russia	0.92	0.00	
Belgium	0.53	−0.07	−0.15
Spain	0.69	−0.09	−0.17
Taiwan	0.51	−0.16	
Mexico	0.70	−0.03	−0.21
India	0.78	−0.12	−0.20
Sweden	0.66	−0.08	−0.13
Australia	0.84	−0.04	−0.06
Brazil	0.86	−0.05	−0.10
Austria	0.65	−0.10	−0.17
Minimum	0.51	−0.18	−0.21
Median	0.72	−0.08	−0.14
Maximum	0.92	0.02	−0.04

Sources: World Input-Output Database (WIOD) and author's calculations, Johnson and Noguera (2014).
Notes: The column "WIOD 2008" is the ratio of value-added exports to gross exports for each country in 2008 from the World Input-Output Database. The column "WIOD change 1995–2008" is the change in this ratio from 1995 to 2008. The column "Johnson–Noguera change 1970–2008" is the change in the ratio of value-added exports to gross exports for each country from 1970 to 2008, from Johnson and Noguera (2014). Blank entries in that column reflect missing data. Exporting countries are ordered top to bottom by total gross exports in 2008.

from 0.6 to 1 across destinations. Second, value-added exports to some destinations exceed gross exports. For example, Japanese value-added exports to the United States are 7 percent larger than their gross exports. This reflects the fact that Japan exports intermediate goods to third countries (such as China) that then re-exports those intermediates to the United States embodied in final goods. Third, the ratio of value-added to gross exports tends to be lower within regions or regional trade agreement blocs than across them. For example, US value-added exports are 64 percent as large as gross exports to Mexico and Canada, while US value-added exports are about 90 percent as large as gross exports to the European Union or Japan. Similar patterns hold for Japan and Germany among their Asian and European Union partners, respectively.

Table 3
Ratio of Bilateral Value-Added to Gross Exports for Top 4 Exporting Countries

	Partner country or region						
Source country	*United States*	*Canada and Mexico*	*European Union*	*China*	*Japan*	*Other Asia*	*Other*
China	0.84	0.71	0.79		0.73	0.52	0.73
Germany	0.99	0.80	0.60	0.77	1.00	0.70	0.74
Japan	1.07	0.86	1.06	0.69		0.53	0.76
United States		0.64	0.87	0.83	0.91	0.69	0.79

Sources: World Input-Output Database (WIOD) and author's calculations.
Notes: Data are for 2008. "Other Asia" includes Indonesia, South Korea, and Taiwan. "Other" includes all other destinations not listed in table.

Fact 5: Changes in value-added relative to gross exports have been heterogeneous across countries and bilateral trade partners.

Table 2 also reports changes in the ratio of value-added to gross exports for each of the top 20 exporters for two time periods. Column 2 records the change in this ratio over the 1995–2008 period, while column 3 records changes over the longer 1970–2008 period. To summarize, some countries have seen declines on the order of 20 percentage points, while others have seen no change—or even increases. In general, declines have been larger in fast-growing emerging markets than other countries, largely due to the rapid increase in the share of manufactures in their gross exports over time (Johnson and Noguera 2014).

Drilling down to the bilateral level, changes in the ratio of value-added to gross exports are also very different across bilateral trade partners. One stylized fact is that the ratio has declined more for nearby countries and countries within the same region (Johnson and Noguera 2012b, 2014). A second fact is that the ratio has declined more for countries that have adopted regional trade agreements with one another (Johnson and Noguera 2014).

International Macroeconomics

Using value-added export data in place of gross exports sheds new light on some old questions in international macroeconomics. Here, I consider three of those questions. First, how large are the spillover effects of changes in foreign final expenditure on domestic economic activity? Second, how do international relative price changes—for example, due to exchange rate movements—influence competitiveness? Third, how large must price changes be in order to close trade imbalances?

Tracking Foreign Expenditure Changes Back Home

How much does US GDP fall when foreign final expenditure falls? Would the US economy be hit harder by a fall in expenditure in Italy or Canada? To answer these questions, analysts traditionally look at US gross exports, and assume that those exports are produced entirely within the United States. They use the share of multilateral or bilateral exports in GDP to summarize the exposure of the US economy to foreign expenditure changes.

A value-added perspective on trade highlights several flaws in this approach. First, a dollar of US exports does not generate a dollar of US value added. As a result, the ratio of exports to GDP will overstate how much GDP falls when exports decline. Second, bilateral gross exports do not capture how much value added the United States sells in particular destinations. For example, a significant share of US exports to Canada are used to produce Canadian goods consumed in the United States, so gross exports overstate US exposure to Canadian demand shocks. Alternatively, gross exports may understate exposure in other cases. For example, the US exports inputs to Germany that are used to produce German goods consumed in Italy. Thus, the US economy is more exposed to changes in Italian demand than gross exports would indicate.

Looking directly at value-added exports side-steps these problems. Value-added exports directly link foreign final expenditure to demand for domestic value added, removing gross exports as the "middle man" in the calculation. Though this intuition is straightforward, explaining how it emerges directly from standard macro-models takes some additional effort. There are two alternative theoretical approaches.

The first approach is to write down the model entirely in value-added terms, ignoring trade in intermediate inputs entirely. Though this approach may initially sound strange, it is in fact completely standard—for example, the canonical international real business cycle fits this description (Backus, Kehoe, and Kydland 1994). On the supply side, producers combine primary factors (labor and capital) to produce value added. On the demand side, consumers directly purchase and consume value added originating from different source countries. Given this structure, value-added exports are the appropriate data to use in measuring trade and calibrating preference parameters.[4] And the share of bilateral value-added exports in total value added is the appropriate weight to use in estimating how much demand for domestic value added falls in response to changes in foreign final expenditure.

The second approach is to embrace input trade, and write down the model in gross terms (Ambler, Cardia, and Zimmerman 2002; Johnson forthcoming). In this case, producers would combine primary factors with intermediate inputs to produce gross output, which may be dedicated to either final or intermediate use. And preferences would be defined over consumption of final goods. In Bems

[4] This observation is closely related to a recent point raised by Herrendorf, Rogerson, and Valentinyi (2013). In a closed economy, they argue that expenditure on value added from each sector, rather than expenditure on final goods from each sector, should be used to calibrate preferences in multisector models that feature value-added production functions for sectoral output.

and Johnson (2012), we show that value-added export shares are the appropriate weights to attach to foreign final expenditure changes in this type of model as well.[5]

Using value-added exports in place of gross exports has three implications. First, all countries appear less exposed to foreign expenditure changes, many substantially so. Remember, the ratio of value-added to gross exports is less than one, and these adjustments are getting larger over time due to declines in value-added to export ratios. Second, at the sector-level, the manufacturing sector looks substantially less exposed, and nonmanufacturing sectors look substantially more exposed to foreign shocks, because manufacturing exports are smaller and services exports larger in value-added terms. Third, the importance of shocks originating in particular export destinations differs—with some countries becoming more important, while others are becoming less important, than one would guess based on gross bilateral exports. This follows from differences in bilateral value-added to export ratios across partners.

To illustrate the magnitude of these adjustments, I graph the ratios of gross and value-added exports to GDP for the top four exporters in Figure 2. Aggregating across sectors, the ratio of value-added exports to GDP is generically smaller than the ratio of gross exports to GDP. At the sector level, the ratio of value-added exports from the manufacturing sector to manufacturing GDP is dramatically smaller than the ratio of gross exports from the manufacturing sector to manufacturing GDP, about half as large for these countries. Further, differences in openness across sectors are reduced when measured using the ratio of value-added exports to GDP, rather than the ratio of gross exports to GDP. Manufacturing openness drops a lot, and nonmanufacturing openness rises (doubling in three of the four countries). This convergence in measured openness will be important below in thinking through the mechanics of trade balance adjustment.[6]

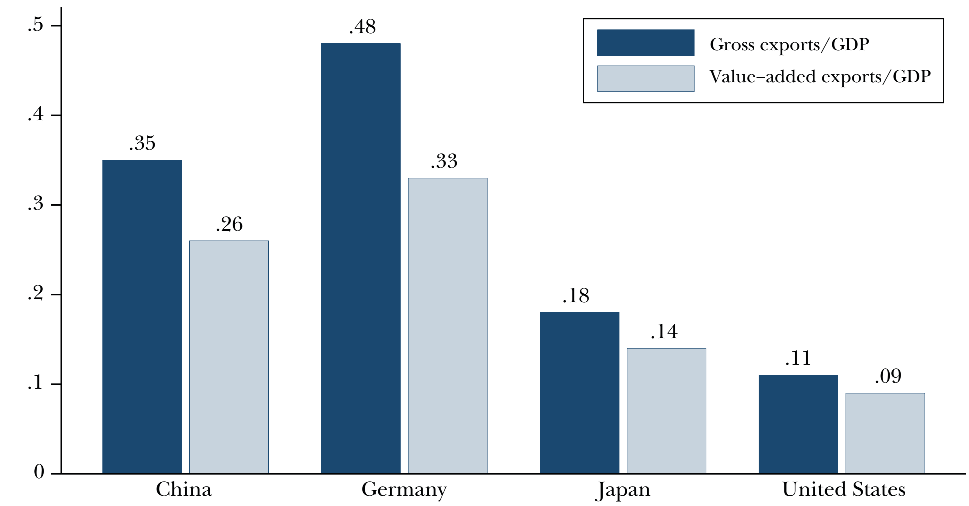
Turning to bilateral data, there are also differences between bilateral value-added versus gross exports to GDP ratios, particuarly for manufacturing. For the United States, the ratio of bilateral gross manufacturing exports to manufacturing GDP is about 0.17 for Canada and only .07 for value-added exports. For exports to the European Union, the comparable figures are 0.11 for gross exports and 0.06 for value-added exports. Therefore, while Canada looks like a more important export destination in gross terms, the European Union is equally important when we focus on how much US value added is actually being consumed in each country. The reason, of course, is that so much of US gross exports to Canada are embodied in Canadian exports back to the US economy. These value-added adjustments should be taken into account in evaluating the strength of bilateral demand linkages.

[5] In Bems, Johnson, and Yi (2010, 2011), we use a Leontief assumption to derive the same result from a global input-output accounting framework. This is a special case of the more general model in Bems and Johnson (2012).

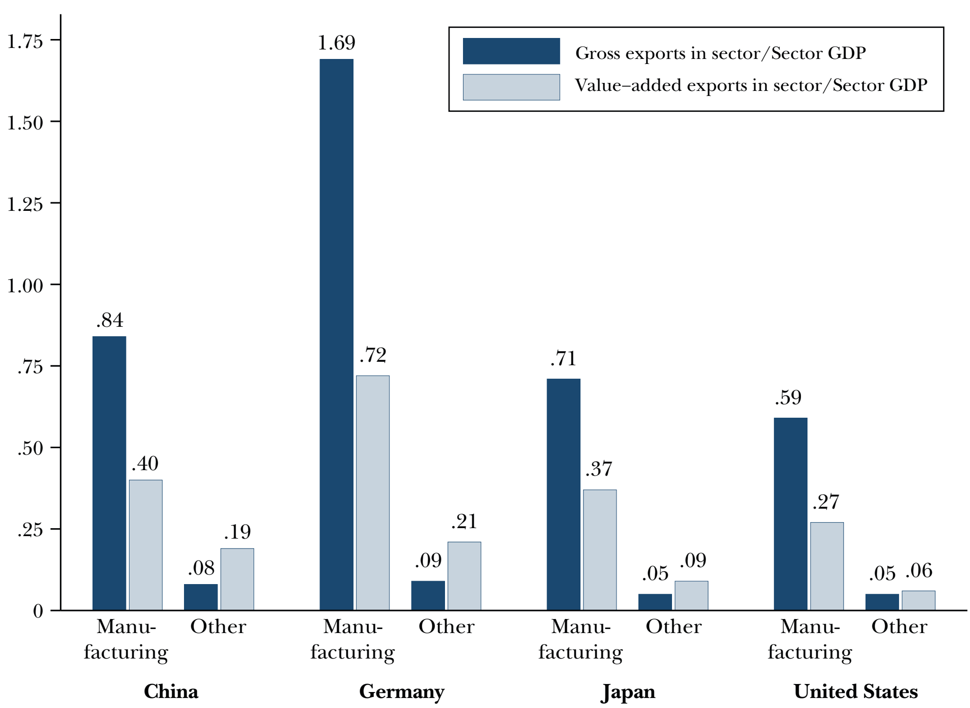
[6] In the absence of value-added trade data, one might be tempted to use the ratio of gross exports to gross output in calibrating openness. This is not only wrong in theory, it is also troublesome in practice because it makes the economy and individual sectors look too closed. For example, the aggregate ratio of gross exports to gross output in China is 0.11, less than half the ratio of value-added exports to GDP.

Figure 2
Aggregate and Sector-Level Openness for Top Four Exporting Countries

A: Aggregate Exports/GDP



B: Sector Exports/Sector GDP



Sources: World Input-Output Database (WIOD) and author's calculations.
Notes: Data are for 2008. The category labeled "Other" includes all nonmanufacturing industries.

Relative Price Changes and Competitiveness

How do changes in relative prices influence demand for value added from particular source countries? For example, how much would a reminbi appreciation lower demand for Chinese value added? What about if the reminbi appreciates against the yen, but holds its value against the dollar? How should we aggregate those heterogenous bilateral relative price changes to evaluate Chinese competitiveness?

The answers to these questions can be somewhat different depending on whether one takes a value-added or conventional view of trade. For example, suppose the renminbi appreciates (against all countries) and factor prices in all countries are fixed in producer currencies. How much this appreciation raises the relative price of Chinese exports depends on how much Chinese value added is embodied in them. Since China imports intermediate inputs to produce exports, China's export price depends on both the price of Chinese and foreign value added. As a result, a lower value-added to export ratio means that the appreciation will have a lower pass-through rate into export prices. The less these prices rise, the less demand for Chinese exports, and hence Chinese value added, falls.

Matters become more complicated when three countries are linked via production chains. For example, consider a scenario in which Japan exports computer parts to China, who then assembles them into a laptop and exports the laptop to the United States. If the Japanese yen depreciates against the US dollar (while the Chinese renminbi is fixed against the US dollar), then this brings down the price of Chinese-assembled laptops in the United States. This implies increased demand for laptops, which generates additional demand for value added from the Chinese computer assembly industry. Thus, even though there is no bilateral movement in the renminbi–dollar exchange rate, vertical input trade linkages imply that exchange rates vis-a-vis third-country input suppliers influence export competitiveness and hence demand for one's own value added.

As these examples illustrate, sorting out the effects of exchange rate movements (or other shocks to relative prices) on demand for exports and value added can be complicated. Fortunately, data on value-added exports can help cut through the fog. Since countries ultimately produce and trade value added, a natural approach would be to use price changes for value added originating from different countries, combined with trade weights based on value-added exports, to construct "real effective exchange rates" for value added (Bems and Johnson 2012).[7] These composite exchange rate indexes capture the effect of changes in relative value-added prices on demand for value added from each country.

In practice, this value-added approach to evaluating exchange rate movements leads to quantitatively different conclusions than conventional approaches. For example, in Bems and Johnson (2012), we find that, from 2000 to 2009, China's

[7] Though value-added real exchange rates can be motivated directly by appealing to value-added models, in Bems and Johnson (2012), we derive value-added weights from a constant elasticity of substitution model written in gross terms under the assumption that elasticities of substitution are equal in preferences and production functions.

value-added real effective exchange rate appreciated by 20 percentage points more than the conventional index used by the IMF. We also find that appreciations in value-added exchange rates for the European periphery prior to the euro-crisis were larger than implied by conventional indexes. The value-added perspective thus indicates that China's exchange rate has become less misaligned (consistent with rebalancing) and intra-EU rates were more misaligned (consistent with the build-up of imbalances within the European Union) than conventional indexes would indicate. The most important reason for these differences is that conventional indices are constructed using consumer price indexes, which are poor guides in practice to how the relative price of value added across countries, and hence demand for value added, changes over time.

Adjustment of Trade Imbalances

The geopolitics of external adjustment are often acrimonious. Not surprisingly therefore, the fact that bilateral trade balances are not equal in gross and value-added terms has attracted substantial attention in policy circles (Xing and Detert 2010; Lamy 2011; Johnson and Noguera 2012a). The value-added view of trade also has important implications for adjustment of multilateral trade balances—an insight that is less commonly appreciated, but perhaps of greater practical importance.

At the outset, it is crucial to emphasize that a country's multilateral trade balance is identical when measured in gross and value-added terms. The national accounts GDP identity states that total value added produced minus total final expenditure (including domestic and imported final goods) is equal to the gross trade balance. Because all final expenditure is ultimately value added purchased from some source, then this is the same as saying that value added produced minus value-added consumed (including domestic and imported value added) equals the gross trade balance. Since value added produced minus value added consumed is equal to value-added exports less value-added imports—that is, the value-added trade balance—the value-added trade balance equals the gross trade balance by construction.

This mechanical equality does not imply that the value-added view has nothing to contribute in analyzing external adjustment. To focus the discussion, consider a standard question asked by Obstfeld and Rogoff (2005, 2007): how much does the consumption real exchange rate—that is, relative consumer price levels—need to change to close the trade imbalance? The answer to this question depends on whether one uses value-added or gross trade data in calibrating the underlying macroeconomic model. Bems (2013) points out three distinct channels that can lead to different results.

First, the economy looks more closed when one uses value-added exports to GDP, rather than the ratio of gross exports to GDP, as the measure of how much output is exported. With a more closed economy (equivalently, stonger home bias in consumption), the "transfer problem" associated with closing imbalances is worse. Specifically, the decline in home expenditure relative to foreign expenditure needed to close home's deficit leads to a larger decline in home's terms of trade (the price of home relative to foreign tradables), thus increasing the size of the required real exchange rate change.

Second, manufacturing and nonmanufacturing sectors look more similar in terms of openness in value-added terms. This tends to reduce the size of the intranational, cross-sector relative price adjustment associated with closing the external imbalance, and hence reduce the required real exchange rate adjustment. Essentially, reducing the asymmetry in openness across sectors means that demand for output declines more uniformly across sectors (and hence cross-sector relative price changes are smaller) following the decline in home expenditure associated with closing the imbalance.

The third channel concerns elasticities, not openness. Typically, macroresearchers plug elasticities (like the elasticity of substitution between home and foreign goods) that are estimated using gross data into value-added models. Bems (2013) argues this approach overstates the appropriate elasticities for cross-sector or cross-country substitution of value added. Converting the estimated gross elasticities of substitution into levels appropriate for value-added models, he shows that the resulting value-added elasticities are lower than the gross elasticities typically used in the literature. Using these lower value-added elasticities increases the size of the real exchange rate adjustment needed to close imbalances.

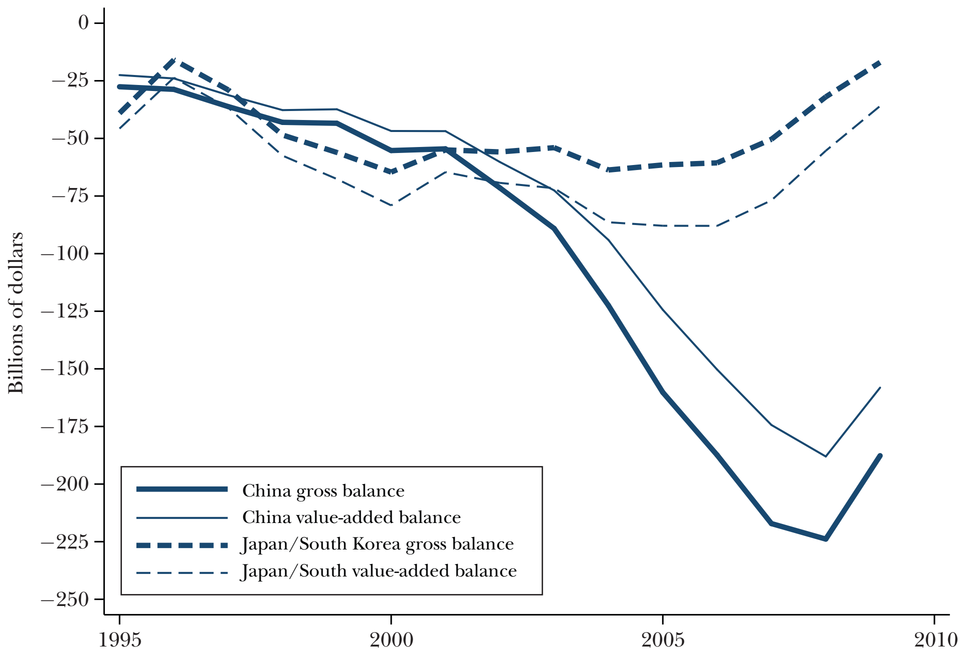
How these channels net out depends on the particular country under examination. Bems (2013) works out the net effects for a range of countries. Not surprisingly, accounting correctly for intermediate inputs in calibration matters most for countries like China, Mexico, or South Korea that are deeply integrated into global supply chains. For a decline in the trade surplus equal to 1 percent of GDP, the real exchange rate appreciates by 15–25 percent more in a model parameterized to be consistent with the value-added data, relative to the conventional approach that mixes value-added and gross data. Specific numbers aside, this analysis points to the usefulness of looking at value-added export data in studying the mechanics of external adjustment.

Shifting our attention to the bilateral level, bilateral trade balances are generally not equal in gross and value-added terms. This is true not only if bilateral gross trade is unbalanced, but holds even if bilateral gross trade is balanced.[8] For illustration, Figure 3 plots United States bilateral gross and value-added trade balances with China and the composite of Japan and South Korea using World Input-Output Database data. The US trade deficit with China looks smaller in value-added terms than it does in gross terms, while the deficit with Japan and South Korea looks correspondingly larger. The maximal difference in percentage terms between the gross and value-added US–China imbalance is about 23 percent in 2004 ($124.5 billion versus $94 billion). In terms of absolute values, the gap peaks at $42 billion in 2007.

Almost surely, this figure understates the true reallocation of trade imbalances. The reason is that the World Input-Output Database (like most other available

[8] With balanced bilateral trade, differences in bilateral value-added to export ratios for exports from country i to country j versus from j to i can generate imbalanced bilateral value-added trade. With imbalanced bilateral trade, the average level of value added to export ratios for a given country pair will scale up/down the value-added imbalance relative to the gross imbalance (Johnson and Noguera 2012a).

Figure 3
United States Trade Deficits with China, Japan, and South Korea



Sources: World Input-Output Database (WIOD) and author's calculations. Deficits for Japan and South Korea are combined in the figure.

input-output data) does not account for the high share of pure "processing trade" in Chinese exports. Specifically, just over half of Chinese exports are produced under its processing trade regime, where firms are allowed to import inputs duty-free if the resulting output is exported. Given these incentives, the imported input intensity of these firms is substantially higher than the average Chinese firm. Standard input-output tables report input requirements for the average firm only, however. Therefore, they understate the import content of exports for China, and thus overstate domestic value-added in Chinese exports.[9]

Adjusting value-added calculations to account for this bias, Koopman, Wang, and Wei (2012) find that the Chinese domestic content in exports from the processing trade sector was only about 25 percent in 2002, as compared to about 90 percent for normal nonprocessing exports. Correctly accounting for these discrepancies

[9] Although I focus on pure processing trade here, the core idea is more general. Micro-data indicate that export and import participation are highly correlated at the firm-level. Therefore, the imported input intensity of exporting firms is likely higher than that of the average firm in most countries. As in the case of processing trade, ignoring this fact (as standard input-output tables do) leads one to overestimate the domestic value-added content of exports.

lowers the ratio of Chinese domestic content exports from about 0.75 to 0.55 in 2002. Drawing on this work, in Johnson and Noguera (2012a), we implement an adjustment for processing trade in China within the global input-output framework and find that it leads the China–US trade balance to shrink by an additional 10 percentage points. Therefore, we find that the difference between the gross and value-added US–China imbalance was actually likely closer to 30–40 percent in 2004, roughly doubling the unadjusted calculation.

These adjustments to bilateral balances suggest that the burden of adjustment associated with closing the US trade balance would be redistributed away from China and toward Japan and Korea, in line with the reallocation of value-added trade balances. To date, however, there has been no work assessing how important these adjustments are quantitatively. This is a topic for future work.

International Trade

Value-added exports also provide a new perspective on traditional topics in international trade. I highlight applications related to the impact of frictions on trade, specialization patterns, the factor content of trade, and trade policy.

Trade Frictions

What is the impact of frictions—tariffs, nontariff barriers, transport costs, and others—on patterns of consumption versus production across countries? This question is typically addressed by examining the effect of frictions on gross production and trade. As a result, we know a lot about where gross output is produced and the destinations to which it is shipped. We know very little, however, about how trade frictions influence trade in value added, and hence differences between where value added is produced versus where it is consumed.

To launch this discussion, it is helpful to refer back to the five facts laid out earlier in this paper. We noted the significant differences between bilateral gross and value-added exports, and that these bilateral differences are systematically related to common proxies for trade costs. For example, the ratio of bilateral value-added to gross exports tends to be lower for country pairs located in the same region, and pairs that are separated by shorter distances. It is also lower for pairs of countries that have adopted regional trade agreements, and even lower for pairs that have adopted "deep" agreements, such as customs unions, common markets, and economic unions.

These underlying patterns all suggest that trade frictions have different effects on value-added trade versus gross trade. One way to think about this is that standard trade frictions impede gross trade, and hence induce the patterns of final and intermediate goods trade that we observe in the data. This trade in final and intermediate goods gives rise to the global input-output structure. As we use that input-output structure to compute value-added trade flows, we are implicitly aggregating the effect of frictions on gross trade to measure the composite impact of those frictions on value-added trade. The input-output structure is a device to map gross trade frictions into implied value-added trade frictions, which measure the reduced form impact of the full set of gross frictions in determining value-added consumption patterns.

One insight from thinking through this aggregation process is that both bilateral trade costs and trade costs between third countries directly influence bilateral value-added exports, whereas only bilateral trade costs directly influence bilateral gross trade (Noguera 2012). For this reason, value-added trade frictions are a manifestation not only of bilateral frictions, but rather the entire matrix of trade frictions among all countries.

Two points follow. First, one important reason that value-added exports are less sensitive to bilateral distance between countries than gross exports is that value added can be traded via third countries. For example, the United States can export intermediate inputs to Europe that are embodied in final European goods shipped to Russia. In a sense, Russia is then effectively "closer" to the United States than it looks on a map. Second, changes in trade costs between third countries can have a direct impact on bilateral value-added exports to other countries. For example, a tariff cut between Japan and China would have a direct effect on value-added exports from Japan to the United States. This point has interesting implications for policy discussions, which typically focus on bilateral rather than third-country barriers. I return to this point below.

Turning from cross-sectional to time series facts, we have seen large declines in the ratio of value-added to gross exports over the past few decades, with particularly large declines in fast-growing emerging markets. An important question is: do changes in gross trade frictions explain this divergence? The answer, by and large, is yes.

Using a multisector gravity model with trade in both final and intermediate goods, in Johnson and Noguera (2014), we decompose changes in the global input-output structure into components attributable to changes in trade frictions, changes in endowments and productivity, and changes in generic sector-to-sector input linkages or sector-level final expenditure shares. We find that changes in trade frictions explain nearly the entire decline in the ratio of value-added to gross exports for the world as a whole. We also explain differences across countries, where countries with large declines in trade frictions have seen particularly large declines in value-added relative to gross exports.

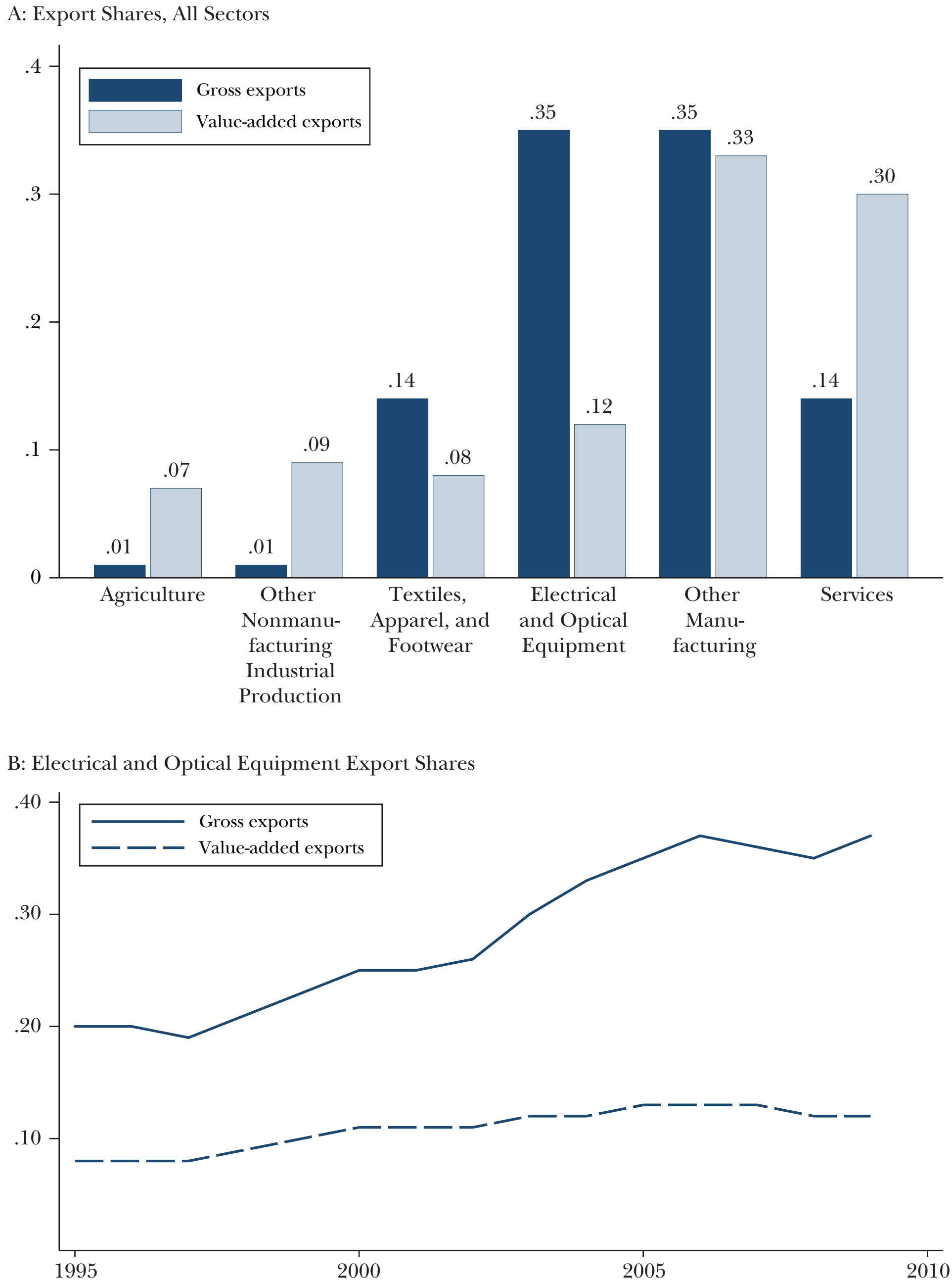
These results are consistent with the idea that value-added trade frictions have declined more slowly than gross trade frictions, leading to disproportionate growth in gross relative to value-added trade. Together with the discussion of bilateral differences, they point to new ways to think about the impact of frictions on trade.

Specialization Patterns

As we have seen, the sector-composition of gross exports can be quite different than the sector-composition of value-added exports. Looking at value-added composition forces us to revisit what we know about patterns of specialization.

This point is driven home by considering the example of China. The top panel of Figure 4 records the share of individual sectors in Chinese gross and value-added

Figure 4
Sector-Level Export Shares for China



Sources: World Input-Output Database (WIOD) and author's calculations. Export shares for all sectors are for 2008.
Note: "Agriculture" means agriculture, hunting, forestry, and fishing.

exports. As in most countries, nonmanufacturing sectors are substantially more important in value-added than gross terms. The more striking fact is that the share of Electrical and Optical Equipment shrinks dramatically, from about one-third of China's exports to just over one-tenth. This difference is of course consistent with the fact that these goods tend to be produced from imported intermediates that are assembled in China. Further, the bottom panel of Figure 4 plots the share of Electrical and Optical Equipment in value-added and gross exports over time. The share of this sector in gross exports has almost doubled since 1995, while the value-added share has barely changed. As such, gross and value-added trade provide very different pictures about what China genuinely produces and sells to the rest of the world.

This example illustrates a general point: what countries export may be very different from what they actually contribute to the production process. Countries that look like dominant exporters in particular sectors may in fact contribute very little value added to those exports. This basic point should be borne in mind in analyses of comparative advantage (Koopman, Wang, and Wei 2014). It should also factor into efforts to evaluate export sophistication across countries (Schott 2008, Wang and Wei 2010), or whether it matters for economic growth what countries export (Hausman, Hwang, and Rodrick 2007).

The Factor Content of Trade

Thus far, we have focused on the value-added content of international trade. Beneath this trade in value added lies trade in primary factors or production tasks. If one knows the quantities of factors needed to produce a unit of GDP in each sector, then one can use these to convert value-added export flows into factor flows. The difference between the quantity of domestic factors needed to produce value-added exports and the quantity of foreign factors needed to produce value-added imports is equal to the net factor content of trade, which measures factors embodied in GDP minus factors embodied in consumption. Moreover, the preceding logic is identical if one uses task contents in place of factor contents.

Value-added export data are useful for performing factor/task content calculations for two reasons. First, using them sidesteps an important conceptual problem with conventional approaches to factor content calculations. Specifically, these approaches made strong, increasingly untenable assumptions—either that gross exports are produced entirely from domestic gross output, or that imported inputs are produced with identical input requirements as domestic output. Relaxing these assumptions requires tracking trade in intermediates across countries and sectors, just as global input-output frameworks are designed to do. Therefore, Reimer (2006) and Trefler and Zhu (2010) proposed methods to compute the multilateral net factor content of trade (that is, the net quantity of factors each country exports to the rest of the world) using global input-output tables. While their approach cannot be used to recover bilateral factor trade, bilateral value-added export data can be used for this purpose. This is another advantage to using value-added data. Measuring bilateral trade in factors enables one to test bilateral predictions of the factor-contents theory that emerge when factor price equalization breaks down (Debaere 2003; Choi and Krishna 2004).

What are the implications of using value-added exports to compute factor contents? The conventional approach overstates factor trade because it assumes that gross exports of a country are produced using that country's technology alone. Instead, with traded inputs, gross exports of each country are produced using a convex combination of domestic and foreign technologies. By making effective production techniques more similar at home and abroad, traded inputs attenuate measured factor trade (Reimer 2006; Johnson 2011). Thus, appropriately accounting for intermediates lowers the measured factor content of trade relative to measurements that allow for differences in production techniques but do not incorporate traded intermediates. More work is needed to quantify the effect these adjustments have on tests of factor-contents theory.

Trade Policy Analysis

The rise of global supply chains has altered the costs and benefits of protection in a variety of ways (Baldwin 2012; Blanchard 2013). Yet empirical research on two-way interactions between trade policy and global supply chains is sparse. I expect that improvements in global input-output data and new measures of trade in value added will facilitate future work in this area. Therefore, I want to highlight a few specific ways in which the value-added analysis, and global input-output data more generally, can inform trade policy analysis.

First, the fact that gross exports and imports contain both foreign and domestic value-added is a core element of the value-added view of trade. The presence of domestic value added in imports gives rise to domestic constituencies that ought to favor liberalization. For example, exporters of intermediate goods that are then embodied in imported final goods should favor lower tariffs on those final goods imports. On the flip side, the presence of foreign value-added in exports ought to give rise to lobbying by exporters to liberalize imports of intermediates. As global supply chains become more important, these pressures should grow. With the new availability of data on input-output linkages across borders, the time seems ripe to investigate the role of these forces in determining trade policy.

Second, an important benefit of value-added export data is that it tracks value added to the final consumer even as it moves through third countries. This role for third parties has implications for trade policy. For example, a regional trade agreement between countries A and B is likely to increase trade in value added between countries C and A when C is an input supplier to country B. This trade-creating effect of the regional trade agreement, and third-country liberalizations more generally, ought to figure into policy analysis.

Third, global input-output tables and value-added trade data can potentially help quantify the extent to which global supply chains magnify the impact of trade barriers, an effect which is reminiscent of an older literature on the "effective rate of protection" (Yi 2003, 2010). In models of multistage production, trade costs are paid multiple times as goods pass across borders through a global supply chain, and trade costs imposed on the value of gross output impose a heavy burden when evaluated relative to the actual value added being traded. Building on this intuition, Koopman, Wang, and Wei (2014) call for value-added data to be used in quantifying these effects.

While the potential role of supply chains in magnifying trade barriers deserves attention, several caveats ought to be borne in mind. First, commonly used multisector models with "roundabout production" can match both gross and value-added trade simultaneously, yet they imply zero magnification of trade barriers. Second, in Johnson and Moxnes (2013), we caution that even models with sequential multistage production, which allow for magnification effects, do not deliver significant magnification when calibrated to match observed levels of final and intermediate goods trade. Given the potential importance of amplification effects in understanding the costs of protection, this area demands more research.

Concluding Remarks

The rise of global supply chains has led to far-reaching changes in the nature of international trade. In this article, I have focused on one particular implication: gross trade is not equal to trade in value added. While this fact has been known for some time, gaps between gross and value-added trade have only recently been quantified. These gaps are markers for differences in global supply chain activity across countries and over time. They are also important to keep in mind in quantitative work. Researchers should beware of mixing gross trade data with value-added production data, or using gross trade data in applications where the underlying theory is based on value added concepts. For both these reasons, I expect value-added export data to figure prominently in international macroeconomic and trade research and policy analysis going forward.

Because research using global input-output frameworks is still relatively new, much remains to be done not only in analyzing trade in value added, but also with regard to improving the data underlying its measurement. Enhanced international cooperation to measure global supply chain activity more accurately would be ideal. Even in its absence, however, much could be done to improve measurement on a country-by-country basis. For example, value-added export measurement would be improved by additional work on quantifying differences in imported input use across exporting versus nonexporting firms and incorporating these into input-output tables. Enhanced data collection for countries with large "processing trade" sectors—like China, Mexico, and other emerging markets—would be a good start. Another issue that deserves attention is how we track imported input use behind the border. That is, we need better data on where inputs from particular source countries go (that is, which firms/sectors use them) after they enter the country. Addressing these issues would increase the accuracy of value-added measurements.

Finally, though I have focused on using global input-output frameworks to compute value-added exports, the underlying data is also valuable in other applications. Most obviously, the data can be used to parameterize trade models written in gross terms with both cross-sector and cross-country input linkages. These models are useful in their own right. For example, trade policy is typically conducted using instruments levied on gross trade, like tariffs, so it is natural to start by analyzing trade policy in gross models. Nonetheless, the deep goal ought to be to better understand how gross policy instruments induce changes in value-added trade, since value added is directly connected to both factor income and final expenditure (and hence welfare). Accomplishing this goal requires a better understanding of the theoretical mapping between gross and value-added representations of international trade.

■ *I thank Rudolfs Bems, Emily Blanchard, Teresa Fort, Nina Pavcnik, and Kei-Mu Yi for helpful discussions, and I thank David Autor, Chang-Tai Hsieh, and Timothy Taylor for many helpful suggestions.*

References

Ambler, Steve, Emanuela Cardia, and Christian Zimmerman. 2002. "International Transmission of the Business Cycle in a Multi-Sector Model." *European Economic Review* 46(2): 273–300.

Backus, David K., Patrick J. Kehoe, and Finn E. Kydland. 1994. "Dynamics of the Trade Balance and the Terms of Trade: The J-Curve?" *American Economic Review* 84(1): 84–103.

Baldwin, Richard. 2012. "WTO 2.0: Global Governance of Supply-Chain Trade." Centre for Economic Policy Research, Policy Insight No. 64.

Baldwin, Richard, and Javier Lopez-Gonzalez. 2013. "Supply-Chain Trade: A Portrait of Global Patterns and Several Testable Hypotheses." NBER Working Paper 18957.

Bems, Rudolfs. 2013. "Intermediate Inputs, External Rebalancing and Relative Price Adjustment." Unpublished paper, International Monetary Fund.

Bems, Rudolfs, and Robert C. Johnson. 2012. "Value-Added Exchange Rates." NBER Working Paper 18498.

Bems, Rudolfs, Robert C. Johnson, and Kei-Mu Yi. 2010. "Demand Spillovers and the Collapse of Trade in the Global Recession." *IMF Economic Review* 58(2): 295–326.

Bems, Rudolfs, Robert C. Johnson, and Kei-Mu Yi. 2011. "Vertical Linkages and the Collapse of Global Trade." *American Economic Review* 101(3): 308–12.

Blanchard, Emily J. 2013. "What Global Fragmentation Means for the WTO: Article XXIV, Behind-the-Border Concessions, and A New Case for WTO Limits on Investment Incentives." WTO Working Paper ERSD-2014-03. http://www.wto.org/english/res_e/reser_e/ersd201403_e.pdf.

Choi, Yong-Seok, and Pravin Krishna. 2004. "The Factor Content of Bilateral Trade: An Empirical Test." *Journal of Political Economy* 112(4): 887–914.

Costinot, Arnaud, and Andrés Rodríguez-Clare. 2013. "Trade Theory with Numbers: Quanitfying the Consequences of Globalization." NBER Working Paper 18896.

Daudin, Guillaume, Christine Rifflart, and Daniele Schweisguth. 2011. "Who Produces for Whom in the World Economy?" *Canadian Journal of Economics* 44(4): 1403–37.

De Backer, Koen, and Sébastien Miroudot. 2013. "Mapping Global Value Chains." OECD Trade Policy Papers, No. 159. http://www.oecd-ilibrary.org/trade/oecd-trade-policy-working-papers_18166873.

Debaere, Peter. 2003. "Relative Factor Abundance and Trade." *Journal of Political Economy* 111(3): 589–610.

Dedrick, Jason, Kenneth L. Kraemer, and Greg Linden. 2010. "Who Profits from Innovation in Global Value Chains? A Study of the iPod and Notebook PCs." *Industrial and Corporate Change* 19(1): 81–116.

Fally, Thibault. 2012. "Production Staging: Measurement and Facts." Unpublished paper, University of California, Berkeley.

Hausman, Ricardo, Jason Hwang, and Dani Rodrick. 2007. "What You Export Matters." *Journal of Economic Growth* 12(1): 1–25.

Herrendorf, Berthold, Richard Rogerson, and Ákos Valentinyi. 2013. "Two Perspectives on Preferences and Structural Transformation." *American Economic Review* 103(7): 2752–89.

Hiratsuka, Daisuke, and Yoko Uchida, eds. 2010. *Input Trade and Production Networks in East Asia.* Cheltenham, UK: Edward Elgar Publishing.

Hummels, David, Jun Ishii, and Kei-Mu Yi. 2001. "The Nature and Growth of Vertical Specialization in World Trade." *Journal of International Economics* 54(1): 75–96.

Institute of Developing Econmics–Japan External Trade Organization (IDE-JETRO), and World Trade Organization (WTO). 2011. *Trade Patterns and Global Value Chains in East Asia: From Trade in Goods to Trade in Tasks.* Geneva: World Trade Organization.

Johnson, Robert C. 2011. "Factor Trade Forensics with Traded Intermediate Goods." Unpublished paper, Dartmouth College.

Johnson, Robert C. Forthcoming. "Trade in Intermediate Inputs and Business Cycle Comovement." *American Economic Journal: Macroeconomics.*

Johnson, Robert C., and Andreas Moxnes. 2013. "Technology, Trade Costs, and the Pattern of Trade with Multi-Stage Production." Unpublished paper, Dartmouth College.

Johnson, Robert C., and Guillermo Noguera. 2012a. "Accounting for Intermediates: Production Sharing and Trade in Value Added." *Journal of International Economics* 82(2): 224–36.

Johnson, Robert C., and Guillermo Noguera. 2012b. "Proximity and Production Fragmentation." *American Economic Review* 102(3): 407–411.

Johnson, Robert C., and Guillermo Noguera. 2014. "A Portrait of Trade in Value Added over Four Decades." Unpublished paper, Dartmouth College.

Koopman, Robert, Zhi Wang, and Shang-Jin Wei. 2012. "Estimating Domestic Content in Exports When Processing Trade Is Pervasive." *Journal of Development Economics* 99(1): 178–89.

Koopman, Robert, Zhi Wang, and Shang-Jin Wei. 2014. "Tracing Value-Added and Double Counting in Gross Exports." *American Economic Review* 104(2): 459–94.

Lamy, Pascal. 2011. "'Made in China' Tells Us Little about Global Trade." *Financial Times,* January 24.

Linden, Greg, Kenneth L. Kraemer, and Jason Dedrick. 2009. "Who Captures Value in a Global Innovation Network? The Case of Apple's iPod." *Communications of the ACM* 52(3): 140–44.

Mattoo, Aaditya, Zhi Wang, and Shang-Jin Wei, ed. 2013. *Trade in Value Added: Developing New Measures of Cross-Border Trade.* Washington, DC: Centre for Economic Policy Research and the World Bank.

Noguera, Guillermo. 2012. "Augmented Gravity: Accounting for Production Sharing." Unpublished paper, University of Warwick.

Obstfeld, Maurice, and Kenneth Rogoff. 2005. "Global Current Account Imbalances and Exchange Rate Adjustments." *Brookings Papers on Economic Activity,* no. 1, pp. 67–146.

Obstfeld, Maurice, and Kenneth Rogoff. 2007. "The Unsustainable U.S. Current Account Position Revisited." Chap. 9 in *G7 Current Account Imbalances: Sustainability and Adjustment,* edited by Richard H. Clarida. University of Chicago Press.

Puzzello, Laura. 2012. "A Proportionality Assumption and Measurement Biases in the Factor Content of Trade." *Journal of International Economics* 87(1): 105–111.

Reimer, Jeffrey J. 2006. "Global Production Sharing and Trade in the Services of Factors." *Journal of International Economics* 68(2): 384–408.

Schott, Peter K. 2008. "The Relative Sophistication of Chinese Exports." *Economic Policy,* vol. 23, pp. 5–49.

Timmer, Marcel, ed. 2012. "The World Input-Output Database (WIOD): Contents, Sources and Methods." WIOD Working Paper 10.

Timmer, Marcel P., Bart Los, Robert Stehrer, and Gaaitzen J. de Vries. 2013. "Fragmentation, Incomes, and Jobs: An Analysis of European Competitiveness." *Economic Policy* 28(76): 613–61.

Trefler, Daniel, and Susan Chun Zhu. 2010. "The Structure of Factor Content Predictions." *Journal of International Economics* 82(2): 195–207.

Wang, Zhi, and Shang-Jin Wei. 2010. "What Accounts for the Rising Sophistication of China's Exports?" In *China's Growing Role in World Trade,* edited by Feenstra, Robert, and Shang-Jin Wei. University of Chicago Press.

Xing, Yuquing, and Neal Detert. 2010. "How the iPhone Widens the United States Trade Deficit with the People's Republic of China." ADBI Working Paper 257, Asian Development Bank Institute.

Yi, Kei-Mu. 2003. "Can Vertical Specialization Explain the Growth of World Trade?" *Journal of Political Economy* 111(1): 52–102.

Yi, Kei-Mu. 2010. "Can Multi-Stage Production Explain the Home Bias in Trade?" *American Economic Review* 100(1): 364–93.

Journal of Economic Perspectives—Volume 28, Number 2—Spring 2014—Pages 143–152

Raj Chetty: 2013 Clark Medal Recipient

Martin Feldstein

Raj Chetty wrote to me during his first week as a freshman at Harvard College in 1997 to ask for a job as a research assistant. Although my research assistants were usually much further along in their studies, the high school essay that he sent to me—a critical comment on Robert Fogel and Stanley Engerman's (1974) *Time on the Cross: The Economics of American Negro Slavery*—was so well done that I decided to interview him. A brief discussion convinced me that he was unusually bright and would be both productive as a research assistant and fun to work with. But after a few months working with Raj as a research assistant, I realized that he was quite exceptional and should be investing in his own intellectual development rather than helping me with my current statistical research. I suggested a variety of things that he might read and we could discuss together. These included not only some papers in public economics but also risk theory and statistical decision theory.

Raj completed his Harvard BA in three years, graduating summa cum laude with a thesis on interest rates and business investment that was eventually the basis for a 2007 article in the *Review of Economic Studies.* He went on to complete his PhD at Harvard in the next three years. He then went to Berkeley as an assistant professor in 2003. Harvard lured him back in 2009.

Raj Chetty is eminently deserving of being awarded the John Bates Clark Medal at the age of 33. His research has transformed the field of public economics. His work is motivated by important public policy issues in the fields of taxation, social insurance, and public spending for education. He approaches his subjects with a creative redefinition of the problems that he studies, and his empirical methods often draw

■ *Martin Feldstein is the George F. Baker Professor of Economics, Harvard University, Cambridge, Massachusetts. His email address is mfeldstein@harvard.edu.*

http://dx.doi.org/10.1257/jep.28.2.143 doi=10.1257/jep.28.2.143

Table 1
Selected Papers by Raj Chetty

1. "Interest Rates, Irreversibility, and Backward-Bending Investment." 2007. *Review of Economic Studies* 74(1): 67–91.
2. "Moral Hazard versus Liquidity and Optimal Unemployment Insurance." 2008. *Journal of Political Economy* 116(2): 173–234.
3. "Is the Taxable Income Elasticity Sufficient to Calculate Deadweight Loss? The Implications of Evasion and Avoidance." 2009. *American Economic Journal: Economic Policy* 1(2): 31–52.
4. "Salience and Taxation: Theory and Evidence," (with Adam Looney and Kory Kroft). 2009. *American Economic Review* 99(4): 1145–77.
5. "Dividend and Corporate Taxation in an Agency Model of the Firm," (with Emmanuel Saez). 2010. *American Economic Journal: Economic Policy* 2(3): 1–31.
6. "How Does Your Kindergarten Classroom Affect Your Earnings? Evidence from Project STAR," (with John Friedman, Nathaniel Hilger, Emmanuel Saez, Diane Whitmore Schanzenbach, and Danny Yagan). 2011. *Quarterly Journal of Economics* 126(4): 1593–1660.
7. "Bounds on Elasticities with Optimization Frictions: A Synthesis of Micro and Macro Evidence on Labor Supply." 2012. *Econometrica* 80(3): 969–1018.
8. "Using Differences in Knowledge across Neighborhoods to Uncover the Impact of the EITC on Earnings," (with John N. Friedman and Emmanuel Saez). 2013. *American Economic Review* 103(7): 2683–2721.
9. "Active vs. Passive Decisions and Crowd-Out in Retirement Savings Accounts: Evidence from Denmark," (with John Friedman, Søren Leth-Petersen, Torben Heien Nielsen, and Tore Olsen). *Quarterly Journal of Economics*, forthcoming.
10. "Measuring the Impacts of Teachers I: Evaluating Bias in Teacher Value-Added Estimates," (with John N. Friedman and Jonah E. Rockoff). NBER Working Papers 19423.
11. "Measuring the Impacts of Teachers II: Teacher Value-Added and Student Outcomes in Adulthood," (with John N. Friedman and Jonah E. Rockoff). NBER Working Paper 19424.

on experimental evidence or unprecedentedly large sets of integrated data. While his work is founded on basic microeconomics, he modifies this framework to take into account behavioral and institutional considerations. The American Economic Association (2013) in its announcement of the Clark Medal declared: "He has established himself in a few short years as arguably the best applied microeconomist of his generation."

Chetty is a prolific scholar. It is difficult to summarize all of Chetty's research or even to capture the details of his most significant papers. I have therefore chosen a selection of Chetty's important papers dealing with taxation, social insurance, and education that contributed to his selection as the winner of the John Bates Clark Medal. These examples from different aspects of public economics indicate Chetty's combination of selecting important issues, creatively extending existing theory, and applying novel empirical methods. Table 1 presents a list of the papers by Chetty cited in this essay, and as is the style of this journal, I will refer to the papers by number.

Raj Chetty

Taxation

In studies of individual taxation, Chetty has taken on questions of how people perceive taxes and how they react to them. Chetty's most widely cited paper, "Salience and Taxation: Theory and Evidence" [4], written with Adam Looney and Kory Kroft, is a good example of Raj's innovative style in conceptualizing a question and bringing novel evidence to bear. He begins by posing the question of whether consumers react to the taxes that are imposed on retail products in the way assumed by standard theory. He shows that they do not. He then explores the implications of the more realistic description of their behavior for the incidence and welfare loss of taxes.

Their first clever strategy for assessing how consumers reacted to taxes on the products that they buy was to compare the reaction to state taxes on beer that are levied in some states as excise taxes, and therefore are built into the price that consumers see on the shelf, with the reactions to taxes on beer levied in other states as sales taxes, which are collected at the cash register after the individual has already made a purchase decision. They found that the more "salient" price effect of the excise tax had a bigger impact on the quantity of beer purchased than a comparable-size sales tax collected at the cash register.

Chetty and his coauthors confirmed this "salience" effect by conducting a major experiment in a large grocery store. They persuaded the store's management to post tax-inclusive prices for several hundred randomly selected products for several weeks. Using scanner data, they found that the information about the higher prices reduced demand relative to control products and nearby stores where the tax was only levied at the cash register.

This novel question and the imaginative source of evidence implies that consumers are making suboptimal purchase decisions and therefore that the tax imposes a higher utility loss to the extent that the consumers ignore the taxes. The consumer also bears more of the burden of the tax because there is less of a decline in the quantity purchased. Chetty and his coauthors then show how to modify the traditional incidence and deadweight loss formula to take into account the consumers' suboptimal behavior.

Many economists have noted that small changes in prices or taxes do not induce changes in behavior comparable to the effects of larger changes in prices or taxes. More formally, the elasticity of response appears to be greater in response to larger tax and wage changes. Chetty [7] explains this apparent "threshold" effect by the cost that individuals face in adjusting to such changes. He develops a formal model that calibrates these adjustment costs and uses this model to reconcile differences between micro and macro models of labor supply. Chetty then goes beyond this application to positive economics to show how traditional calculations of deadweight loss can be modified to take these adjustment costs into account.

Another impressive application of an expanded model of taxation was Chetty's analysis [3] of the implications of taxpayer evasion of taxes. In a basic model of the deadweight loss of taxation in a model without tax evasion, higher marginal income tax rates create a deadweight loss in three ways: 1) reduced labor supply, broadly defined to include a number of effects like lower participation, reduced hours, less effort, the choice of lower productivity jobs; 2) a shift from taxable cash to untaxed fringe benefits, nice offices, and other favorable working conditions; and 3) a shift of consumer spending to categories that are tax deductible like mortgage interest, local government taxes, and others. The resulting deadweight loss can be calculated using the effect of the higher marginal tax rate on total taxable income because all three effects are responses to the same marginal tax rate, therefore implying that the taxable income is a type of Hicksian composite good (Feldstein 1999).

Chetty [3] extended this analysis to allow for the realistic possibility that taxpayers may modify their reporting of income in order to avoid paying their full tax liabilities and that taxpayers may incorrectly estimate the probability of getting caught for this tax evasion. In his paper titled, "Is the Taxable Income Elasticity Sufficient to Calculate the Deadweight Loss? The Implications of Evasion and Avoidance," Chetty showed that when evasion is taken into account, the taxable income elasticity is no longer a sufficient statistic with which to calculate the deadweight loss. He goes on to show how the deadweight loss formula based on the taxable income elasticity can be modified to derive a measure of deadweight loss as a weighted average of the elasticity of taxable income and the elasticity of total earnings with respect to the marginal tax rate. The relative weights placed on these two components depend on the extent to which changes in taxable incomes are driven by responses with real resource costs, such as changes in the form of compensation and of tax-deductible spending, versus responses such as underreporting taxable income that have private costs to the individual (such as the risk of paying a fine) but no net social cost.

Tax rules in the United States and other countries discriminate against saving by taxing the return to capital. However, the United States and other countries also offset this effect to some extent with special rules for reducing the tax on the return to saving relative to other forms of income. In the United States, such provisions include the 401(k) employer plans (in which employees have the opportunity to exclude part of their earnings from current taxation if employers deposit those funds in long-term investment accounts) and the Individual Retirement Accounts (in which individuals can choose to exclude part of their earnings by depositing those funds in similar investment accounts). Employees with 401(k) plans and Individual Retirement Accounts do increase their saving in that form; however, it remains controversial whether this represents a net increase in the employees' total saving, or a transfer from other forms of saving, or a rise in one form of saving that is offset by increased mortgage borrowing. This question cannot be fully resolved based on US experience because of limits on the data on total individual assets and liabilities.

For the data to tackle this issue, Chetty along with John Friedman, Søren Leth-Petersen, Torben Heien Nielsen, and Tore Olsen [9] reached out to Denmark, where there are complete records on the savings, asset, and liabilities of everyone in the country. The Danish tax system has both types of savings incentive accounts: the 401(k)-type and IRA-type. Moreover, Denmark changed the rules of these programs during the sample period of the study, thus offering a source of variation for the analysis to exploit. In this analysis, Chetty and his coauthors distinguished between "passive savers" who do not respond to any of the changes in the tax rules applicable to saving and "active savers" who do respond. Employer contributions to 401(k)-type plans do raise the total saving of passive savers in Denmark, but changes in the incentives in IRA-type plans do not alter the savings of this "passive saver" group. Active savers do respond to saving incentives in IRA-type plans but also offset their increased saving in these accounts by reducing their net saving in other accounts. This evidence suggests that IRA-type saving incentives in Denmark do not increase national saving.

The analysis of Chetty and his coauthors in this study [9] focused on the implications for overall national saving, not on the welfare effect of changes in the effective marginal tax rate on saving through IRA-type plans. Taxes on the return to saving create a deadweight loss by altering the relative price of current and future consumption. Since saving is only the current outlay to purchase future consumption, a tax on saving creates a deadweight loss even if does not alter the amount of saving (Feldstein 1978). The analysis in [9] shows that the active savers do not increase their total saving in response to more favorable IRA-type rules, but shift saving to a form that delivers a higher net-of-tax return and therefore higher future consumption. In doing so, the more favorable IRA rules do reduce the deadweight loss associated with the existing taxation of saving.

The four studies that I have summarized so far all deal with the effects of taxes on the behavior of individuals. Chetty shifts his focus to the behavior of firms in

"Dividend and Corporate Taxation in an Agency Model of the Firm" [5], written with Emmanuel Saez. This paper begins by using the experience with the 2003 reduction in the rate of personal income tax on dividends to confirm earlier evidence that such tax changes lead to higher dividend payouts by firms (Feldstein 1970). The authors then show that the increased dividend payout is greatest in firms where senior managers and the board of directors have substantial share ownership. Chetty and Saez then develop a model in which shareholders and managers have conflicting interests over the desirability of dividends versus retained earnings. This agency model is used to calculate the deadweight burden when higher taxes on dividends cause an increase in retained earnings.

Social Insurance and the Safety Net

The design and optimal level of social insurance typically involves balancing the benefits provided by insurance protection against the loss of output and other sources of deadweight loss caused by the moral hazard created by the social insurance.

Chetty [2] made a major contribution to understanding this dynamic in a study of unemployment insurance. There is substantial evidence that higher unemployment insurance benefits lead to longer periods of unemployment. Chetty begins by decomposing the reasons for the increased duration into liquidity effects and moral hazard effects. He finds that increases in benefits have much bigger effects for households that are liquidity constrained. He also shows that lump-sum unemployment benefits have bigger effects on liquidity-constrained households.

The optimal level of unemployment insurance benefits depends on balancing the insurance protection that benefits provide with the loss of output caused by the moral hazard of the induced increase in unemployment. In turn, the gain from social insurance depends on the risk aversion of the unemployed individual. Earlier studies used the measures of risk aversion derived from investment portfolio decisions, which are inherently measures of long-term risk aversion. Chetty [2] notes that an unemployed individual at the time of unemployment typically has substantial fixed commitments—like rent or mortgage payments—that cannot be changed in the short run. In this situation, a loss of income associated with unemployment represents a larger fraction of the individual's uncommitted income, implying that the relevant risk aversion parameter is larger than it would be when considering long-term unconstrained decisions.

Chetty then derives a formula for the optimal level of unemployment benefits that depends on the reduced form elasticity with respect to liquidity and the moral hazard elasticity. Because the implied level of risk aversion is higher than in other previous studies, Chetty concludes that the optimal level of unemployment benefits is substantially higher than previous studies found and also higher than the benefit levels that are typical in the United States.

The Chetty calculation of the optimal level of unemployment benefits assumes a structure of the unemployment insurance system that provides government-financed benefits proportional to previous wages. My own research on unemployment insurance led me to favor an alternative approach of mandatory unemployment insurance saving accounts that provide liquidity when an individual is unemployed based on previous mandatory saving by that individual (Feldstein and Altman 2007). Chetty's evidence on the importance of liquidity for the unemployed reinforces the desirability of such a structure that provides liquidity with relatively little risk of moral hazard—because an unemployed individual will be drawing upon that person's own previously accumulated funds.

While unemployment insurance is designed to provide cash benefits to individuals with a short-term loss of income, other programs like the Earned Income Tax Credit (EITC) are intended to provide cash to individuals whose income is low for a sustained period of time. The structure of the EITC makes benefits a function of the individual's earned income. If individuals understand the rules, they can adjust their income or the amount of income that they report to maximize their benefit.

In "Using Differences in Knowledge across Neighborhoods to Uncover the Impacts of EITC on Earnings," Chetty with John Friedman and Emmanuel Saez [8] explore the extent to which individuals adjust the income that they report by using a remarkable set of administrative tax return data on all individuals who filed for EITC in the 14 years from 1996 to 2009. The parameters of the EITC differ by family status, but the program provides a payment that is a percentage of earned income up to a certain level (for a single parent with two children in 2013, 40 percent of income up to $13,430 earned), then the amount of the credit paid does not change over a certain income range (for a single parent, two children, up to $17,530 earned), and then the credit phases out as additional income is earned (for a single parent, two children, the phase out rate is 21.06 percent) (Maag and Carasso 2013). With the enormous size of the dataset, Chetty and his coauthors can study the extent to which individuals in different geographic areas bunch their earnings at the income level that maximized their EITC payments.

More specifically, the study begins by focusing on self-employed individuals because that group has the greatest control over their reported earnings. They find that the degree of "sharp bunching" at the benefit-maximizing level of income differs substantially among geographic areas. Chetty and his coauthors infer that differences among areas are not random but reflect local knowledge. Self-employed individuals will become more informed about the EITC schedule if they move from an area with less "sharp bunching" to an area with more "sharp bunching." The study then uses this area measure of "sharp bunching" to assess the extent to which employed individuals (in addition to the self-employed) are able to report W-2 wage incomes that come close to maximizing their benefits. They find that recipients of the EITC do have an ability to learn from the knowledge in their neighborhood and to "optimize" in this way. This personal optimization may reflect actual earnings or fraudulent reporting, a difference that cannot be determined from the data.

Education

Spending on education is one of the most substantial and important activities of government. For example, state and local governments spent $574 billion on primary and secondary education during the 2009–2010 school year, and $243 billion on colleges and universities (National Center for Education Statistics, undated). Economists are interested in ways to make that educational spending more productive and in understanding the effects of educational quality on later life outcomes. To study these important issues, Chetty and various coauthors amassed and linked amazing sets of data so they could trace the success of students in primary school to their college attendance and to their earnings later in life.

In the first of these education studies, Chetty with John Friedman, Nathaniel Hilger, Emmanuel Saez, Diane Whitmore Schanzenbach, and Danny Yagan [6] evaluated the impact of children's early primary education on their important outcomes later in life, including college attendance and incomes at age 27, using data from Project STAR, a careful experiment in Tennessee that randomly assigned students in grades kindergarten through third grade to different classrooms with different teachers. (Age 27 was the most recent year of the tax data that they could link to the primary school records.) Their first dramatic finding was that kindergarten test scores are highly correlated with adult outcomes, including college attendance and earnings, at age 27. This analysis also showed that students who were randomly assigned to small classes, as well as students with more experienced teachers, had higher earnings later in life. There is also a contagion effect: some classes in kindergarten through third grade produced better lifetime performance for the class as a whole.

In a second major education study, Chetty with John Friedman and Jonah Rockoff [10] studied whether teachers in a major urban school system who increase their pupils' test scores actually help them with achievements later in life. This study is another tour de force of combining massive independent datasets. There is no fancy economic theory, but rather thoroughly convincing econometric evidence based on bringing large amounts of relevant data to bear in a sophisticated econometric analysis. They start with data on 2.5 million pupils in grades 3 through 8 linked to tax records of their parents and of themselves as adults. Also integrated into the linked data are the identities of the colleges that they later attend, their incomes, and even the places where they live as adults. The teachers of these 2.5 million pupils are evaluated based on their "value added"—that is, the increase in test scores of these pupils in primary school. Using school district data for each pupil and taking into account each student's prior test scores, Chetty and his coauthors first study the effect of changing teacher assignments to show that there is little or no statistical bias in assessing the value added of individual teachers.

In a follow-up study [11], Chetty and the same coauthors also find that pupils who had high-value-added teachers are more likely to attend college, attend higher-ranked colleges, live in neighborhoods that measure higher in socioeconomic status, and earn higher salaries as adults. Perhaps the most amazing finding in this study is that replacing a teacher in the bottom 5 percent of the value-added distribution with an average teacher would "increase the present value of students' lifetime income by more than $250,000 for the average classroom in our sample." In short, test-score effects identify teachers that can consistently improve student performance, and those improved test scores have very substantial positive effects on the students' lifetime earnings.

Colleague, Mentor, Teacher

Although the Clark Medal is awarded primarily on the basis of the economist's research contribution, Chetty also contributes fully as a colleague, mentor, and teacher.

Chetty has been the editor of the *Journal of Public Economics* since 2009, and Co-Director (with Amy Finkelstein) of the Public Economics Program of the National Bureau of Economic Research since 2008. When Chetty came to Harvard as a professor, he organized the Lab for Economic Applications and Policy (http://leap.fas.harvard.edu/) to encourage active collaboration among graduate students and faculty of the economics department as well as researchers in other Harvard faculties and visitors from other universities.

Chetty has made and is continuing to make a major contribution to the teaching of public economics. Much of Chetty's work, including several of the papers summarized here, was produced by Chetty and small teams of researchers, many of whom are very bright graduate students. Their experience creates a group of researchers who will at least aspire to follow Chetty's methods and approach.

At Harvard, Chetty not only teaches public economics to both graduate students and undergraduates, but for the first term of the second-year graduate course in public economics at Harvard, he has also made his Fall 2012 course available online, including 24 lectures of 90 minutes each, together with lecture slides and a reading list, at http://www.rajchetty.com/index.php/lecture-videos. It is certainly the best modern "textbook" of public economics.

References

American Economic Association (AEA) Honors and Awards Committee. 2013. "Report on Raj Chetty." March. http://www.aeaweb.org/honors_awards/bios/Raj_Chetty.php.

Feldstein, Martin S. 1970. "Corporate Taxation and Dividend Behaviour." *Review of Economic Studies* 37(1): 57–72.

Feldstein, Martin. 1978. "The Welfare Cost of Capital Income Taxation." *Journal of Political Economy* 86(2, Part 2): S29–51.

Feldstein, Martin. 1999. "Tax Avoidance and the Deadweight Loss of the Income Tax." *Review of Economics and Statistics* 81(4): 674–80.

Feldstein, Martin, and Daniel Altman. 2007. "Unemployment Insurance Saving Accounts." In *Tax Policy and the Economy*, Vol. 21, edited by James M. Poterba, 35–64. MIT Press.

Fogel, Robert William, and Stanley L. Engermann. 1974. *Time on the Cross: The Economics of American Negro Slavery.* W. W. Norton & Company.

Maag, Elaine, and Adam Carasso. 2013. "The Tax Policy Center Briefing Book: Taxation and the Family: What is the Earned Income Tax Credit?" January 12. Tax Policy Center. http://www.taxpolicycenter.org/briefing-book/key-elements/family/eitc.cfm.

National Center for Education Statistics. Undated. Table 31: Direct general expenditures of state and local governments for all functions and for education, by level of education and state: 2008–9 and 2009–10. http://nces.ed.gov/programs/digest/d12/tables/dt12_031.asp.

Journal of Economic Perspectives—Volume 28, Number 2—Spring 2014—Pages 153–176

Fluctuations in Uncertainty†

Nicholas Bloom

Uncertainty is an amorphous concept. It reflects uncertainty in the minds of consumers, managers, and policymakers about possible futures. It is also a broad concept, including uncertainty over the path of macro phenomena like GDP growth, micro phenomena like the growth rate of firms, and noneconomic events like war and climate change. In this essay, I address four questions about uncertainty.

First, what are some facts and patterns about economic uncertainty? Both macro and micro uncertainty appear to rise sharply in recessions and fall in booms. Uncertainty also varies heavily across countries—developing countries appear to have about one-third more macro uncertainty than developed countries.

Second, why does uncertainty vary during business cycles? The types of exogenous shocks that can cause recessions—like wars, oil price jumps, and financial panics—typically also increase uncertainty. Uncertainty also appears to endogenously increase during recessions, as lower economic growth induces greater micro and macro uncertainty.

Third, do fluctuations in uncertainty affect behavior? Greater uncertainty appears to reduce the willingness of firms to hire and invest, and consumers to spend. However, there is also some evidence that uncertainty can stimulate research and development—faced with a more uncertain future, some firms appear more willing to innovate.

■ *Nicholas Bloom is Professor of Economics, Stanford University, Stanford, California, Co-Director of the Productivity Program, National Bureau of Economic Research, Cambridge, Massachusetts, and Research Fellow, Center for Economic Performance, London School of Economics, London, United Kingdom, and Research Fellow, Center for Economic and Policy Research, London, United Kingdom. His email address is nbloom@stanford.edu.*

† To access the Appendix and disclosure statement, visit
http://dx.doi.org/10.1257/jep.28.2.153 doi=10.1257/jep.28.2.153

Fourth, has higher uncertainty worsened the Great Recession and slowed the recovery? A 2008 jump in uncertainty was likely an important factor exacerbating the size of the economic contraction, accounting for maybe one-third of the drop in the US GDP.

Much of this discussion is based on research on uncertainty from the last five years, reflecting the recent growth of the literature. This surge in research interest in uncertainty has been driven by several factors. First, the jump in uncertainty in 2008 and its likely role in shaping the Great Recession has focused policy attention onto the topic. Second, the increased availability of empirical proxies for uncertainty, such as panels of firm-level outcomes, online news databases, and surveys, has facilitated empirical work. Third, the increase in computing power has made it possible to include uncertainty shocks directly in a wide range of models, allowing economists to abandon assumptions built on "certainty equivalence," which refers to the amount of money that would be required as compensation for risk. While there has been substantial progress, a range of questions remain open around the measurement, cause, and effect of uncertainty, making this a fertile area for continued research.

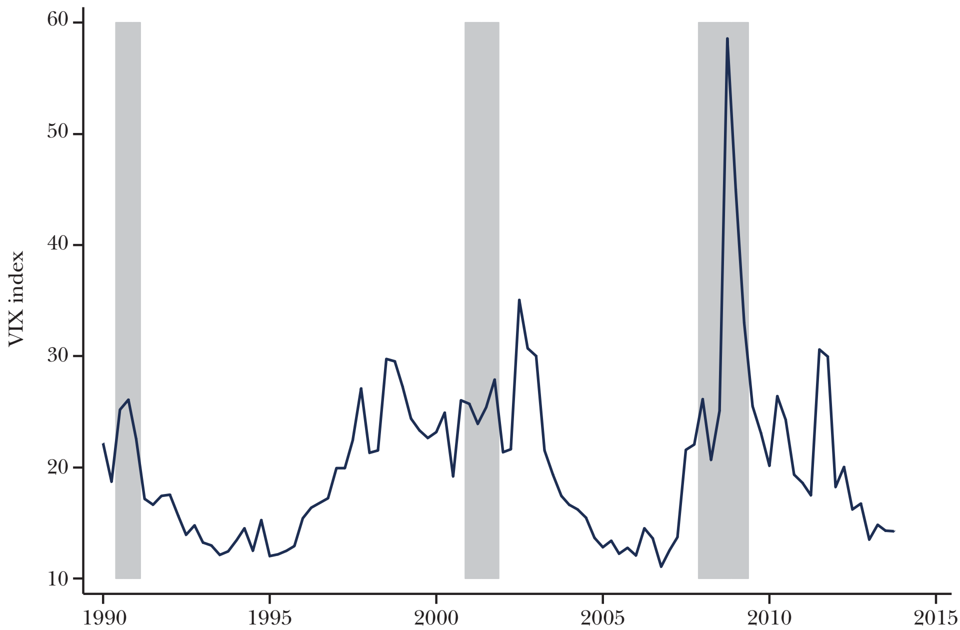
The Facts of Uncertainty

Frank Knight (1921), the famous Chicago economist, created the modern definition of *uncertainty*. Knight started by defining the related concept of *risk*, which he argued describes a known probability distribution over a set of events. In his terminology, flipping a coin is risky—for a fair toss there is a 50 percent chance of heads and a 50 percent chance of tails. In contrast, Knight defined *uncertainty* as peoples' inability to forecast the likelihood of events happening. For example, the number of coins ever produced by mankind is uncertain. To calculate this would require estimating the distribution of coins minted across the hundreds of countries that exist today and throughout history, a task where most people would have no idea even how to begin.

In this article, I'll refer to a single concept of uncertainty, but it will typically be a stand-in for a mixture of risk and uncertainty. Given this broad definition of uncertainty, it should be unsurprising that there is no perfect measure but instead a broad range of proxies. The volatility of the stock market or GDP is often used as a measure of uncertainty because when a data series becomes more volatile it is harder to forecast. Other common measures of uncertainty include forecaster disagreement, mentions of "uncertainty" in news, and the dispersion of productivity shocks to firms. I start by highlighting four key facts about uncertainty based on these proxies.[1]

[1] All the data used in this paper is available in an online Appendix available with this paper at http://e-jep.org, and also at my website in this zip file: http://www.stanford.edu/~nbloom/JEPdata.zip.

Figure 1
Stock-Market Implied Volatility is Higher in Recessions



Source: Author using data from the Chicago Board of Options and Exchange.
Notes: Figure 1 shows the VIX index of 30-day implied volatility on the Standard & Poor's 500 stock market index. The VIX index is traded on the Chicago Board Options Exchange. It is constructed from the values of a range of call and put options on the Standard & Poor's 500 index, and represents the market's expectation of volatility over the next 30 days. Gray bars are NBER recessions.

Fact 1: Macro Uncertainty Rises in Recessions

The volatility of stock markets, bond markets, exchange rates, and GDP growth all rise steeply in recessions. In fact, almost every macroeconomic indicator of uncertainty I know of—from disagreement amongst professional forecasters to the frequency of the word "uncertain" in the *New York Times* (Alexopolous and Cohen 2009)—appears to be countercyclical.

As one example, Figure 1 shows the VIX index of 30-day implied volatility on the Standard & Poor's 500 stock market index. The VIX index is traded on the Chicago Board Options Exchange. It is constructed based on the values of a range of call and put options on the Standard & Poor's 500 index and represents the market's expectation of volatility over the next 30 days. The VIX index is clearly countercyclical, rising by 58 percent on average in recessions (the shaded areas in the figure) as dated by the National Bureau of Economic Research.

One explanation for this surge in stock market volatility in recessions is the effect of leverage. In recessions, firms usually take on more debt, which increases their stock-returns volatility. However, Schwert (1989) calculates that the leverage effect can explain at most 10 percent of this rise in uncertainty during recessions.

Another explanation is that increased risk aversion during recessions will tend to increase the prices of options (because options provide insurance against large price movements), biasing up this measure of uncertainty. However, these fluctuations in the VIX are too large to be explained by plausible movements in risk aversion (Bekaert, Hoerova, and Lo Duca 2013). Moreover, it is not just stock markets that become more volatile in recessions. Other financial prices like exchange rates and bond yields also experience surging volatility in recessions.

An alternative proxy of uncertainty is disagreement amongst professional forecasters. Periods when banks, industry, and professional forecasters hold more diverse opinions are likely to reflect greater uncertainty. Examining data from the Philadelphia Federal Reserve panel of about 50 forecasters shows that between 1968 and 2012 the standard deviation across forecasts of US industrial production growth was 64 percent higher during recessions, similar to results from European countries (Bachmann, Elstner, and Sims 2010). So forecaster disagreement is sharply higher in downturns.

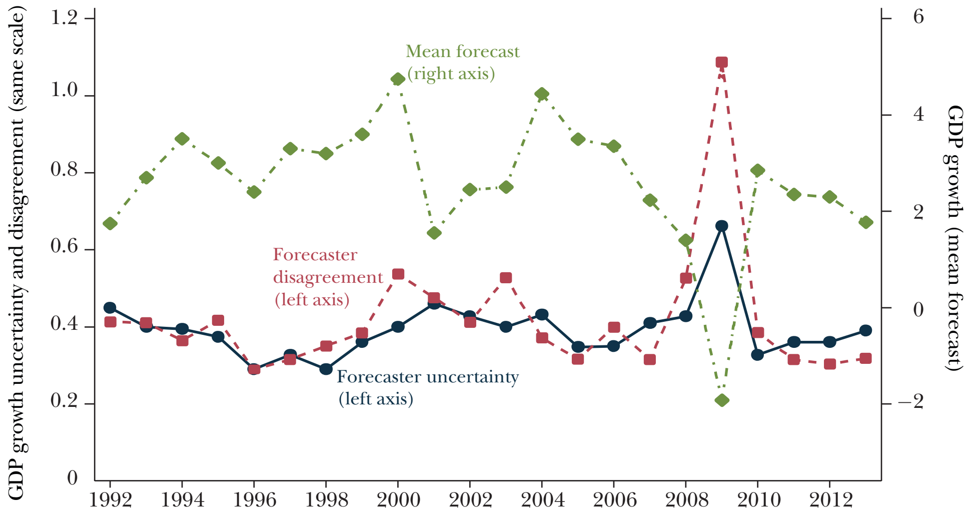
A related proxy is how uncertain forecasters are about their own forecasts, which is called subjective uncertainty. The Philadelphia Federal Reserve has since 1992 asked forecasters to provide probabilities for GDP growth (in percent) falling into ten different bins: "< -2," "-2 to -1.1," "-1 to -0.1," and so forth up to "> 6." Figure 2 plots the mean of forecasters' subjective uncertainty calculated using these probabilities (solid line) alongside the forecast mean (dot-dash line), plus for comparison the disagreement across forecasters (dash line). We see that both uncertainty and disagreement rose sharply during the Great Recession.

Yet another proxy for uncertainty is the frequency of newspaper articles about economic uncertainty. Figure 3 shows the Baker, Bloom, and Davis (2012) measure of economic policy uncertainty, which counts the frequency of articles containing the words "uncertain or uncertainty" and "economy or economics" and one of six policy words across ten leading US newspapers. Again, this measure is clearly countercyclical, with its level 51 percent higher on average during recessions. A related proxy is the count of the word "uncertain" in the Federal Reserve's Beige Book. The Beige Book is a 15,000 word overview of the US economy published after each meeting of the Federal Reserves Open Market Committee. Even here we see evidence for higher uncertainty in recessions: the word "uncertainty" is used 52 percent more often during recessions (Baker, Bloom, and Davis 2012).

An eclectic mix of other indicators of macro uncertainty also rises in recessions. Scotti (2013) measures the size of the surprise when economic data is released: that is, she compares the pre-release date expectations (from Bloomberg's median forecast) for categories like nonfarm payroll and quarterly GDP with their release values. She finds these surprises are 36 percent larger in recessions, suggesting forecasts are less reliable in downturns. Jurado, Ludvigson, and Ng (2013) use data on hundreds of monthly economic data series in a system of forecasting equations and look at the implied forecast errors. By their calculations, forecast errors rise dramatically in large recessions, most notably during

Figure 2

GDP Growth Forecaster Uncertainty and Disagreement Both Rose Significantly during the Great Recession

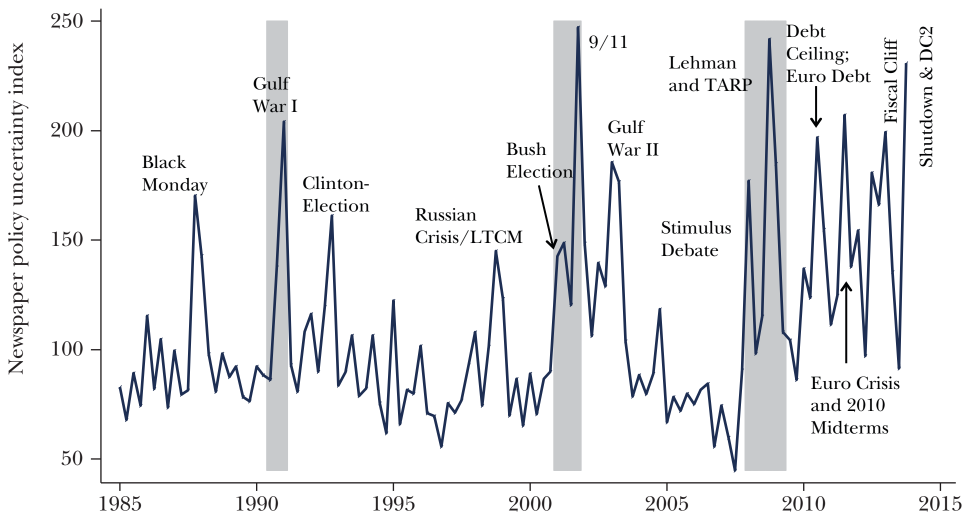


Source: Author using data on the forecaster probability distributions of GDP growth rates from the Philadelphia Survey of Professional Forecasters.

Notes: "Mean forecast" is the average forecaster's expected GDP growth rate, "Forecaster disagreement" is the cross-sectional standard deviation of forecasts, and "Forecaster uncertainty" is the median within forecaster subjective variance. Data are only available on a consistent basis since 1992Q1, with an average of 48 forecasters per quarter.

Figure 3

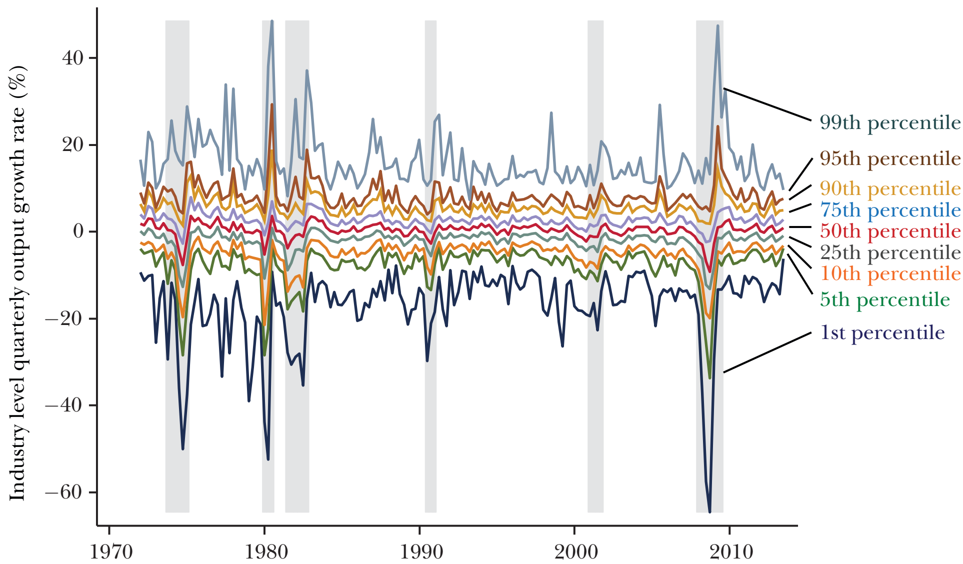
Newspaper Policy Uncertainty Index is 51 percent Higher in Recessions



Source: Data is from Baker, Bloom, and Davis (2012).

Notes: The figure shows the Baker, Bloom, and Davis (2012) measure of economic policy uncertainty, which counts the frequency of articles containing the words "uncertain or uncertainty" and "economy or economics" and one of six policy words in ten leading US newspapers. Data from 1985Q1 to 2013Q4, normalized to 100 for the period 1985 to 2009. Gray bars are NBER recessions.

Figure 4
Industry Growth Rate Spreads Increase in Recessions



Notes: The figure shows the 1st, 5th, 10th, 25th, 50th, 75th, 90th, 95th and 99th percentiles of three-month percentage growth rates of industrial production for all 196 manufacturing NAICS sectors in the Federal Reserve Boards' industry database. Data spans 1972Q1–2013Q3. Gray bars are NBER recessions.

the OPEC I recession (1973–1974), the early 1980s rust-belt recession (1982), and the Great Recession (2007–2009). Nakamura, Sergeyev, and Steinsson (2012) used over 100 years of consumption data from 16 OECD countries to estimate short- and long-run fluctuations in volatility, again finding that this volatility rises strikingly in periods of lower growth.

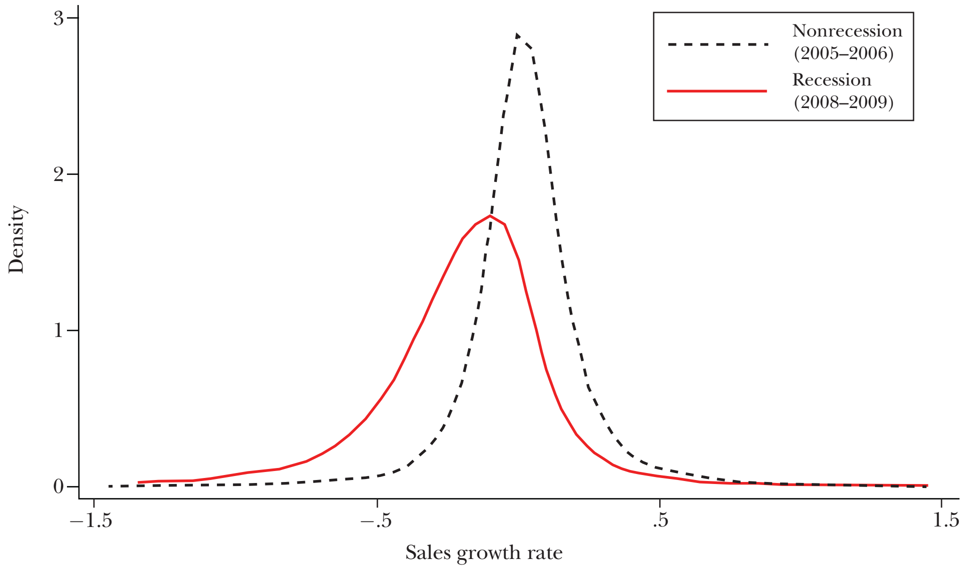
Fact 2: Micro Uncertainty Rises in Recessions

We can drop down a level of aggregation from looking at macro data to looking at micro data on individual industries, firms, and plants. At every level, uncertainty appears to rise during recessions. This result is in some senses "fractal"—that is, uncertainty rises in recession at each level of disaggregation.

For example, Figure 4 is based on a panel of about 200 manufacturing industries. The lines are based on the rate of industry output growth, and they show how different percentiles perform across these industries. During recessions, these percentiles widen out as some industries do well while others are hit hard. This increased dispersion is a proxy for industry-level uncertainty because it suggests that industries are getting larger industry-level shocks during recessions.

Uncertainty as proxied by dispersion at the firm and plant level also surges in recessions. For example, Campbell, Lettau, Malkiel, and Xu (2001) report that

Figure 5
Plant Uncertainty—Sales Growth Dispersion

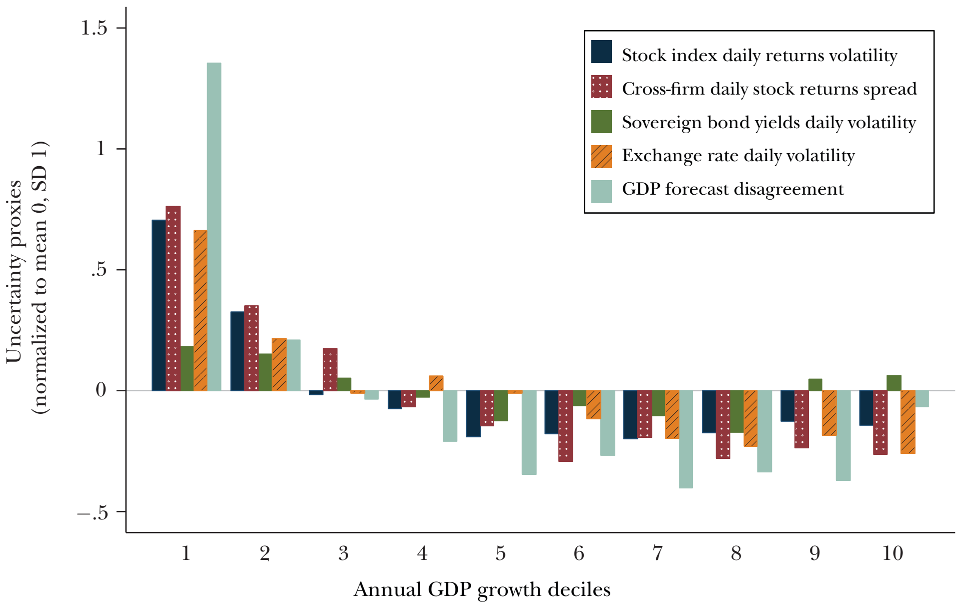


Source: Bloom, Floetotto, Jaimovich, Saporta-Eksten, and Terry (2012).
Notes: Figure 5 plots the dispersion of sales growth rates for a panel of plants within the US manufacturing for Great Recession of 2008–2009 (the solid line) against their values for the pre-recession period of 2005–2006 (the dashed line). Constructed from the Census of Manufactures and the Annual Survey of Manufactures using a balanced panel of 15,752 establishments active in 2005–2006 and 2008–2009. Moments of the distribution for nonrecession (recession) years are mean 0.026 (−0.191), variance 0.052 (0.131), coefficient of skewness 0.164 (−0.330), and kurtosis 13.07 (7.66). The year 2007 is omitted because according to the NBER the recession began in December 2007, so 2007 is not a clean "before" or "during" recession year.

cross-firm stock-return variation is almost 50 percent higher in recession than booms. Likewise, the dispersion of plant-level shocks to total factor productivity rises sharply in recessions (Kehrig, 2011; Bloom, Floetotto, Jaimovich, Saporta-Eksten, and Terry 2012). For example, Figure 5 plots the dispersion of sales growth rates for a balanced panel of about 16,000 plants within the US manufacturing sector for the Great Recession of 2008–2009 (the solid line) against their values for the pre-recession period of 2005–2006 (the dashed line). The variance of plants' sales growth rates rose by a massive 152 percent during the Great Recession, a striking jump in sales dispersion.

Digging down even further to individual product prices, yet again we find a similar story. Vavra (2013) analyzed price changes from the Bureau of Labor Statistics on tens of thousands of products, such as a one-liter bottle of Coca-Cola or a pack of four Duracell AAA batteries. They find price changes for even these kinds of items were about 50 percent more volatile during recessions.

Figure 6
Uncertainty Measures Are Countercyclical Across Countries



Source: Baker and Bloom (2013).
Notes: Figure 6 is based on annual data for 60 developing and developing countries over the period 1970 to 2012. Each country-year is placed into a bin based on the decile of their annual growth rates, with bins from 1 to 10, where 1 is the lowest decile of growth and 10 is the highest decile. So, for example, for the United States, bin 1 is growth rates of below −0.3 percent, bin 2 is growth rates of −0.3 percent to 1.2 percent, bin 3 are growth rates of 1.2 percent to 1.9 percent, and so on, while for the United Kingdom bin 1 is growth rates of below −0.8 percent, bin 2 is growth rates of −0.8 percent to 0.6 percent, and so on. The uncertainty measures plotted for each bin are averages for each country-year in the bin. Each decile shows five different measures of uncertainty: stock market volatility, firm stock-returns dispersion, bond-yield volatility, exchange rate volatility, and macro forecaster disagreement—with each measure normalized to a mean 0 and standard deviation 1.

This increase in both macro and micro uncertainty during recessions is also true on the global scale. Figure 6 is based on annual data for 60 developed and developing countries over the period 1970 to 2012. Each country-year is placed into a bin based on the deciles of a country's annual growth rates, with bins from 1 to 10 where 1 is the lowest decile of growth and 10 is the highest decile. So, for example, for the United States, bin 1 is for growth rates of below −0.3 percent, bin 2 is for growth rates of −0.3 percent to 1.2 percent, bin 3 is for growth rates of 1.2 percent to 1.9 percent, and so on, while for the United Kingdom, bin 1 is for growth rates of below −0.8 percent, bin 2 is for growth rates of −0.8 percent to 0.6 percent, and so on. The uncertainty measures plotted for each bin are averages over each country-year in the bin. Each decile shows five different measures of uncertainty: stock market volatility, firm stock-returns dispersion, bond-yield volatility, exchange rate volatility, and macro forecaster disagreement—with each measure normalized to a mean 0 and standard deviation 1.

All five of these measures of uncertainty are higher when country growth is lower, particularly when growth is in its lowest decile, which is typically during a recession. This highlights the global robustness of the link between recessions and uncertainty.

Fact 3: Wages and Income Volatility Appear to Be Countercyclical

Unemployment rises during a recession, so the volatility of household incomes will rise as well. But perhaps less expected is that wages for even those who are employed also become more volatile during recessions (Meghir and Pistaferri 2004; Storesletten, Telmer, and Yaron 2004; Heathcote, Perri, and Violante 2010). This is particularly true for lower-wage workers, whom Guvenen, Ozkan, and Song (forthcoming) show face a particularly large surge in income volatility during recession. Thus, the increasing volatility of macro, industry, firm, and plant outcomes in recessions translates into higher volatility of wages for employees.

Fact 4: Uncertainty Is Higher in Developing Countries

Low-income countries in regions like Africa and South America tend to have the most volatile GDP growth rates, stock markets, and exchange rates. In fact, the World Bank's *World Development Report 2014*, themed "Risk and Opportunity," focused on how households and firms in developing countries face a huge variety of macro and micro risks (World Bank 2013). In the panel of 60 countries with available growth and financial data I examined, those with low incomes (less than $10,000 GDP per capita) had 50 percent higher volatility of growth rates, 12 percent higher stock-market volatility, and 35 percent higher bond-market volatility, so overall developing countries experience about one-third higher macro uncertainty.

Why Does Uncertainty Vary?

What factors might be causing these variations in uncertainty? I will first focus on factors that might cause uncertainty to fluctuate over time. I'll then turn to some reasons for the higher uncertainty in low-income and emerging economies. Of course, identifying possible causes of uncertainty is only one step; the later discussion will consider evidence on the effects of uncertainty.

Bad events often seem to increase uncertainty, events like oil-price shocks, terrorist attacks, and wars. For example, in Bloom (2009), I defined 17 uncertainty shocks from 1962 to 2008 on the basis of jumps in stock market volatility and found that all but one was bad news (in that they lowered expected growth). These uncertainty shocks included the assassination of President Kennedy, the Cuban missile crisis, the OPEC oil price shocks, the 9/11 attack, and the Gulf Wars. All these dramatic shocks seemed to shake people's confidence in their forecasts of economic growth, raising macro and micro uncertainty. The only uncertainty shock in this series associated with good news was the October 1982 business cycle turning point, a relatively minor uncertainty shock.

Why does good news so rarely cause an uncertainty shock? Perhaps good news often develops more gradually—like the fall of the Berlin Wall or the development of the Internet. These change beliefs more smoothly over time instead of causing large jumps in uncertainty. Indeed, it is hard to come up with any large good news shocks in recent US history. Or alternatively, perhaps bad news itself may induce uncertainty. We know from the previous section that recessions are associated with increased uncertainty. Maybe this is because slower growth increases uncertainty.

The theory literature highlights four mechanisms through which recessions might increase uncertainty. First, when business is good, firms are trading actively, which helps to generate and spread information (Van Nieuwerburgh and Veldkamp 2006; Fajgelbaum, Schaal, and Tashereau-Dumouchel 2013). But when business is bad, this activity slows down, reducing the flow of new information and thereby raising uncertainty. Second, individuals are more confident in predicting the future when "business as usual" prevails in a growing economy. Forecasting is harder during recessions (Orlik and Veldkamp 2014). This arises from the fact that recessions are rare events, so that people are unfamiliar with them. Third, public policy that is unclear, hyperactive, or both, may raise uncertainty. Pastor and Veronesi (2011) argue that when the economy is doing well, politicians prefer to stay largely with their current policies, following the old adage "if it isn't broke, don't fix it." But when the economy turns down, politicians are tempted to experiment, elevating economic policy uncertainty. Indeed, Baker, Bloom, and Davis (2012) find that policy uncertainty rises during recessions, particularly during the Great Recession. Fourth, when business is slack, it is cheap to try out new ideas and to divert unused resources to research and development (Bachman and Moscarini 2011; D'Erasmo and Moscoso-Boedo 2011). This dynamic leads to heightened micro uncertainty, potentially feeding into higher macro uncertainty.

When considering the reasons for higher uncertainty in lower-income countries, three mechanisms are typically mentioned (Koren and Tenereyo 2007; World Bank Development Report 2013). First, developing countries tend to have less-diversified economies—for example, they may export only a small number of products—so their entire economy is more exposed to fluctuations in the output and price of those goods. Second, many of the goods on which developing countries focus also have quite volatile prices: commodities like rubber, sugar, oil, and copper. Finally, developing countries appear to have more domestic political shocks like coups, revolutions, and wars; are more susceptible to natural disasters like epidemics and floods; and have less-effective fiscal and monetary stabilization policies.

Why Might Fluctuations in Uncertainty Matter: Theory

Having established that uncertainty fluctuates over time, to what extent does this matter? I will start by discussing the theory concerning the impact of shocks to uncertainty and then turn to the empirical evidence. The theoretical literature emphasizes two negative channels for uncertainty to influence growth, but also highlights two positive channels of influence.

Real Options

The largest body of theoretical literature about the effects of uncertainty focuses on "real options" (Bernanke 1983; Brennan and Schwartz 1985; McDonald and Siegel 1986). The idea is that firms can look at their investment choices as a series of options: for example, a supermarket chain that owns an empty plot of land has the option to build a new store on the plot. If the supermarket becomes uncertain about the future—for example, because it is unsure if a local housing development will go ahead—it may prefer to wait. If the housing development proceeds, the supermarket can develop the site, and if not, it can continue to wait and avoid (for now) a costly mistake. In the language of real options, the option value of delay for the supermarket chain is high when uncertainty is high. As a result, uncertainty makes firms cautious about actions like investment and hiring, which adjustment costs can make expensive to reverse.

Investment adjustment costs have both a physical element (equipment may get damaged in installation and removal) and a financial element (the used-good discount on resale). Ramey and Shapiro (2001) and Cooper and Haltiwanger (2006) estimate these investment adjustment costs are extremely large at roughly 50 percent of the value of capital.[2] Hiring adjustment costs include recruitment, training, and severance pay, which in Nickell (1986) and Bloom (2009) are estimated at about 10 to 20 percent of annual wages. Schaal (2010) also emphasizes search frictions, showing how uncertainty can interact with search costs to impede labor markets in recessions.

However, real options effects are not universal. They arise only when decisions cannot be easily reversed; after all, reversible actions do not lead to the loss of an option. Thus, firms may be happy to hire part-time employees even when uncertainty is extremely high, because if conditions deteriorate, they can easily lay off these employees. In fact, because part-time employees are so flexible, firms may switch from hiring full-time to part-time employees during periods of high uncertainty, as indeed happens in recessions (Valetta and Bengali 2013).

Real options effects also rely on firms having the ability to wait. But if firms are racing, perhaps to be the first to patent a new idea or launch a new product, this option disappears. If delay would be extremely costly, then the option to delay is not valuable, breaking the negative real options effect of uncertainty on investment.

[2] The literature distinguishes two families of adjustment costs. There are lumpy "nonconvex" adjustment costs, which are fixed costs (a one-off cost to buy/sell capital) and partial irreversibility (a cost per unit of capital sold). These "nonconvex" adjustment costs generate real options effects. There are also smooth "convex" adjustment costs like quadratic adjustment costs (a cost that increases in the squared rate of investment), which do not generate real options. For details, see Dixit and Pindyck (1994) and Abel and Eberly (1996).

Finally, real options require that actions that are taken now influence the returns to actions taken later. But in some situations—like firms producing with a constant-returns-to-scale technology and selling into a perfectly competitive market—the choice of investment this period will have no effect on the profitability of investment next period, leading to no option value from waiting. Thus, another requirement of the real options literature is that firms are selling into imperfectly competitive markets and/or operating with a decreasing-returns-to-scale technology.

Turning from investment to consumption, an analogous channel arises for uncertainty to cause postponed consumption. When consumers are making decisions on buying durables like housing, cars, and furniture, they can usually delay purchases relatively easily (see, for instance, Eberly 1994). For example, people may be thinking about moving to another house, but they could either move this year or wait until next year. This option value of waiting will be much more valuable when income uncertainty is higher—if, for example, you are unsure about whether a major promotion will arrive by the end of this year, it makes sense to wait until this is decided before undertaking an expensive house move. Delaying purchases of nondurables like food and entertainment is harder, so the real options effects of uncertainty on nondurable consumption will be lower.

The real option argument not only suggests that uncertainty reduces *levels* of investment, hiring, and consumption, but it also makes economic actors *less sensitive* to changes in business conditions. This can make countercyclical economic policy less effective. For example, in low-uncertainty periods, the elasticity of investment with respect to interest rates might be −1, but when uncertainty is very high, this elasticity could fall to −0.25. Similarly, higher uncertainty should also make consumers' durables expenditures less sensitive to demand and prices signals, something Foote, Hurst, and Leahy (2000) and Bertola, Guiso, and Pistaferri (2005) report in studies of US and Italian consumers.

In other words, just as the economy is heading into recession, higher uncertainty can make monetary and fiscal stabilization tools less effective. Firms and consumers are likely to respond more cautiously to interest-rate and tax cuts when they are particularly uncertain about the future, dampening the impact of any potential stimulus policy. Because of this shift, stimulus policy may need to be more aggressive during periods of higher uncertainty. A related argument is that aggressive stimulus policies are helpful for reducing uncertainty by providing reassurance that the government is taking action to stabilize the economy.

This channel whereby uncertainty reduces firms' sensitivity also provides an explanation for procyclical productivity, an empirical regularity found in many modern studies of business cycles (King and Rebelo 1999). When uncertainty is high, productive firms are less aggressive in expanding and unproductive firms are less aggressive in contracting. The high uncertainty makes both of them more cautious. This caution produces a chilling effect on the productivity-enhancing reallocation of resources across firms. Because reallocation appears to drive the majority of aggregate productivity growth (for example, Foster, Haltiwanger, and Krizan 2000, 2006), higher uncertainty can stall productivity growth. This productivity impact of uncertainty shocks underlies the theories of uncertainty-driven business cycles, which emphasize how uncertainty shocks reduce investment, hiring, and productivity (Bloom et al. 2012). The difference with more-traditional real business cycle models (for example, Kydland and Prescott 1982) is that in the uncertainty-driven theory, the fall in productivity growth is an outcome of the uncertainty shock, rather than the shock itself.

Risk Aversion and Risk Premia

Investors want to be compensated for higher risk, and because greater uncertainty leads to *increasing risk premia*, this should raise the cost of finance. Furthermore, uncertainty also increases the probability of default, by expanding the size of the left-tail default outcomes, raising the default premium and the aggregate deadweight cost of bankruptcy. This role of uncertainty in raising borrowing costs can reduce micro and macro growth, as emphasized in papers on the impact of uncertainty in the presence of financial constraints (Arellano, Bai, and Kehoe 2010; Christiano, Motto, and Rostagno 2014; Gilchrist, Sim, and Zakrasjek 2011).

Another mechanism related to risk premia is the *confidence* effect of uncertainty in models where consumers have pessimistic beliefs (for example, Hansen, Sargent, and Tallarini 1999; Ilut and Schneider 2011). In these models, agents are so uncertain about the future they cannot form a probability distribution. Instead they have a range of possible outcomes and act as if the worst outcomes will occur, displaying a behavior known as "ambiguity aversion." As the range of possible outcomes (uncertainty) expands, the worst possible outcome gets worse, so agents cut back on investment and hiring. Of course this assumes agents are pessimistic, but if instead agents are optimistic (that is, they assume the best case) as Malmendier and Tate (2005) hint at for CEOs, then uncertainty can actually have a positive impact.

A rise in uncertainty risk should also lead consumers to increase their *precautionary saving*, which reduces consumption expenditure (for example, Bansal and Yaron 2004). This effect is likely contractionary for an economy in the short run, but the long-run effects are less clear. After all, at least in theory, lower consumption and greater saving may allow a rise in investment, which could then benefit long-term growth. However, in most open economies some of this increased saving will flow abroad, reducing domestic demand. For this reason, Fernández-Villaverde, Guerrón-Quintana, Rubio-Ramirez, and Uribe (2011) argue that rising uncertainty can be crippling for growth in smaller highly open countries, as domestic money flees the country.

What about the effect of a rise in precautionary saving in larger and more closed countries like the United States? At first, it would seem that uncertainty may have potentially positive effects—by encouraging consumers to save, this will increase investment (because savings equals investment in closed economies). But as several recent papers have noted, if prices are sticky (as New Keynesian models commonly assume), uncertainty shocks can lead to recessions even in closed economies because prices do not fall enough to clear markets (for example, Leduc and Liu 2012; Basu and Bundick 2011; Fernández-Villaverde, Guerrón-Quintana, Kuester, and Rubio-Ramirez 2011). The intuition is that uncertainty increases the desire of consumers to save, which should cut interest rates and output prices, stimulating an offsetting rise in investment; but if prices are sticky, this effect does not happen—prices and interest rates do not fall enough to encourage the offsetting rise in investment—so that output falls. This effect of uncertainty can be particularly damaging if interest rates are constrained at zero by the lower bound, as has been the case during much of the last five years.

Another precautionary effect of uncertainty may affect firms through the incentives of their chief executive officers. Most top corporate executives are not well diversified: both their personal financial assets and their human capital are disproportionately tied up in their firm. Hence, when uncertainty is high, these executives may become more cautious in making long-run investments. For example, the chief executive officer of an oil exploration company may become increasingly nervous when the price of oil becomes volatile, leading that firm to take a more cautious position on oil exploration. Panousi and Pananikolaou (2012) have shown in a panel of US firms that when uncertainty is higher, investment drops, particularly in firms where the chief executive officers hold extensive equity in the firm and so are highly exposed to firm-level risk.

Growth Options

There are two mechanisms through which uncertainty can potentially have a positive effect on long-run growth. The "growth options" argument is based on the insight that uncertainty can encourage investment if it increases the size of the potential prize. For example, Bar-Ilan and Strange (1996) note that if firms have long delays in completing projects—perhaps because of time-to-build or time-to-develop—then uncertainty can have a positive effect on investment. As an illustration, consider a pharmaceutical company developing a new drug that notices that a mean-preserving increase in demand uncertainty has occurred. The costs of bad draws (for example, the drug turns out to be ineffective or unsafe) have a limited lower bound because the firm can cancel the product losing only its sunk research and development costs. But good draws (the product turns out to be even more useful and profitable than expected) are not constrained in this way. In this situation, a rise in mean-preserving risk means higher expected profit when the product goes to market.[3]

Growth options were often invoked to explain the dot-com boom of the late 1990s. Firms were unsure about the Internet but that uncertainty encouraged investment. The worst outcome for firms starting new websites was losing their development costs, while the best outcome looked ever more profitable as the range

[3] This is sometimes called the "good news principle" that only good news matters in growth options because bad news is capped by closing down the project. This phrase originates from Bernanke (1983), who discussed the reverse "bad news principle" in terms of the classic real-options negative effects of uncertainty on investment. In a recent working paper, Segal, Shaliastovich and Yaron (2013) find interesting evidence for both these good news (growth option) and bad news (real option) effects of uncertainty in aggregate investment.

of uncertainty about the Internet expanded. Because developing websites took time, building one was seen as investing in a "call-option" on the future success of the Internet. Likewise, a literature on the value of oil drilling leases shows how these are call options on possible future extraction, so oil price uncertainty increases their value (Paddock, Siegel, and Smith 1988). More recently Kraft, Schwartz, and Weiss (2013) have shown how growth options are particularly important for research and development–intensive firms, so much so that higher uncertainty can raise their stock value.

Oi–Hartman–Abel Effects

The other channel I examine through which uncertainty can potentially increase growth is known as the Oi–Hartman–Abel effect (after Oi 1961; Hartman 1972; Abel 1983). This effect highlights the possibility that if firms can expand to exploit good outcomes and contract to insure against bad outcomes, they may be risk loving. For example, if a factory can easily halve production volumes if the price of its products falls and double production volumes if the price rises, it should desire a mean-preserving increase in uncertainty, because the firm gets 50 percent during bad outcomes and 200 percent during good outcomes. In effect, the factory is partly insured against bad outcomes by being able to contract and has the option to increase its advantage from good outcomes by being able to expand. (Formally, if profits are convex in demand or costs, then demand or cost uncertainty increases expected profits.) However, for this mechanism to work, firms need to be able to easily expand or contract in response to good or bad news, so while the Oi–Hartman–Abel effects are typically not very strong in the short run (because of adjustment costs), they can be more powerful in the medium and long run.

How Much Might Fluctuations in Uncertainty Matter: Empirics

The evidence on the impact of uncertainty is limited because of the difficulties in stripping out cause and effect. A central challenge in the uncertainty literature (as in macroeconomics as a whole) is to distinguish the impact of uncertainty from the impact of recessions. We know that uncertainty moves with the business cycle, which raises the question of how to distinguish the separate causal effects of higher uncertainty.

To identify the causal impact of uncertainty on firms and consumers, the literature has taken three approaches. One approach relies on timing: that is, estimating the movements in output, hiring, and investment that follow jumps in uncertainty. This approach works reasonably well for unexpected shocks to uncertainty but is more problematic if changes in uncertainty are predicted in advance or are correlated with other unobserved factors. A second approach uses structural models calibrated from macro and micro moments to quantify the potential effect of uncertainty shocks. This approach is conceptually well grounded, but like many structural models, it is sensitive to somewhat debatable modelling assumptions. A third approach exploits natural experiments like disasters, political coups, trade changes, or movements in energy and exchange rates. The challenge here is over the generalizability of these results, and the extent to which these events influence firms and consumers beyond just changes in uncertainty.

My overall view is that this literature provides suggestive but not conclusive evidence that uncertainty damages short-run (quarterly and annual) growth, by reducing output, investment, hiring, consumption, and trade. The longer-run evidence of the effect of uncertainty on output is far more limited, and while my personal view is that uncertainty is damaging for growth, it is extremely hard to show this definitively. One reason is that while uncertainty appears to reduce short-run hiring and investment, it may also stimulate research and development, as some recent empirical work suggests. This may be because of the "growth options effect"—the idea that uncertainty increases the upside from innovative new products. As such, more empirical work on the effects of uncertainty would be valuable, particularly work which can identify clear causal relationships.

Timing Approaches to Estimating the Effect of Uncertainty Shocks

A standard approach in macroeconomic analysis has been to look at short-term economic fluctuations separately from long-term trends in economic growth. The classic macro study of uncertainty by Ramey and Ramey (1995) challenged this separation. They looked both at a broad sample of 92 countries from 1960 to 1985 and also at a narrower sample of high-income countries from 1950 to 1988. They considered an equation for forecasting GDP by country, and find that economies which depart most strongly from that forecast equation—an idea that they equate with a rise in uncertainty—experience lower growth rates. This negative volatility link with growth has been confirmed in a number of subsequent studies using more advanced estimations techniques (Engel and Rangel 2008) or different measures of uncertainty (Bloom 2009).

Other studies have considered how rising uncertainty might affect other macroeconomic outcomes. For example, Romer (1990) argues that the uncertainty created by the stock market crash of 1929 led to a drop in consumer spending on durable goods. Indeed, she finds a negative correlation between stock market volatility and purchases of consumer durables throughout the prewar period. Handley and Limão (2012) model the role of uncertainty in how firms make investment choices related to export markets. When they apply the model to the example of Portugal joining the European Community in 1986, they find that the removal of uncertainty accounted for a substantial rise in firm investment spending. Finally, Novy and Taylor (2014) use US data since the 1960s to examine the differential impact of uncertainty shocks across sectors to show that uncertainty significantly depresses trade flows and that this effect may explain about half of the collapse of global trade in 2008–2009.

A corresponding micro literature focuses on how uncertainty affects individual firms and households, again typically finding that higher uncertainty has a negative impact. For example, Leahy and Whited (1996) examined a panel of several hundred US publicly listed manufacturing firms and found a strong relationship between uncertainty, proxied by the stock-price volatility for that firm, and investment, which they argue is consistent with theories of firms looking at investment as an irreversible choice. In Bloom, Bond, and Van Reenen (2007), we confirmed this result in data for 672 UK manufacturing firms from 1972–1991, using lagged firm accounting and financial data outcomes as instruments. Guiso and Parigi (1999) used a survey of Italian firms in 1993 in which the firms themselves reported the distribution of their expectations of future demand, and using this measure of uncertainty, they find a large negative relationship between uncertainty and investment.

Structural Models Estimating the Effect of Uncertainty Shocks

One structural approach is to build micro-to-macro general equilibrium models of firms and the economy, calibrating the key parameters against micro and macro data moments. For example, in Bloom et al. (2012) we build a general equilibrium model with heterogeneous firms with labor and capital adjustment costs and countercyclical micro and macro uncertainty. We find that the average increase in uncertainty that happens during recessions reduces output by about 3 percent during the first year, but with a rapid recovery in the second year. The reason for this rapid drop in output is that higher uncertainty leads firms to pause hiring and investment, cutting aggregate capital and labor through depreciation and attrition. Productivity growth also drops as reallocation freezes (productive plants do not expand and unproductive plants do not contract). However, once uncertainty starts to drop, pent-up demand for hiring and investment leads to a rapid rebound. Hence, uncertainty shocks generate short, sharp drops and rebounds in output.

These results, however, appear sensitive to assumptions on some of the parameter values in the model. For example, Bachmann and Bayer (2012, 2013) model general equilibrium models with heterogeneous agents and capital adjustment, finding much smaller impacts of uncertainty on growth. Their models differ from ours in that they exclude labor adjustment costs, place more weight on micro compared to macro uncertainty shocks, and have smaller fluctuations in uncertainty. Which set of assumptions is right is not obvious, and this highlights the need for richer micro and macro models to pin down these types of questions.

Another structural approach models individual firms' behavior, such as Kellogg's (forthcoming) study of drilling oil wells in Texas. He finds that jumps in oil price uncertainty lead firms to pause new drilling activity, with this response to uncertainty increasing their expected value from drilling new oil wells by up to 25 percent. Hence, for oil firms, it is extremely important to consider both the level and the uncertainty of future oil prices before drilling wells. Intriguingly, Kellogg also shows firms appear to use oil futures and derivatives from the New York Mercantile Exchange to predict future oil prices and volatility (rather than simply extrapolating from historic prices), suggesting sophisticated forward-looking behavior on the part of drilling firms.

Using Natural Experiments to Estimate the Effect of Uncertainty

A recent approach to estimating the impact of uncertainty shocks has tried to exploit various macro and micro natural experiments. For example, in Baker and Bloom (2013), we sought to use natural disasters, terrorist events, and political shocks as instruments for uncertainty. We defined these events in terms of a minimum share of the population killed, a minimum share of GDP lost, or as resulting in a political regime change and considered data from 60 countries from 1970–2012. Stock market and news data shows that these events were not anticipated. We use the events to predict stock market volatility, and then use the stock market volatility that can be predicted from these shocks to forecast GDP growth. Across countries, the rise in volatility from these events explains about half of the variation in growth.

In another approach along these lines, Stein and Stone (2012) used the exposure of US firms to exogenous variations in energy and currency volatility as an instrument for the uncertainty that they face. They find that those firms exposed to greater uncertainty have lower investment, hiring, and advertising. Indeed, they estimate that uncertainty accounts for roughly a third of the fall in capital investment and hiring that occurred in 2008–2010, a subject taken up in greater detail in the next section. Interestingly, they also find that uncertainty seems to increase research and development spending, something that the growth options mechanism—the idea the more uncertainty yields a larger upside for long-run growth—can explain.

Has Higher Uncertainty Worsened the Great Recession and Slowed the Recovery?

Finally, I turn to the question of the importance of uncertainty in driving the recent Great Recession and sluggish recovery. Certainly, policymakers believe uncertainty has played an important role. For example, the Federal Reserve Open Market Committee (2008) noted that "participants reported that uncertainty about the economic outlook was leading firms to defer spending projects until prospects for economic activity became clearer." In 2009, Chief Economist of the International Monetary Fund (IMF) Olivier Blanchard wrote in *The Economist:* "Uncertainty is largely behind the dramatic collapse in demand. Given the uncertainty, why build a new plant, or introduce a new product? Better to pause until the smoke clears." The Chair of the Council of Economic Advisers, Christina Romer, noted in her 2009 testimony to the US Congress Joint Economic Committee: "Volatility, according to some measures, has been over five times as high over the past six months as it was in the first half of 2007. The resulting uncertainty has almost surely contributed to a decline in spending."

Such claims about the damaging impact of uncertainty have continued, with policymakers arguing it has also been responsible for the slow recovery. For example, in 2012 the IMF Managing Director, Christine Lagarde, argued: "There is a level of uncertainty which is hampering decision makers from investing and from creating jobs" (IMF 2012). A joint European Union and OECD article in 2013 similarly noted that "high uncertainty is all the more damaging for growth as it magnifies the effect of credit constraints and weak balance sheets, forcing banks to rein in credit further and companies to hold back investment" (Buti and Padoan 2013). The International Labor Organization (ILO 2013) argued that "indecision of policy makers in several countries has led to uncertainty about future conditions and reinforced corporate tendencies to increase cash holdings or pay dividends rather than expand capacity and hire new workers."

But while policymakers clearly think uncertainty has played a central role in driving the Great Recession and slow recovery, the econometric evidence is really no more than suggestive. It is certainly true that every measure of economic uncertainty rose sharply in 2008. As one might guess from Figures 1 to 4, the level of uncertainty around 2008–2009 was more than triple the size of an average uncertainty shock and about twice as persistent as during an average recession. This jump in uncertainty reflects its role as both an impulse and a propagation mechanism for recessions. The shocks initiating the Great Recession—the financial crisis and the housing collapse—increased uncertainty. In particular, it was unclear how serious the financial and housing problems were, or what their impact would be nationally and globally, or what the appropriate policy responses should be. Furthermore, the Great Recession itself further increased uncertainty, leading the initial slowdown to be propagated and amplified over time.

For a rough calculation of the magnitude of the impact of uncertainty, I start with the drop in GDP during the Great Recession. This appears to be about 9 percent, consisting of the 3 percent drop in GDP over 2008 and 2009 versus the 6 percent rise that would have occurred if GDP had followed trend growth. Next, we need to estimate the impact of uncertainty on GDP growth. We can do this several ways, all of which yield reassuringly similar answers of about a 3 percent drop in GDP (around one-third of the total decline). One way is to take the 1 percent drop in GDP as estimated from vector autoregressions after an *average* uncertainty shock (Bloom 2009), and triple this, remembering that the 2008–2009 rise in uncertainty was about triple the "normal" uncertainty shock. Another approach is to take the structural model estimates from Bloom et al. (2012) of a 1.3 percent drop in GDP in the year after an *average* recessionary uncertainty shock, and again triple this. Finally, we can use the estimates from Stein and Stone (2012) who aggregate up from micro-data instrumental variable results, again finding that an uncertainty shock the size of the one experienced during the Great Recession reduced output by about 3 percent.

Concluding Thoughts

A range of evidence shows that uncertainty rises strongly in recessions, at both the macro and micro levels. More speculatively, I have argued this is because increases in uncertainty are both part of the impulse arising from bad news shocks that start recessions, and because uncertainty amplifies recessions by rising further as growth slows.

The empirical literature on uncertainty is still at an early stage with many open research questions. First and most immediately, the question over the causality of uncertainty and growth is still unclear, and more work exploiting both natural experiments and structural models would be very valuable. Second, our measures of uncertainty are far from perfect and in fact are best described as proxies rather than real measures. Developing a wider set of uncertainty measures is important. For example, there is little data on the time horizon of uncertainty (short-run versus long-run uncertainty), on types of uncertainty (demand versus supply, technology versus policy), or on the nature of uncertainty (risk versus Knightian).

The literature on the policy implications of uncertainty is also at an early stage. The basic lessons seem to be twofold. First, uncertainty shocks appear to lead to short, sharp drops and recoveries in output, which if a policymaker wanted to stabilize, would require a similarly short, sharp macroeconomic stimulus to achieve stabilization. Second, policy should try to address the root cause of the uncertainty—an approach more likely to be effective than treating the symptoms (the drop in output). For example, during the Great Recession, I believe that one of the most important policy responses was to stabilize the financial system, helping to stem the rise in financial uncertainty.

But many policy questions remain. If public policy becomes more rule based, would this help to reduce policy uncertainty, or, by limiting flexibility, would rules impede the ability of policymakers to address uncertainty by judicious interventions? For example, quantitative easing has been used heavily by US monetary authorities to try and stabilize demand, but is clearly different from the recent history of interest rate manipulation. If public policy was communicated more transparently, would this act to reduce uncertainty, or would it introduce greater volatility by generating more frequent jumps in financial markets after each policy pronouncement? The Federal Reserve is grappling with these questions as it seeks to be more transparent in signaling the path of monetary policy.

While the empirical progress on fluctuations in uncertainty over the last decade has been exciting, there is still much about uncertainty about which we remain uncertain.

■ *I would like to thank David Autor, Steve Davis, Itay Saporta-Eksten, Luke Stein, Timothy Taylor, Stephen Terry, and Ian Wright for detailed comments on the review. I would like to thank the National Science Foundation for ongoing funding which supported this research.*

References

Abel, Andrew B. 1983. "Optimal Investment Under Uncertainty." *American Economic Review* 73(1): 228–33.

Abel, Andrew B., and Janice C. Eberly. 1996. "Optimal Investment with Costly Reversibility." *Review of Economic Studies* 63(4): 581–93.

Alexopoulos, Michelle, and John Cohen. 2009. "Uncertain Times, Uncertain Measures." Working Paper 352, University of Toronto, Department of Economics. February.

Arellano, Cristina, Yan Bai, and Patrick Kehoe. 2010. "Financial Markets and Fluctuations in Uncertainty." Federal Reserve Bank of Minnesota Research Department Staff Report.

Bachmann, Ruediger, and Christian Bayer. 2012. "Investment Dispersion and the Business Cycle." NBER Working Paper 16861.

Bachmann, Rüdiger, and Christian Bayer. 2013. "Wait-and-See Business Cycles?" *Journal of Monetary Economics* 60(6): 704–719.

Bachmann, Ruediger, Steffen Elstner, and Erik R. Sims. 2010. "Uncertainty and Economic Activity: Evidence from Business Survey Data." NBER Working Paper 16143.

Bachmann, Ruediger, and Giuseppe Moscarini. 2011. "Business Cycles and Endogenous Uncertainty." 2011 Meeting Papers no. 36, Society for Economic Dynamics.

Baker, Scott R., and Nicholas Bloom. 2013. "Does Uncertainty Drive Business Cycles? Using Disasters as Natural Experiments." NBER Working Paper 19475.

Baker, Scott R., Nicholas Bloom, and Steve J. Davis. 2012. "Measuring Economic Policy Uncertainty." Unpublished paper.

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

Bar-Ilan, Avner, and William Strange. 1996. "Investment Lags." *American Economic Review* 86(3): 610–22.

Basu, Susanto, and Brent Bundick. 2011. "Uncertainty Shocks in a Model of Effective Demand." Boston College Working Papers in Economics 774, September.

Bekaert, Geert, Marie Hoerova, and Marco Lo Duca. 2013. "Risk, Uncertainty, and Monetary Policy." *Journal of Monetary Economics* 60(7): 771–88.

Bernanke, Ben S. 1983. "Irreversibility, Uncertainty, and Cyclical Investment." *Quarterly Journal of Economics* 98(1): 85–106.

Bertola, Giuseppe, Luigi Guiso, and Luigi Pistaferri. 2005. "Uncertainty and Consumer Durables Adjustment." *Review of Economic Studies* 72(4): 973–1007.

Blanchard, Olivier. 2009. "(Nearly) Nothing to Fear but Fear Itself." *The Economist,* January 29.

Bloom, Nicholas. 2009. "The Impact of Uncertainty Shocks." *Econometrica* 77(3): 623–85.

Bloom, Nick, Stephen Bond, and John Van Reenen. 2007. "Uncertainty and Investment Dynamics." *Review of Economic Studies* 74(2): 391–415.

Bloom, Nicholas, Max Floetotto, Nir Jaimovich, Itay Saporta-Eksten, and Stephen J. Terry. 2012. "Really Uncertain Business Cycles." NBER Working Paper 18245.

Brennan, Michael J., and Eduardo S. Schwartz. 1985. "Evaluating Natural Resource Investments." *Journal of Business* 58(2): 135–57.

Buti, Marco, and Pier Carlo Padoan. 2013. "How to Make Europe's Incipient Recovery Durable: End Policy Uncertainty." VOX, September 12.

Campbell, John Y., Martin Lettau, Burton G. Malkiel, and Yexiao Xu. 2001. "Have Individual Stocks become More Volatile? An Empirical Exploration of Idiosyncratic Risk." *Journal of Finance* 56(1): 1–43.

Christiano, Lawrence J., Roberto Motto, and Massimo Rostagno. 2014. "Risk Shocks." *American Economic Review* 104(1): 27–65.

Cooper, Russell W., and John C. Haltiwanger. 2006. "On the Nature of Capital Adjustment Costs." *Review of Economic Studies* 73(3): 611–33.

D'Erasmo, Pablo N., and Hernan J. Moscoso-Boedo. 2011. "Intangibles and Endogenous Firm Volatility over the Business Cycle." Virginia Economics Online Papers no. 400.

Dixit, Avinash K., and Robert S. Pindyck. 1994. "Investment under Uncertainty." Princeton: Princeton University Press.

Eberly, Janice. 1994. "Adjustment of Consumers' Durables Stocks: Evidence from Automobile Purchases." *Journal of Political Economy* 102(3): 403–36.

Engle, Robert, and J. Gonzalo Rangel. 2008. "The Spline-GARCH Model for Low-Frequency Volatility and Its Global Macroeconomic Causes." *Review of Financial Studies* 21(3): 1187–1222.

Fajgelbaum, Pablo, Eduardo Schaal, and Mathieu Taschereau-Dumouchel. 2014. "Uncertainty Traps."

Federal Open Market Committee (FOMC). 2008. Federal Open Market Committee Minutes, April 29–30, 2008.

Fernández-Villaverde, Jesús, Pablo Guerrón-Quintana, Keith Kuester, and Juan Rubio-Ramirez. 2011. "Fiscal Volatility Shocks and Economic Activity." NBER Working Paper 17317.

Fernández-Villaverde, Jesús, Pablo Guerrón-Quintana, Juan F. Rubio-Ramirez, and Martin Uribe. 2011. "Risk Matters: The Real Effects of Volatility Shocks." *American Economic Review* 101(6): 2530–61.

Foote, Christopher, Erik Hurst, and John Leahy. 2000. "Testing the (*S*, *s*) Model." *American Economic Review* 90(2): 116–19.

Foster, Lucia, John Haltiwanger, and C. J. Krizan. 2000. "Aggregate Productivity Growth: Lessons From Microeconomic Evidence." Chap. 8 in *New Developments in Productivity Analysis*, edited by C. R. Hulten, E. R. Dean, and M. J. Harper. University of Chicago Press.

Foster, Lucia, John Haltiwanger, and C. J. Krizan. 2006. "Market Selection, Reallocation and Restructuring in the U.S. Retail Trade Sector in the 1990s." *Review of Economics and Statistics* 88(4): 748–58.

Gilchrist, Simon, Jae Sim, and Egon Zakrajsek. 2011. "Uncertainty, Financial Frictions and Investment Dynamics." Unpublished paper.

Guiso, Luigi, and Guiso Parigi. 1999. "Investment and Demand Uncertainty." *Quarterly Journal of Economics* 114(1): 185–227.

Guvenen, Fatih, Serdar Ozkan, and Jae Song. Forthcoming. "The Nature of Countercyclical Income Risk." *Journal of Political Economy.*

Handley, Kyle, and Nuno Limão. 2012. "Trade and Investment under Policy Uncertainty: Theory and Firm Evidence." NBER Working Paper 17790.

Hansen, Lars Peter, Thomas J. Sargent, and Thomas D. Tallarini. 1999. "Robust Permanent Income and Pricing." *Review of Economic Studies* 66(4): 873–907.

Hartman, Richard. 1972. "The Effects of Price and Cost Uncertainty on Investment." *Journal of Economic Theory* 5(2): 258–66.

Heathcote, Jonathan, Fabrizio Perri, and Gianluca L. Violante. 2010. "Unequal We Stand: An Empirical Analysis of Economic Inequality in the United States: 1967–2006." *Review of Economic Dynamics* 13(1): 15–51.

ILO. 2013. "Global Employment Trends 2013: Recovering from a Second Jobs Dip." Geneva: International Labour Office.

Ilut, Cosmin, and Martin Schneider. 2011. "Ambiguous Business Cycles." NBER Working Paper 17900.

International Monetary Fund (IMF). 2012. "Global Recovery, Growth Hampered by Uncertainty—Lagarde." *IMF Survey Magazine,* October 11.

Jurado, Kyle, Sydney C. Ludvigson, and Serena Ng. 2013. "Measuring Uncertainty." NBER Working Paper 19456.

King, Robert, and Sergio Rebelo. 1999. "Resuscitating Real Business Cycles." In *Handbook of Macroeconomics*, edited by John B. Taylor and Michael Woodford, 928–1002. Elsevier.

Kehrig, Matthias. 2011. "The Cyclicality of Productivity Dispersion." US Census Bureau Center for Economic Studies Paper no. CES-WP-11-15.

Kellogg, Ryan. Forthcoming. "The Effect of Uncertainty on Investment: Evidence from Texas Oil Drilling." *American Economic Review.*

Knight, Frank H. 1921. *Risk, Uncertainty, and Profit*. Boston, MA: Hart, Schaffner & Marx; Houghton Mifflin Company.

Koren, Miklós, and Sylvania Tenreyro. 2007. "Volatility and Development." *Quarterly Journal of Economics* 122(1): 243–87.

Kraft, Holger, Eduardo S. Schwartz, and Farina Weiss. 2013. "Growth Options and Firm Valuation." NBER Working Paper 18836.

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

Leahy, John, and Toni Whited. 1996. "The Effect of Uncertainty on Investment: Some Stylized Facts." *Journal of Money, Credit and Banking* 28(1): 64–83.

Leduc, Sylvain, and Zhang Liu. 2012. "Uncertainty Shocks are Aggregate Demand Shocks." Federal Reserve Bank of San Francisco Working Paper 2012-10.

Malmendier, Ulrike, and Geoff Tate. 2005. "CEO Overconfidence and Corporate Investment." *Journal of Finance* 60(6): 2661–2700.

McDonald, Rob, and Daniel Siegel. 1986. "The Value of Waiting to Invest." *Quarterly Journal of Economics* 101(4): 707–728.

Meghir, Costas, and Luigi Pistaferri. 2004. "Income Variance Dynamics and Heterogeneity." *Econometrica* 72(1): 1–32.

Nakamura, Emi, Dmitriy Sergeyev, and Jón Steinsson. 2012 "Growth-Rate and Uncertainty Shocks in Consumption: Cross-Country Evidence." NBER Working Paper 18128.

Nickell, Stephen J. 1986. "Dynamic Models of Labor Demand." Chap 9 in *Handbook of Labor Economics*, Vol. 1, ed. by Orley C. Ashenfelter and Richard Layard. Amsterdam: North-Holland.

Novy, Dennis, and Alan M. Taylor. 2014. "Trade and Uncertainty." NBER Working Paper 19941.

Oi, Walter Y. 1961 "The Desirability of Price Instability under Perfect Competition." *Econometrica* 29(1): 58–64.

Orlik, Anna, and Laura Veldkamp. 2014. "Understanding Uncertainty Shocks and the Role of the Black Swan." http://people.stern.nyu.edu/lveldkam/pdfs/uncertaintyOV.pdf.

Paddock, James L., Daniel R. Siegel, and James L. Smith. 1988. "Option Valuation of Claims on Real Assets: The Case of Offshore Petroleum Leases." *Quarterly Journal of Economics* 103(3): 479–508.

Panousi, Vasia, and Dimitri Papanikolaou. 2012. "Investment, Idiosyncratic Risk, and Ownership." *Journal of Finance* 67(3): 1113–48.

Pastor, Lubos, and Pietro Veronesi. 2011. "Political Uncertainty and Risk Premia." Working Paper 17464.

Ramey, Gary, and Valerie A. Ramey. 1995. "Cross-Country Evidence on the Link between Volatility and Growth." *American Economic Review* 85(5): 1138–51.

Ramey, Valerie, and Matthew Shapiro. 2001. "Displaced Capital: A Study of Aerospace Plant Closings." *Journal of Political Economy* 109(5): 958–92.

Romer, Christina. 1990. "The Great Crash and the Onset of the Great Depression." *Quarterly Journal of Economics* 105(3): 597–624.

Romer, Christina. 2009. Testimony to the U.S. Congress Joint Economic Committee at their hearings on the Economic Outlook in the United States, April 30th.

Schaal, Edouard. 2010. "Uncertainty, Productivity and Unemployment in the Great Recession." Unpublished paper.

Schwert, G. William. 1989. "Why Does Stock Market Volatility Change over Time?" *Journal of Finance* 44(5): 1115–53.

Scotti, Chiara. 2013. "Surprise and Uncertainty Indexes: Real-time Aggregation of Real-activity Macro Surprises." International Finance Discussion Paper 1093, Board of Governors of the Federal Reserve System.

Segal, Gill, Ivan Shaliastovich, and Amir Yaron. 2013. "Good and Bad Uncertainty: Macroeconomic and Financial Market Implications." Unpublished paper.

Stein, Luke C. D., and Elizabeth Stone. 2012. "The Effect of Uncertainty on Investment, Hiring, and R&D: Causal Evidence from Equity Options." Unpublished paper.

Storesletten, Kjetil, Chris I. Telmer, and Amir Yaron. 2004. "Cyclical Dynamics in Idiosyncratic Labor Market Risk." *Journal of Political Economy* 112(3): 695–717.

Valletta, Rob, and Leila Bengali. 2013. "What's Behind the Increase in Part-time Work?" San Francisco Federal Reserve Bank Economic Letter, August 26.

Van Nieuwerburgh, Stijn, and Laura Veldkamp. 2006. "Learning Asymmetries in Real Business Cycles." *Journal of Monetary Economics* 53(4): 753–72.

Vavra, Joseph. 2013. "Inflation Dynamics and Time-varying Uncertainty: New Evidence and an Ss Interpretation." Unpublished paper.

World Bank. 2013. *World Bank Development Report 2014: Risk and Opportunity*. World Bank.

Journal of Economic Perspectives—Volume 28, Number 2—Spring 2014—Pages 177–196

The Market for Blood†

Robert Slonim, Carmen Wang, and Ellen Garbarino

Donating blood, "the gift of life," is among the noblest activities and it is performed worldwide nearly 100 million times annually (World Health Organization 2011). Massive blood donations after disasters—like the terrorist attacks on September 11, 2001, Hurricane Katrina in 2005, the Australian bushfires in 2009—exemplify human empathy and altruism. Unfortunately, because most such disasters only minimally affect demand for blood, spikes in blood donation after such disasters result in excess supply and (given blood's limited shelf-life) have led later to destruction of supply (Starr 2002). Conversely, seasonal supply shortages of blood in winter and around holidays are more common. These supply and demand imbalances are not surprising given the lack of market prices (and shadow values) for collecting blood in many countries where donations are predominantly voluntary.

The economic perspective presented here shows how the gift of life, albeit noble and often motivated by altruism, is heavily influenced by standard economic forces including supply and demand, economies of scale, and moral hazard. These forces, shaped by technological advances, have driven the evolution of blood donation markets from thin one-to-one "marriage markets," in which each recipient needed a personal blood donor, to thick, impersonalized, diffuse markets. Today, imbalances between aggregate supply and demand are a major challenge in blood markets, including excess supply after disasters and insufficient supply at other times. These

■ *Robert Slonim is Professor of Economics, University of Sydney, Sydney, Australia. Carmen Wang is a Ph.D. student in economics, Harvard Business School, Boston, Massachusetts. Ellen Garbarino is Professor of Marketing, University of Sydney Business School, Sydney, Australia. Slonim is the corresponding author at robert.slonim@sydney.edu.au.*

† To access the Appendix and disclosure statements, visit
http://dx.doi.org/10.1257/jep.28.2.177 doi=10.1257/jep.28.2.177

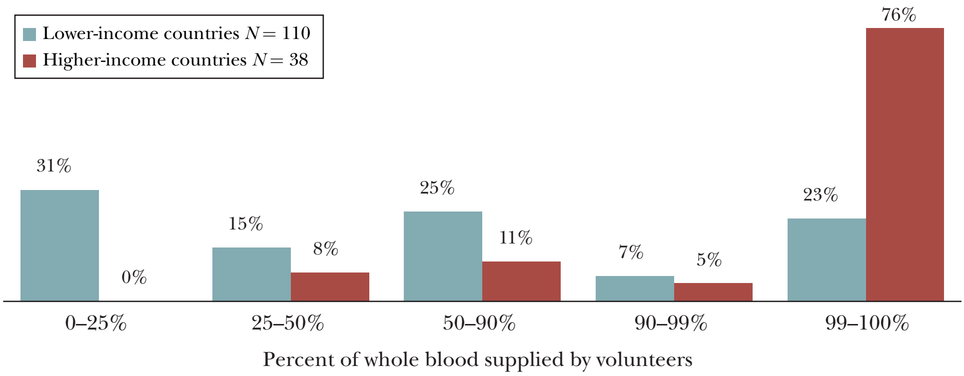
imbalances are not unexpected given that the blood market operates without market prices and with limited storage length (about six weeks) for whole blood. Yet shifting to a system of paying blood donors seems a practical impossibility given attitudes toward paying blood donors and concerns that a paid system could compromise blood safety. Nonetheless, we believe that an economic perspective offers promising directions to increase supply and improve the supply and demand balance even in the presence of volunteer supply and with the absence of market prices.

Background

Blood products, which include whole blood, platelets, plasma, and its fractionated components, provide supplies for transfusions, surgeries, and many routine treatments. The current annual worldwide supply of whole blood is roughly 100 million units at 450 milliliters per unit (World Health Organization 2011). Transfusions of blood and plasma have saved tens of millions of lives, more than doubled the life expectancy of hemophiliacs, and improved health outcomes for many more people (Starr 1998; Hayes 2006). Even with a largely voluntary supply of blood, the blood industry can be regarded as a multibillion–dollar market because hospitals pay for blood products and charge patients for their use. For example, the cost of the components of each unit of blood sold to hospitals in the United States is approximately \$570, with the cost for red blood cells at \$229, platelets at \$300, and plasma at \$40 (Tracy 2010). Hospitals transfuse this blood at estimated costs of between \$522 and \$1,183 per unit in the United States and Europe (Shander et al. 2010; Abraham and Sun 2012). Of course, these prices are likely to underestimate social welfare because they ignore consumer surplus—suffering diminished and lives saved.

Systems for collecting blood are diverse across and within many countries. Wealthy countries rely heavily on unpaid volunteers for whole blood. Volunteer blood collection systems fall into four general subcategories: state-run monopolies, like Britain, France, Ireland, New Zealand, Canada; Red Cross-run monopolies, like Australia, Belgium, Luxembourg, The Netherlands; majority Red Cross-controlled, like the United States, Germany, and Austria; and majority independent blood banks, like Denmark, Italy, Norway, Portugal, and Spain. Healy (2000) discusses these categories, but finds few differences in outcomes between these systems. Several high-income countries also collect plasma through voluntary donation, like Australia, Belgium, France, New Zealand, and Japan, while others at least partially compensate suppliers, including the United States, Germany, Austria, and Lithuania (Eastlund 1998; Farrugia, Penrod, and Bult 2010). In the United States, with the highest percent of plasma products collected from paid suppliers, 81 percent of US plasma products were derived from compensated donors by 2004 (Flood et al. 2006). In poorer countries, blood typically comes from paid donors and "emergency-replacement" donors who are associated with recipients (and usually family and friends).

Figure 1
Distribution of Countries by Share of Whole Blood Collected from Volunteers



Notes: Figure 1 uses the World Bank income classification method (http://data.worldbank.org/about/country-classifications/country-and-lending-groups)." The 8 percent of the high-income countries (three out of 38) with below 50 percent blood supplied by volunteers are Greece, Saudi Arabia, and Lithuania. The 46 percent (51 out of 110) of lower-income countries with below 50 percent of blood supplied by volunteers include Mexico and many Latin American and African countries.

There is large variation across countries in the percentage of whole blood supply collected from volunteer donors. Thirty-seven percent of all countries collect their entire whole blood supply from volunteers, while another 36 percent collect under 50 percent from volunteers (World Health Organization 2011). Among countries with annual income exceeding $12,616 per capita (using the World Health Organization definition for higher-income countries), Figure 1 shows that over three-quarters rely on 100 percent volunteers, while 46 percent of the lower-income countries rely on other systems for over half of their supply.[1]

Technology and Historical Events Shaping the Market for Blood

The first blood transfusion occurred in the 1600s and hundreds were performed by 1900, although most recipients did not survive. Three breakthroughs radically increased the likelihood of survival for recipients. In the 1800s,

[1] The online appendix (available with this article at http://e-jep.org) shows absolute per capita paid and volunteer donations for all 144 countries for which we obtained data. The absolute per capita shares adjust for the fact that higher-income countries have higher per capita donations (as shown in Figure 2 later). The online appendix shows that the gap in absolute per capita donations from paid donors between lower- and higher-income countries is not as wide as suggested in Figure 1, though per capita donations from paid donors remains higher in almost all lower-income countries than higher-income countries.

sterilization gained widespread acceptance, greatly reducing infections caused when blood passed through tubes from donors to recipients. Karl Landsteiner's 1900 discovery of blood types (O, A, and B) mitigated adverse effects from transfusing mismatched blood types. Last, mechanical devices were developed to control the blood flow and pressure entering recipients. By 1914 almost all recipients survived transfusions. With these quality improvements (the dramatically higher survival rates), demand for blood increased significantly and the quantity supplied rose in response, initially on a small scale due to an inability to store blood, resulting in "marriage market" set-ups in which each recipient needed a personal donor (1914–1937), then on a grander scale with the ability to store blood, resulting in today's impersonalized, diffuse markets.[2]

1914–1937: The Blood-on-the-Hoof Era

From 1914 to 1937, transfusions required blood to flow directly from donors to recipients because blood storage was not yet possible. Suppliers had to be present during transfusions and were thus referred to as "blood-on-the-hoof" (Starr 1998). Although requiring suppliers to be present is implausibly inefficient for large-scale demand, blood-on-the-hoof nonetheless had motivational advantages over current practices; being together, donors saw their donations being used and met recipients, thus eliminating social distance that exists today between anonymous suppliers and unknown recipients.

The most critical challenge during the blood-on-the-hoof era was to find people who could be available to donate when needed. To find these potential matches, hospitals and doctors built lists of people who were pre-screened for health and blood type and were readily available. To be readily available, donor-on-demand lists often included phone numbers for easy contact, and because having a phone in the early 1900s meant being relatively wealthy, donors-on-demand were likely to have relatively good health. The lower costs of pre-screening, greater availability when needed, and good health made donor-on-demand groups the main supply source for hospitals, doctors, and Red Cross agencies.

From the outset, there were both volunteer and paid blood donors. In London, Dr. Percy Oliver established one of the first volunteer groups with 20 donors in 1922 that grew to almost 900 by 1926. In New York, family and friends of patients were encouraged to donate, and individuals could earn $35 to $50 per donation. Given average annual income of around $1,200 (Whitley 1999), the *New York Times* (February 11, 1923) labeled donating blood the "1,001st Way to Make a Living;" the donation price attracted people whose benefit from donating, from altruism and compensation, exceeded their costs (time, discomfort, and health risks).

[2] Starr (1998) and Hayes (2006) provide engaging historical perspectives on blood donations. Many of the historical developments described in this essay can be found in greater depth in these books and the other specific references in this essay.

Foreshadowing future policy debates, concerns with transmitting viruses and bacteria existed from the outset. The New York Blood Transfusion Betterment Association provided safety guidelines to all donors: for example, recommending five weeks between donations, regular health checks, and deferrals for anemia. However, little enforcement, regulation, or oversight existed for another half century, indicating that safety concerns were not a dominant factor in the early years of blood collection.

Collecting blood could also be very profitable to those doing transfusions. Transfusions during this era required skilled doctors who could consequently earn a generous wage (as much as $500) for a single transfusion when a doctor's average annual wage was around $3,380 (Paper-Dragon.com undated). Doctors sometimes did whatever was necessary to find blood; some walked the streets offering cash. Thus, both altruistic and economic (or financial) motives were present from the outset in the medical community and among donors.

1937–World War II: Economies of Scale and Impersonalized Diffuse Markets

In the mid to late 1930s, scientific developments occurred allowing for economies of scale in the blood supply. First, researchers developed the use of sodium citrate as an anti-coagulant (to prevent clotting) to store blood cost-effectively for up to two weeks (now six weeks). Blood banks opened immediately to collect and store blood; storage ended the blood-on-the-hoof (marriage market) era and ushered in the modern impersonalized supplier-recipient relationship. The benefits of blood banking—including the lower cost of moving blood rather than people, stockpiling capability, ability to collect supplies before running out, and today, testing and treating donated blood—vastly outweighed any potential negatives from the movement to an impersonalized market.

At the same time that storage became feasible, scientists learned how to separate whole blood into red cells, platelets, and plasma, and how to fractionate plasma into components (Giangrande 2010). Plasma components could then be a) stored cost-effectively for years, b) transported more safely and cost-effectively than whole blood, and c) combined from multiple suppliers into single packets and later distributed to multiple patients. The economies of scale were enormous.

World War II generated a surge in demand for blood products. Plasma and particularly the protein albumin (that assists in blood flow regulation) were vital to treat shock victims. During the war, more than 13 million units of whole blood were drawn by the American Red Cross alone (Hess and Thomas 2003). Such quantities necessitated more cost-effective collection, storage, and transportation; the first large-scale warehousing of plasma emerged, combining plasma from many individuals into single packets. To motivate donors, and with no price mechanism to adjust for excess demand, collection agencies and governments often linked donations to patriotism to increase the shadow value to donors. The initial US campaign to raise blood, "Plasma for Britain," evolved into nationalistic donation campaigns to support US troops.

In the Aftermath of World War II

The volunteer-nationalism association in blood donation activities likely spilled over to post–World War II norms in many countries, including England, France, Poland, Switzerland, and the United States, that consequently relied primarily on volunteer blood supply immediately after the war. Countries where the link between blood supply and patriotism was weaker or nonexistent—for example, Japan, China, and Russia—primarily relied on paid blood supply immediately after the war.

The grim practicalities of World War II likely also affected attitudes towards blood supply safety. During the war, blood quality concerns like the risks of infections from transfusions were overshadowed by the need to increase the quantity of blood supplied, because the benefits of the high odds of surviving shock with a transfusion greatly outweighed the risks of contracting a blood-related infection, which at that time were mostly not life-threatening. Thus, combining donations from many individuals was considered acceptable, even though it meant viruses and bacteria like hepatitis B and syphilis could be transmitted from a single donor to as many as 60 patients.

Following the war, volunteer and paid supply coexisted across (and sometimes within) countries, with the balance often shifting over time. For instance, in the 20 years after World War II, Japan transitioned from paid to volunteer blood supply, while the United States went from almost entirely volunteer supply to having approximately 20 percent of blood products, primarily plasma, collected from paid supply.

1950s–Present: Demand Growth, Safety, Volunteerism

Since the 1950s, demand for whole blood and plasma products has increased dramatically and continues to increase today due to new medical procedures and aging populations. New procedures—including heart surgery, organ transplants, advanced natal care, and many cancer treatments—require increasingly large amounts of blood. The whole blood supply in the United States has increased from 4 million to 16 million units (450 milliliters per unit) from 1950 to 2006 and worldwide supply exceeded 92 million units in 2011 (World Health Organization 2011). Demand for plasma products has also increased rapidly. For instance, the annual supply of albumin—often used in the treatment of shock and severe burns—increased 20-fold from 30,000 pints at the end of the war to over 600,000 pints by 1990 (Peters 1996) and intravenous immunoglobulin—often used to fight infections—has increased more than five-fold from 15 to 80 tons from 1990 to 2006 (Flood et al. 2006).

While whole blood generally remains within country borders due to its relatively high transportation costs and six-week shelf life, plasma components have been actively traded internationally since World War II. Japan developed the first major for-profit plasma collection organization from paid donors, annually shipping up to $1.5 billion of plasma worldwide by the early 1960s. Other countries, including at least seven in Latin America and South Africa, were also exporting plasma from paid donors in the 1960s.

In the late 1960s, the newly discovered plasmapheresis process radically changed the amount of plasma donors could supply. The plasmapheresis process extracts whole blood, uses centrifuge or filtration to extract the plasma, and returns the red cells and platelets to the donor. Because the red cells are returned, donors have minimal risk of developing anemia and can thus donate 650 milliliters of plasma once or twice per week (compared to giving 250 milliliters of plasma once every 8 to 12 weeks from a whole blood donation).[3] A plasmapheresis donor could thus donate 20 (or more) times more plasma than a whole blood donor. The disadvantage to a plasmapheresis donation is that the entire procedure takes nearly two hours (including 45 minutes for the draw and returning red cells) whereas whole blood takes about one hour (including 10 minutes for the draw).

The dramatic increase in the quantity of plasma that could be collected from the same number of donors using plasmapheresis, along with a willingness to pay donors, led the United States to become the dominant worldwide supplier of plasma in the 1970s. In contrast to most high-income countries relying on 100 percent volunteer plasma supply, by 2004, 81 percent of US plasma supply was collected from paid donors (Flood et al. 2006). In 2004, the United States collected almost 70 percent of the world's plasma, with 40 percent eventually used in North America, 32 percent used in Europe, and 19 percent used in Asia (Flood et al. 2006).

To meet the growing demand for blood products, cost-effective methods to collect and store adequate blood supply had to be weighed against the safety of donors and recipients. Prior to the AIDS crisis of the 1980s, blood collection agencies tolerated more safety risks than today for a variety of reasons, including the nonexistence, unreliability, or high cost of testing for most infections, the unknown extent of health concerns and deaths from infected blood products, the lack of treatment for viruses and bacteria in blood products, and the lack of substantive government regulation and oversight of blood collection. Further, unawareness of the risks associated with subpopulations, combined with protecting the privacy of the volunteer donor, made many collection agencies unwilling to ask personal questions that might have helped screen out higher-risk donors.[4] By the late 1960s, however, there was increasing awareness of infections spreading through blood products. For example, reports at that time estimated 17,000 cases of hepatitis per annum in the United States from blood donations, resulting in estimates of deaths ranging from 850 (Starr 1998) to 3,500 (Comptroller General 1976). Even with this awareness, a low-cost procedure to eradicate hepatitis B in plasma developed in the late 1960s was not commonly used until the early 1980s.

[3] The Council of Europe guidelines suggest one plasma donation per week is safe, while guidelines from the US Food and Drug Administration indicate twice per week (Williams 2013).
[4] One can only wonder, had whole blood supply come from paid suppliers rather than volunteer donors whose motives were perceived as beyond reproach, whether protecting the privacy of donors would have outweighed the benefits of screening donors.

The arrival of AIDS in the early 1980s ushered in an era of aggressive screening and intolerance to risks concerning viruses spread through blood donations. When HIV was first detected, the initial assessment of the risks was underestimated, thus leading agencies to decide that destroying the vast stockpiles of plasma was too costly. With no reliable blood test for HIV at that time, a reluctance to screen donors regarding sensitive issues like sexual activity, and unwillingness to destroy stockpiles, AIDS spread rapidly. Over 14,000 people are estimated to have died of AIDS contracted from blood transfusions and 50 to 80 percent of hemophiliacs were infected by 1985 (Donegan 2003). Even with the availability of an effective treatment for preventing the spread of HIV in stockpiled plasma, the Canadian Red Cross implemented a seven month "transition" period before they required all plasma-based clotting factor to be treated, allowing the agency to distribute over eleven million units of untreated material (Starr 1998).

The underestimation of the HIV-related health risks was the tipping point that has made the safety of the blood supply the predominant concern for all blood collection related policies today. It led to aggressive donor screening and testing all donations, erring today if anything on the side of extreme caution. An example of the caution directly stemming from the AIDS era is the current restrictions with respect to CJD (Creutzfeldt–Jakob Disease or "mad cow disease"). The American Red Cross (2013) permanently defers potential blood donors if they spent more than five years in Europe since 1980 or three months in the UK between 1980–1996. Yet, the US Centers for Disease Control considers the risk of CJD infection almost eliminated and only three cases worldwide were ever traced to blood products (Brown et al. 2012).

The shifting views towards blood supply safety took place against the backdrop of an on-going debate in the 1960s and 1970s on paying donors for blood, which focused on blood supply safety and ethical considerations for donors. Titmuss's 1971 book, *The Gift Relationship*, most prominently articulated the concerns. Titmuss argued that blood supply safety would be compromised by paying donors because it would attract higher-risk donors. He further believed that paying for blood donations would reduce donations because volunteers donating for altruistic reasons would be less willing to donate if paid.[5] *The Gift Relationship* profoundly influenced policymakers (Healy 1999), and by 1975 the World Health Organization (2009) issued policy guidelines for countries to have 100 percent non-remunerated volunteer donations. The guidelines stand to this day and have been adopted by many blood collection agencies. However, Lacetera, Macis, and Slonim (2013) note that no empirical evidence using reputable data-gathering and econometric cause-and-effect analysis has ever tested Titmuss's assertion using payments for donating actual blood. Such a test would be very difficult to

[5] Healy (1999) argues that Titmuss was using the blood context to raise broader concerns about markets, bolstering the argument with quotations from Titmuss (1971, p. 531) like this one, "[A]ltruism is morally better for society than the market. Markets are both inefficient and morally bankrupt. If blood remains a gift, then the system will stay efficient and the bonds of community will remain strong."

carry out, because so many blood collection agency guidelines forbid paying for donations of whole blood.

The volunteer system has performed well in most high-income countries, providing higher per capita donations than in poorer countries relying on nonvolunteer supply to meet demand for whole blood (as evidenced below). However, it is impossible to say how well the volunteer system has performed in an absolute perspective; for example, it is possible that if the blood supply was to increase via a market mechanism that priced blood to its marginal value, then healthcare providers would find innovative uses for it such as the recent trials on the use of plasma derivatives to treat Alzheimer's (Jeffrey 2013). In other words, volunteer supply may meet current demand because the health industry is not aggressively pursuing research and development that might lead to greater demand for blood that they recognize the volunteer system cannot supply.

The volunteer system, however, has not done well in meeting plasma demand. The United States is the only country that is totally self-sufficient in all blood and plasma products (Flood et al. 2006) and has accomplished this using a mostly for-profit plasma industry. Most other countries have to import at least some plasma products, with the single biggest importers being Germany, Austria, and Spain (Ayers 2013). Many countries remain unwilling to pay for plasma donations due to concerns regarding safety and ethics, which may explain the dramatically different usage rates for plasma products. For instance, in 2006 the US health care system used 105 grams per/1,000 people of the dominant plasma product (immunoglobulin) which was more than 250 percent of the rate in Italy, the United Kingdom, Germany, the Netherlands, and Japan (Flood et al. 2006), suggesting that the noncompensated plasma collection system might be limiting potential usage in many countries.

The major ethical consideration with paying for blood donations has been the potential exploitation of donors. As Roth (2007) discussed in this journal, certain transactions involving money can be perceived as repugnant, and these perceptions put real constraints on market transactions. While the World Health Organization guidelines are not legally binding, they have greatly reduced the option of offering economic rewards for blood donations in most high-income countries today. With no competitive market price for obtaining whole blood in these countries, supply of whole blood relies almost entirely on altruistic donations.

Repugnance alone cannot explain, however, why the US whole blood supply is almost entirely from volunteer (noncompensated) donors, while 81 percent of plasma is supplied by paid donors. This distinction did not always exist. Until 1978, paid and volunteer donations coexisted for both whole blood and plasma. In 1978, the Food and Drug Administration ruled that blood products had to be labeled as "paid" or "volunteer" (FDA Compliance Manual undated). Paid whole blood disappeared almost immediately, yet paid plasma continued to coexist with volunteer donations (Starr 1998). Objectively, repugnance arguments should apply equally to whole blood and plasma donations; both involve renewable body parts (in contrast for instance to donation of organs like kidneys), pose minimal health risks to donors, and both are open to exploitation concerns. Starr (1998) provides an economic explanation that has little to do with moral arguments. He argues that hospitals believed in 1978 that whole blood from paid donors was more likely to have the hepatitis B virus and thus believed paid supply was inferior to volunteer supply. In contrast, hospitals and drug companies obtaining plasma made no such distinction between paid and volunteer plasma, because plasma donations were already being screened for hepatitis B. Further, because volunteer donors were able to supply sufficient quantities to meet demand for whole blood most of the time, the need for paid whole blood was low, whereas volunteer supply of plasma frequently could not meet demand. Today, despite reliable tests and treatments of whole blood, paying for whole blood remains almost nonexistent in the United States, while paying for plasma is the norm.

Current Conditions: Pricing, Supply, Safety, and Imbalances in Supply and Demand

Blood Prices

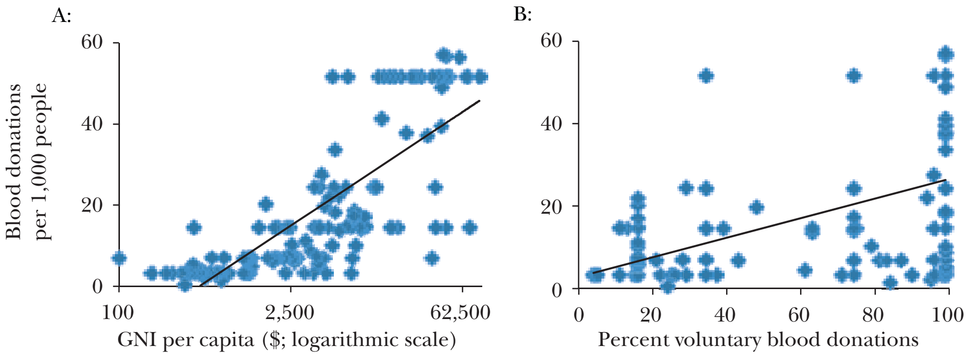
The price paid to agencies that collect whole blood is similar in the United States and Europe. Average prices in 2010 ranged from $154 to $211 per unit (Toner et al. 2011; Shander et al. 2010; Dreaper 2010) to cover operating costs of collecting and storing the blood that include staff, facilities, equipment, and testing. Prices are contracted and hence do not typically change with short-term fluctuations in supply or demand; only 12 percent of US hospitals report prices increasing during periods of shortage. US patients on average pay hospitals $334 per unit for whole blood (Toner et al. 2011).

The United States, Germany, and Austria collect a substantial share of plasma from paid donors, with the United States collecting 70 percent of the world supply (Flood et al. 2006). US plasma donors are typically paid $30 to $60 per donation depending on donation frequency (Blood Plasma Donation Tips 2013). Plasma is then aggregated and fractionated into its component parts and sold to hospitals and drug companies, who pay on average $61 per unit for plasma and $534 per unit for platelets (Toner et al. 2011). Plasma prices fluctuate with supply and demand factors; for example, with increased regulations increasing the costs to fractionators in the mid-1990s, the price of plasma products jumped more than 20 percent from 1996 to 1998 (Flood et al. 2006).

Quantity Supplied and Safety

Many surveys in wealthy countries suggest that the major motivations for people voluntarily donating blood are helping the community, friends, and relatives (Bednall and Bove 2011). These motivations appear sufficiently strong to generate greater per capita quantities and similar safety to other collection systems. Figure 2 shows the relationships between blood donations, income, and volunteerism, based on country-specific data and national blood agency data from approximately 7,000 blood collection agencies covering 144 countries (World Health Organization

Figure 2
Country-Level Blood Donations by Income and Voluntary Status



Notes: Data are for 2008 or the closest possible year with comprehensive and publicly available data. We have four fewer observations in Figure 2 than Figure 1, that is, 144 countries instead of 148; Figure 2 uses GNI per capita data based on the World Bank Atlas method (http://data.worldbank.org/indicator/NY.GNP.PCAP.CD) whereas Figure 1 uses the World Bank income classification method (http://data.worldbank.org/about/country-classifications/country-and-lending-groups). The World Health Organization separates Quebec from the rest of Canada, because it runs its own blood collection operations. We did not include Quebec in these figures; it has higher per capita donations (66 per 1,000 people) than any country listed here.

2011).[6] Data for Figure 2 (as well as Figure 3 and Table 1) are for 2008, or the closest possible year with comprehensive and publicly available data.

The left panel of Figure 2 shows that per capita supply rises proportionally with log per capita gross national income (GNI). The three countries with the highest per capita donations, in descending order, are Australia, the United States, and Denmark. The 10 high-income countries with the lowest per capita donations are, in ascending order, Singapore, Oman, the United Arab Emirates, Qatar, Kuwait, Saudi Arabia, Japan, Poland, and Lithuania. The right panel shows a positive relationship between per capita supply and the percent of voluntary donations. In Panel B, Greece is the outlier country with below 50 percent donations from volunteers with just over 50 donations per 1,000. Table 1 presents some illustrative ordinary least squares regressions using these data. The first column presents ordinary least squares regression estimates of the correlation of country per capita supply with (the log of) gross national income. The estimates indicate that a country's income is significantly positively correlated with per capita donations; 1 percent higher gross national income is associated with 8.8 extra donations per 1,000 people. The second column

[6] A comprehensive list of references for the data can be found in the online appendix. The appendix lists the sources and years reported and where nonstandard procedures were used to obtain country-specific statistics. Figure 1 uses data from 148 countries using the WHO's discrete income classifications whereas in Figures 2 and 3 we include 144 countries using a continuous measure of gross national income using the World Bank classification of country income (that is, the World Bank Classification Method).

Table 1
Blood Donations and Safety, by Income and Volunteerism

	Donations /1,000			Percent TTI/donation		
	(1)	(2)	(3)	(4)	(5)	(6)
log(GNI)	8.83***		7.97***	−3.43***		−3.42***
	(0.65)		(0.63)	(0.53)		(0.53)
Percent volunteer		0.24***	0.12***		0.02	0.00
		(0.03)	(0.02)		(0.03)	(0.02)
R^2	0.62	0.22	0.67	0.40	0.01	0.40
N (countries)	144	144	144	45	45	45

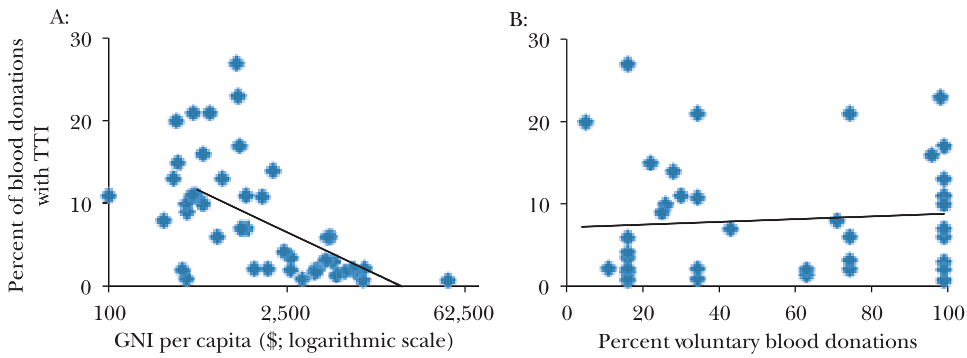
Note: Percent TTI/donation is a measure of how much donated blood is discarded due to detection of a transfusion transmissible infection (TTI). Data are for 2008 or the closest possible year with comprehensive and publicly available data. Calculations use robust standard errors.
*** $p < .01$.

shows a significant positive correlation between the percentage of blood collected from volunteer donations and per capita donations.[7] The third column shows that even after controlling for income, countries with higher proportions of volunteer donations supply more blood; a one percentage point higher proportion of blood collected from volunteer donations is associated with 0.12 additional donations per 1,000 people.

We can use the same database to look at blood safety. To examine safety, we use a common measure of whether donated blood is discarded due to detection of a transfusion transmissible infection. Figure 3 illustrates the results. For countries with public information on transfusion transmissible infections, the left panel of Figure 3 shows a strong negative correlation between a country's (log) gross national income and the percent of blood in which a transfusion transmissible infection was detected. The five countries with 20 percent or higher transfusion transmissible infections, in descending order, are Mauritania, Senegal, Mali, Central African Republic, and Niger. Again, Table 1 offers summary regressions. Column 4 in Table 1 shows the negative correlation between log of gross national income and transfusion transmissible infections (TTIs) is significant; a 1 percent higher gross national income is correlated with 3.3 percentage points fewer transfusion transmissible infections per donation. The right panel of Figure 3 shows little relationship between the percentage of blood collected from volunteer donations and transfusion transmissible infections. Columns 5 and 6 confirm an insignificant relationship

[7] The actual percent of blood collected from volunteer donations is reported in 67 countries. The remaining 77 countries only report the percent collected from volunteers within the categories specified by the World Health Organization: < 25 percent, 25–49.9 percent, 50–89.9 percent, 90–98.9 percent, and 99–100 percent. Alternative specifications in which we use the categories as dummy variables or we limit the analyses to only countries that report actual percentages provide qualitatively similar results.

Figure 3
Country-Level Presence of Transfusion Transmissible Infections (TTI), by Income and Voluntary Status



Note: This figure includes data for 45 countries. Data are for 2008 or the closest possible year with comprehensive and publicly available data.

between volunteer donations and transfusion transmissible infections, and log of gross national income remains significant after controlling for the percentage share collected from volunteers. Thus, while volunteer and nonvolunteer donor characteristics may differ, on the critical issue of blood safety, we find no evidence that countries with higher percentages of volunteer donors provide safer blood.

However, interpretation of safety based on percent of donations with transfusion transmissible infections should be done with caution. For instance, countries with a higher proportion of volunteers may test a higher proportion of donated blood or test for a broader range of infections and viruses. The safety results should also be interpreted with caution since they rely on only 45 countries, and many high-income countries, including the United States, do not report these statistics publicly.[8]

Imbalances in Supply and Demand for Blood

With no market price for whole blood donations and with limited storage length for whole blood, coordinating demand and volunteer supply has been subject to episodes of both excess supply and shortages. Supply spikes often occur after disasters, due to suppliers' altruistic responses and inadequate market signals

[8] Donation decisions may be influenced by perceptions of the effectiveness, safety, and fairness of the system that may or may not match the objective data. To assess whether relationships based on perceptions differed from those based on objective data, we ran a survey using Amazon Mechanical Turk (https://www.mturk.com) to explore attitudes towards blood donations across 78 high- and low-income countries. The survey and overview of the results are available with the online appendix. The results show that, similar to the objective data, controlling for a country's income, respondents from volunteer collection countries are more likely to believe their country is collecting blood efficiently and distributing it fairly ($p < .05$). However, respondent's perceptions of the safety to either the donor or recipient do not vary with the percent of blood collected from volunteers ($p > .85$).

that would have revealed little or no shift in demand. Spikes in donations, combined with six week shelf life for whole blood, along with technical constraints and collection agency policies, have led to destroying blood supply after national disasters. Starr (2002) documents that over 570,000 additional units of blood were collected by the Red Cross immediately after the terrorist attacks of 9/11, with an estimated 100,000 to 300,000 units eventually discarded (plus the time and equipment wasted collecting these units), for an estimated minimum cost of $21–63 million (using the $211 unit cost reported above). Well-publicized images of lines outside blood donor centers immediately after 9/11 likely exacerbated the problem by signaling that donating was the normatively appropriate behavioral response (Cialdini et al. 1993), despite virtually no change in demand for blood.

Of course, one option for collection agencies to address excess supply is to separate the red cells, platelets, and plasma, with the idea that at least the plasma can be stored, although the red cells would still be discarded after six weeks. This option is limited by centrifuge and filtration capacity constraints. When a disaster response is local, supply can be shipped to other locations having centrifuge and filtration capacity. However, when disaster responses are spread over a larger area or are national in scope, there is nowhere to send the excess supply. Yet another option is for collection agencies to turn away volunteer donors, though this policy has not usually been followed for fear that volunteers may take this as a signal that they should not bother donating in the future.

It is also common for volunteer supply to fall below demand, especially during the winter (when suppliers are less able to donate due to higher rates of colds and the flu) and holiday periods (when people travel). The first response to these shortages from blood collection agencies is to employ higher marginal cost strategies to obtain supply, like running additional mass media advertising appeals and increasing direct communications. If such steps prove inadequate, then hospitals which are unable to receive enough supply must prioritize their usage and postpone transfusions and elective surgeries. Toner et al. (2011) report that 58 percent of US hospitals surveyed have postponed transfusions and 46 percent have postponed surgeries, while 14 and 13 percent have cancelled transfusions and surgeries, respectively.

How Economists Can Improve the Market for Blood

The preceding discussion suggests that the whole blood donation market in wealthy countries today is characterized by volunteer supply motivated by altruism. In the absence of a price mechanism to coordinate supply and demand, there is no a priori reason to believe supply and demand will be in balance. Economists can provide solutions to both increase supply and improve the balance.

Recent research suggests several avenues to increase supply. Lacetera and Macis (2010) report higher donations from symbolic rewards (medals) and social recognition (newspaper recognition) among all donors in an Italian town. Goette, Stutzer, and Zehnder (2011), examining 1,838 students, find that requiring people to say yes or no to a donation invitation, rather than offering an option to decide later, increases blood supply. Garbarino, Slonim, and Wang (2013) show that, among 6,000 Australian blood donors who had not donated for at least 28 months, an unconditional gift (a Blood Service pen) increases donations, consistent with preferences for reciprocity. Goette and Stutzer (2011), Iajya, Macis, Lacetera, and Slonim (forthcoming), and Lacetera, Macis, and Slonim (2012; 2013; forthcoming), examining approximately 12,000 Swiss Red Cross donors, 25,000 Argentine nondonors, and 100,000 American Red Cross donors, respectively, provide robust field evidence that offering small gifts like lottery tickets, t-shirts, and gift cards to anyone who presents to donate blood increases supply without affecting future donations. These studies also show that blood safety (using donor deferrals to proxy for level of safety) was unaffected by the incentives, indicating that the Titmuss (1971) concern, that payments would attract lower quality supply, does not seem to apply for small gifts. Lacetera, Macis, and Slonim (2013) stress that the current practice in these studies is that blood donors receive incentives for presenting themselves to donate blood, rather than conditional on actually donating. This distinction could matter for safety because making rewards contingent on successfully donating can provide an incentive to falsify information, whereas providing rewards for showing up removes the incentive to falsify information to receive the rewards.

However, offering substantial or widespread material incentives to increase supply remains an unlikely option in many countries whose institutions retain policies promoting unremunerated donations. Although many countries seem unwilling to pay blood donors themselves, current policies indicate the willingness of high-income countries to pay for the efforts of those who collect, store, and use the blood for treatments.

Non-price Signals and a Blood Registry

In the current volunteer blood donation context, the absence of a market price means that donors may donate when blood is not needed and not donate when it is needed. Because blood donors are a diffuse and independent group who make decisions with limited information on needs (given that there is no price to indicate higher or lower need), they cannot easily coordinate their actions. A central clearinghouse system can provide this coordination in the absence of a price signal, and economic market design principles can be used to fashion these operations.

To address shortages, most blood services use a combination of strategies, including media and telemarketing. These strategies can increase supply, but do not solve coordination problems. Introducing a donor registry to be used during times of shortages can better coordinate donor actions and be more cost-effective to the collection agency. The registry collects information from blood donors on their preferences about when to donate; for example, some people are more willing to donate if there is a particular need for blood of their personal blood type or in their own community. The critical assumptions for the registry's success are that its members 1) are willing to donate when there is a specific need, but unlikely otherwise, and 2) know they will only be invited to donate when there is demand matching their preferences. Thus, the registry informs marginal donors when the shadow value of their donation rises.[9]

To test the efficacy of a donor registry, we ran a field experiment with the Australian Red Cross Blood Service. Australia is a high-income country with a well-established 100 percent volunteer blood supply and has the highest per capita donations of any country. The Blood Service, with approximately 4,000 paid employees that include donor center operational staff like nurses and lab technicians, as well as telemarketers, is responsible for Australia's entire blood collection and distribution. This national monopoly is ideal for studying the market for blood because the Blood Service manages *all* blood donations, communications, appointments, bloodstocks, and demand information for the entire market.

The study examined 13,200 "long-lapsed" donors, defined as past donors who have not donated for at least two years. Most long-lapsed donors are eligible to donate but unlikely to return on their own, having an annual reactivation rate under 1 percent. Because long-lapsed donors stop receiving targeted Blood Service marketing, our study was the only direct communication with them on blood donation. The subjects were randomly chosen from the universe of 44,222 long-lapsed donors between the ages of 23 and 60 and who last donated between 27 and 43 months prior to our first attempted communication with them. We have substantial demographic information and donation behavior including donation history and eligibility. We randomly divided the 13,200 donors into 9,000 registry and 4,200 control subjects each having an equal number of men and women and equal distribution across three past donation categories: one past donation, two or three past donations, and four or more past donations.

The Blood Service's National Call Center called 9,000 of the subjects to invite them to join the registry, telling them that they would only be contacted "when the community has a critical need for blood, for example a need for your own blood type or a need in your local area," and that they would probably only be contacted once or twice a year. The Call Center reached and invited 2,588 of these potential donors who were still eligible to donate; almost all of those not invited were subjects not answering the attempted phone calls. Among the 2,588 past donors contacted, 1,914 (74 percent) agreed to join the registry.[10]

[9] Two closely related mechanisms the registry may also operate through are: 1) the foot-in-the-door technique (Burger 1999) which requires small involvement initially (joining the registry) then the larger sacrifice (donation) later; and 2) as a precommitment device (de Hooge, Breugelmans, and Zeelenberg 2008), with joining the registry being an implicit promise to donate later, thus raising the (psychological) cost to say no later. Distinguishing between these paths and the shadow value are empirically difficult, but we would expect the shadow value path would not fade over the longer term, a finding which is consistent with our results.

[10] Subjects invited to join the registry were divided into three subtreatments to see if the invitation to join the registry crowded out making a donation immediately. Comparing immediate donations among donors invited to both join the registry and donate immediately with a control group that was only invited to donate immediately, we found no crowding. The subtreatments let us further test whether being asked to donate affected joining the registry. Comparing the decision to join the registry among donors invited to join the registry and donate immediately with a group invited to join the registry only, we found no difference.

Three to five months later, the Call Center called all registry members plus 2,400 control subjects (who the center had not attempted to contact previously) to invite them to donate blood during the winter blood shortage period. The script explained that "so many of our regular donors are unable to give due to having colds or the flu around this time, but Australia continues to need over 26,000 donations every week just to meet the ongoing needs of patients." We found that 9.0 percent of registry members but only 5.5 percent of control donors presented themselves to donate within four weeks. (Most attendances are within four weeks of calling, but the results are qualitatively similar using longer time-frames.) This greater registry response was not because registry members were more likely to be contacted than control subjects; examining only eligible subjects who were reached by phone, 32.1 percent of registry members but only 20.8 percent of control subjects presented to donate, a 54 percent (32.1 percent/20.8 percent) relatively higher donation rate.[11] The registry effects reported here are statistically significant in probit regression analysis controlling for many factors including past donations, time since last donation, whether deferred at their last donation, demographics, and experimental design features like including dummy variables for specific Call Center agents and for call dates (detailed regression results are available in the online appendix). Moreover, when called one year later to donate during the following winter shortage, the results were qualitatively similar; among all subjects we attempted to contact in July/August 2013, 9.7 percent of registry members but only 5.9 percent of control subjects presented to donate within four weeks of receiving the calls. Last, in a separate treatment, when the Call Center invitation more strongly justified the need for donating by explaining "the blood levels are very low and donations of your type <A/O> blood are urgently needed. Many Australians with life-threatening conditions will need blood in the next few weeks to stay alive," the registry effects were even larger.

Thus, our registry design that initially invited long-lapsed donors to join the registry and later called those who joined to make a donation during a critical need produces the same number of donations for fewer calls (or more donations for the same number of calls). Moreover, the registry increases donations during shortages and increases donor welfare by inviting donations when the shadow value of donating is higher to the supplier. Future registry formats might improve on our design by exploiting potential heterogeneous donation preferences such as preferring to donate at specific times of the year, for local needs vs. national needs, or for specific purposes (like accident victims or military personnel).

[11] The significance cannot be explained by the 26 percent attrition among subjects who did not join the registry. For instance, even if we remove 26 percent of the control subjects (who chose not to donate) from our analysis, there is still a significantly higher percent of donations among registry members than among the remaining control subjects.

Discussion

The market for blood—along with many other "products" and services such as organ donations, adoption, surrogacy, and dating services—is constrained by deeply entrenched social norms and ethical and safety concerns. In the case of the supply and demand for blood donations, a combination of these concerns, together with historical events, has led to a reliance on volunteer donations for whole blood. Our evidence of the current state of the whole blood market indicates that countries with higher percentages of volunteer donors are associated with a higher quantity of blood donations, even after controlling for per capita income, but do not provide safer blood.

We expect that limited experiments with incentives to donate blood will continue, and perhaps a few countries will move toward remunerated systems. After all, although the United States is essentially all-volunteer for whole blood, the US plasma market has long been dominated by commercial operations which have made the United States the world's major supplier of plasma products meeting the needs of countries that rely on volunteer plasma. Social norms evolve over time, but the use of explicit prices for whole blood, at least on any widespread basis, will remain infeasible in most high-income countries unless current policies change. Hence, given the social welfare importance of the market for blood and the lack of substitutes, market innovations that improve the balance between supply and demand within the constraints of volunteer systems are needed. Nonprice signals that operate within social and ethical constraints should seek to provide a higher shadow value to motivate marginal blood donors to donate when the quantity of blood supplied falls below the quantity demanded. We described the success of one design that introduces a registry to improve the central clearinghouse functions of blood collection agencies. More innovations in this direction, using economic market design tools, have the potential to improve welfare significantly.

■ *We received many helpful suggestions from seminar participants at the University of New South Wales, the University of Technology, and the editors of this journal. Michael Major provided exceptional research assistance. We greatly appreciate the generous financial support of the Australian Research Council, #1095222, and the Australian Red Cross Blood Service (Blood Service) for their continuous support. We appreciate the support from many Blood Service personnel including Perfecto Diaz, Bianca Folber, David Irving, Anita Smith, Geoff Smith, and Dan Waller who have generously given their time and energy with institutional details, guidance, and service. We would like to acknowledge the Australian Red Cross Blood Service and Australian governments that fully fund the Blood Service for the provision of blood products and services to the Australian community.*

References

Abraham, I., and D. Sun. 2012. "The Cost of Blood Transfusion in Western Europe as Estimated from Six Studies." *Transfusion* 52(9): 1983–88.

American Red Cross. 2013. "Eligibility Criteria by Alphabetical Listing." http://www.redcrossblood.org/donating-blood/eligibility-requirements/eligibility-criteria-alphabetical-listing#arc5. Accessed December 12, 2013.

American Red Cross. Undated. "Blood Facts and Statistics." http://www.redcrossblood.org/learn-about-blood/blood-facts-and-statistics. Accessed December 12, 2013.

Ayers, Christin. 2013. "Who is Making Money from the Bay Area's Blood Donation?" CBS San Francisco, November 1. http://sanfrancisco.cbslocal.com/2013/11/01/who-is-making-money-from-bay-area-blood-donations/.

Bednell, T., and L. Bove. 2011. "Donating Blood: A Meta-analytic Review of Self-reported Motivators and Deterrents." *Transfusion Medicine Reviews* 25(4): 317–34.

Blood Plasma Donation Tips. 2013. "How Much Do You Get Paid to Donate Plasma?" March 3. http://plasmadonating.blogspot.com.au/2013/03/how-much-do-you-get-paid-to-donate.html.

Brown, Paul, Jean-Philippe Brandel, Takeshi Sato, Yosikazu Nakamura, Jan MacKenzie, Robert G. Will, Anna Ladogana, Maurizio Pocchiari, Ellen W. Leschek, and Lawrence B. Schonberger. 2012. "Iatrogenic Creutzfeldt-Jakob disease, final assessment." *Emerging Infectious Diseases* 18(6): 901–907. http://wwwnc.cdc.gov/eid/article/18/6/pdfs/12-0116.pdf.

Burger, J. M. 1999. "The Foot-in-the-Door Compliance Procedure: A Multiple-Process Analysis and Review." *Personality and Social Psychology Review* 3(4): 303–25.

Cialdini, Robert B., Wilhelmina Wosinska, Daniel W. Barrett, Jonathan Butner, and Malgorzata Gornik-Durose. 1999. "Compliance with a Request in Two Cultures: The Differential Influence of Social Proof and Commitment/Consistency on Collectivists and Individualists." *Personality and Social Psychology Bulletin* 25(10): 1242–53.

Comptroller General of the United States. 1976. "Hepatitis from Blood Transfusions: Evaluations of Methods to Reduce the Problem." February 13.

de Hooge, I. E., S. M. Breugelmans, M. Zeelenberg. 2008. "Not So Ugly after All: When Shame Acts as a Commitment Device." *Journal of Personality and Social Psychology* 95(4): 933–43.

Donegan, Elizabeth. 2003. "Transmission of HIV by Blood, Blood Products, Tissue Transplantation, and Artificial Insemination." HIV InSite Knowledge Base Chapter, University of California, San Francisco. http://hivinsite.ucsf.edu/InSite?page=kb-07-02-09#S2X.

Dreaper, Jane. 2010. "Lower Price for Donated Blood Saves NHS Hospitals £50m." BBC, Aug 20. http://www.bbc.co.uk/news/health-11038849.

Eastlund, T. 1998. "Monetary Blood Donation Incentives and the Risk of Transfusion-transmitted Infections." *Transfusion* 38(9): 874–82.

Farrugia, A, J. Penrod, and J. M. Bult. 2010. "Payment, Compensation and Replacement—The Ethics and Motivation of Blood and Plasma Donation." *Vox Sanguinis* 99(3): 202–211.

FDA Compliance Manual. Undated. "Inspections, Compliance, Enforcement, and Criminal Investigations." http://www.fda.gov/ICECI/ComplianceManuals/CompliancePolicyGuidanceManual/ucm122798.htm. Accessed December 13, 2013.

Flood, Philip, Peter Wills, Peter Lawler, Graeme Ryan, and Kevin A. Rickard. 2006. "2006 Review of Australia's Plasma Fractionation Arrangements." Australian Government Department of Health. http://www.donateblood.com.au/files/pdfs/Review%20of%20Australia%27s%20plasma%20fractionation%20arrangements%20Dec06.pdf.

Garbarino, Ellen, Robert Slonim, and Carmen Wang. 2013. "The Multidimensional Effects of a Small Gift: Evidence from a Natural Field Experiment." *Economics Letters* 120(1): 83–86.

Giangrande, Paul L. F. 2010. "The History of Blood Transfusion." *British Journal of Haematology* 110(4): 758–67.

Goette, L., and A. Stutzer. 2011. "Blood Donation and Incentives: Evidence from a Field Experiment." Unpublished paper.

Hayes, Bill. 2006. *Five Quarts: A Personal and Natural History of Blood*. Random House.

Healy, Kieran. 1999. "The Emergence of HIV in the U.S. Blood Supply: Organizations, Obligations, and the Management of Uncertainty." *Theory and Society* 28(4): 529–58.

Healy, Kieran. 2000. "Embedded Altruism: Blood Collection Regimes and the European Union's Donor Population." *American Journal of Sociology* 105(6): 1633–57.

Hess, J., and M. Thomas. 2003. "Blood Use in War and Disaster: Lessons from the Past Century." *Transfusion* 43(11): 1622–33.

Iajya, Victor, Nicola Lacetera, Mario Macis, and Robert Slonim. 2013. "The Effects of Information, Social and Financial Incentives on Voluntary Undirected Blood Donations: Evidence from a Randomized Controlled Trial in Argentina." *Social Science and Medicine* 98 (December): 214–33.

Jeffrey, Susan. 2013. "GAP: IVIG Negative in Alzheimer's, But Some Hints of Benefit." Medscape.com, July 17. http://www.medscape.com/viewarticle/807970.

Lacetera, Nicola, and Mario Macis. 2010. "Social Image Concerns and Prosocial Behavior: Field Evidence from a Nonlinear Incentive Scheme." *Journal of Economic Behavior and Organization* 76(2): 225–37.

Lacetera, Nicola, Mario Macis, and Robert Slonim. 2012. "Will There Be Blood? Incentives and Substitution Effects in Pro-Social Behavior." *American Economic Journal: Economic Policy* 4(1): 186–223.

Lacetera, Nicola, Mario Macis, and Robert Slonim. 2013. "Economic Rewards to Motivate Blood Donation." *Science* 340(6135): 927–28.

Lacetera, Nicola, Mario Macis, and Robert Slonim. Forthcoming. "Rewarding Volunteers: A Field Experiment." *Management Science.*

Paper-Dragon.com. Undated. "Dirty 30s." http://www.paper-dragon.com/1939/priceguide.html. Accessed December 11, 2013.

Peters, Theodore. 1996. *All About Albumin: Biochemistry, Genetics and Medical Applications.* Elsevier: Academic Press.

Roth, Alvin. 2007. "Repugnance as a Constraint on Markets." *Journal of Economic Perspectives* 21(3): 37–58.

Shander, A., A. Hoffmann, S. Ozawa, O. M. Theusinger, H. Gombotz, and D. R. Spahn. 2010. "Activity-based Costs of Blood Transfusions in Surgical Patients at Four Hospitals." *Transfusion* 50(4): 753–65.

Starr, Douglas. 1998. *Blood: An Epic History of Medicine and Commerce.* Alfred A. Knopf.

Starr, Douglas. 2002. "Bad blood: The 9/11 Blood-donation Disaster." *The New Republic,* July 29, 227(5): 13–16.

Stutzer, A., L. Goette, and M. Zehnder. 2011. "Active Decisions and Prosocial Behaviour: A Field Experiment on Blood Donation." *Economic Journal* 121(556): F476–F493.

Titmuss, Richard M. 1971. *The Gift Relationship.* London: Allen and Unwin.

Toner, R. W., L. Pizzi, B. Leas, S. K. Ballas, A. Quigley, and N. I. Goldfarb. 2011. "Costs to Hospitals of Acquiring and Processing Blood in the US." *Applied Health Economics and Health Policy* 9(1): 29–37.

Tracy, Dan. 2010. "Blood is Big Business: Why Does It Cost So Much?" *Orlando Sentinel*, April 5. http://articles.orlandosentinel.com/2010-04-05/news/os-blood-cost-anne-chinoda-20100405_1_blood-is-big-business-community-blood-centers-donors/2.

Whitley, Peggy. 1999. *1920–1929 American Cultural History.* Lone Star College-Kingwood Library. http://kclibrary.lonestar.edu/decade20.html.

Williams, Alan E. 2013. "FDA Considerations Regarding Frequent Plasma Collection Procedures." Presentation at 15th International Haemovigilance Seminar, January 20–22. http://www.ihn-org.com/wp-content/uploads/2012/08/4-WilliamsFreqPlasma-2-21-131.pdf.

World Health Organization. 2009. "WHO Global Consultation: 100% Voluntary Non-remunerated Donation of Blood and Blood components." June. http://www.who.int/bloodsafety/events/consultation_vnrbd/en/.

World Health Organization. 2011. "Global Database on Blood Safety." http://www.who.int/bloodsafety/global_database/GDBS_Summary_Report_2011.pdf. Accessed December 10, 2013.

Journal of Economic Perspectives—Volume 28, Number 2—Spring 2014—Pages 197–212

Retrospectives
The Cyclical Behavior of Labor Productivity and the Emergence of the Labor Hoarding Concept

Jeff E. Biddle

This feature addresses the history of economic terms and ideas. The hope is to deepen the workaday dialogue of economists, while perhaps also casting new light on ongoing questions. If you have suggestions for future topics or authors, please write to Joseph Persky of the University of Illinois at Chicago at jpersky@uic.edu.

Introduction

The concept of "labor hoarding," at least in its modern form, was first fully articulated in the early 1960s by Arthur Okun (1963). By the 1980s, it could be found in undergraduate economics textbooks, where it was presented as a profit-maximizing response by employers to costs of hiring, firing, and training workers, and thus as an explanation of procyclical labor productivity (for example, Dornbusch and Fisher 1981; Hamermesh and Rees 1984; Flanagan, Smith, and Ehrenberg 1984). By the end of the 20th century, the concept of "labor hoarding" had become an accepted part of economists' explanations of the workings of labor markets and of the relationship between labor productivity and economic fluctuations.

The emergence of this modern concept of labor hoarding involved the conjunction of three key elements: a fact, a perceived contradiction, and an explanation. The *fact* was that measured labor productivity (output per worker or per hour worked) was found to be procyclical, rising during expansions and falling during

■ *Jeff E. Biddle is Professor of Economics, Michigan State University, East Lansing, Michigan. His email address is biddle@msu.edu.*

http://dx.doi.org/10.1257/jep.28.2.197 doi=10.1257/jep.28.2.197

contractions. That fact was perceived to be in *contradiction* with a widely held theory: In the basic neoclassical theory of the firm in a competitive industry, short-run fluctuations in demand are met by increases or decreases in the variable labor input, with the fixed capital stock unchanged. Given the assumption of the diminishing marginal productivity of labor, this should lead average labor productivity to move countercyclically. A possible *explanation*, rooted in optimizing behavior on the part of firms, held that costs of hiring, firing, and training employees make it optimal for employers facing a short-run drop in demand to retain more workers than technically necessary to produce current output. This dampens the amplitude of employment fluctuations in response to demand fluctuations, leading average output per worker to fall when demand falls.

Each of these three elements—fact, contradiction, and explanation—has a history of its own, dating back to at least the opening decades of the 20th century. Telling the story of the emergence of the modern labor hoarding concept requires recounting these three histories, histories that involve the work of economists motivated by diverse purposes and often not mainly, if at all, concerned with the questions that the labor hoarding concept was ultimately used to address. As a final twist to the story, the long-standing positive relationship between labor productivity and output in the US economy began to disappear in the late 1980s; and during the Great Recession, labor productivity rose while the economy contracted. In the conclusion, I offer some observations on how some economists have reacted to this shift.

The Discovery of Procyclical Labor Productivity

For the first half of the 20th century, the conventional wisdom among economists was that labor productivity was countercyclical. Three main arguments were offered in support of this belief. First, rising demand for labor during expansions would force employers to hire lower-quality workers, reducing average productivity, while during recessions productivity would rise because the least productive workers would be discharged. Second, workers were fatigued by long hours typically demanded during economic expansions, making them less productive on average. Third, workers were motivated to work harder when they feared the prospect of job loss, as in a recession, and tended to slack off when labor markets were tight and good alternative jobs readily available. Wesley Mitchell, in his influential 1913 book *Business Cycles*, phrased the arguments this way (pp. 477–78):

> (L)ess efficient employes [sic] are the first to be discharged after a crisis. Hence the relatively small working forces of depression are the picked troops of the industrial army. When a revival has grown into full prosperity, on the contrary, employers are constrained to accept any help to be had . . . A deterioration of the average efficiency of the working force inevitably follows. . . . Now overtime labor is especially expensive to employers, not only because

> it often commands extra rates of wages, but because it is tired labor . . . (A)fter a time all hours of every day find the men less alert and less energetic—unable to accomplish as much work per hour as in less busy seasons. . . .
>
> Quite apart from this difficulty of overtime, men cannot be induced to work at so fast a pace when employment is abundant as when it is scarce.

Mitchell's interest in the cyclical behavior of productivity was linked to his desire to construct a theory of the business cycle in which people's economic behavior during contractions sowed the seeds for subsequent expansions, and vice versa. The alleged upward pressure on labor costs during economic expansions as labor efficiency declined, and the corresponding decline in costs during recessions, were potentially important mechanisms in his theory. But Mitchell, although a staunch empiricist, was only able to present anecdotal evidence to support this theory. However, Paul Douglas (1922) presented productivity statistics from a few industries showing labor productivity falling during the World War I expansion and increasing in the subsequent recession, and over the next several decades, other economists would frequently echo Mitchell's reasoning regarding the likely cyclical behavior of labor productivity.

The Bureau of Labor Statistics began compiling industry-level measures of labor productivity in the mid-1920s (Woirol 2006), but by the 1930s the National Bureau of Economic Research (NBER), where Mitchell was Director of Research, had become the center for the study of the accumulating productivity data.[1] However, reliable generalizations about labor productivity over the cycle were difficult to come by, as annual and monthly productivity measures still depended heavily on interpolated data and were considered untrustworthy by NBER researchers (for example, Hultgren 1960, p. xv). As a result, most of the analyses of productivity coming out of the NBER in the 1940s and 1950s dealt with long-term trends. Still, NBER researchers were beginning to see patterns in the data that contradicted the conventional wisdom that labor productivity was countercyclical. Hultgren (1948, p. 182), in his study of the cyclical behavior of the transportation sector, examined rail industry data and concluded: "From 1921 onward, the productivity of labor, defined as traffic units per man-hour, rose in every expansion, fell in every contraction." Hultgren's results led Mitchell to doubt his 1913 hypothesis (Burns 1952). In his posthumously published *What Happens During Business Cycles*, Mitchell (1951, pp. 132–133) acknowledged "two sets of cyclical fluctuations" in labor productivity, "one positively, the other inversely, related to production." The main factor cited as making for a positive relationship was that "modern plants attain their highest technical efficiency when operated steadily at the capacity for which they were designed."

[1] Mills (1933) is one early NBER research project concerned with measuring and analyzing productivity statistics. In 1937, the Bureau received a grant from the Maurice and Laura Falk Foundation to fund a program on the measurement of production and productivity (NBER 1938, p. 29).

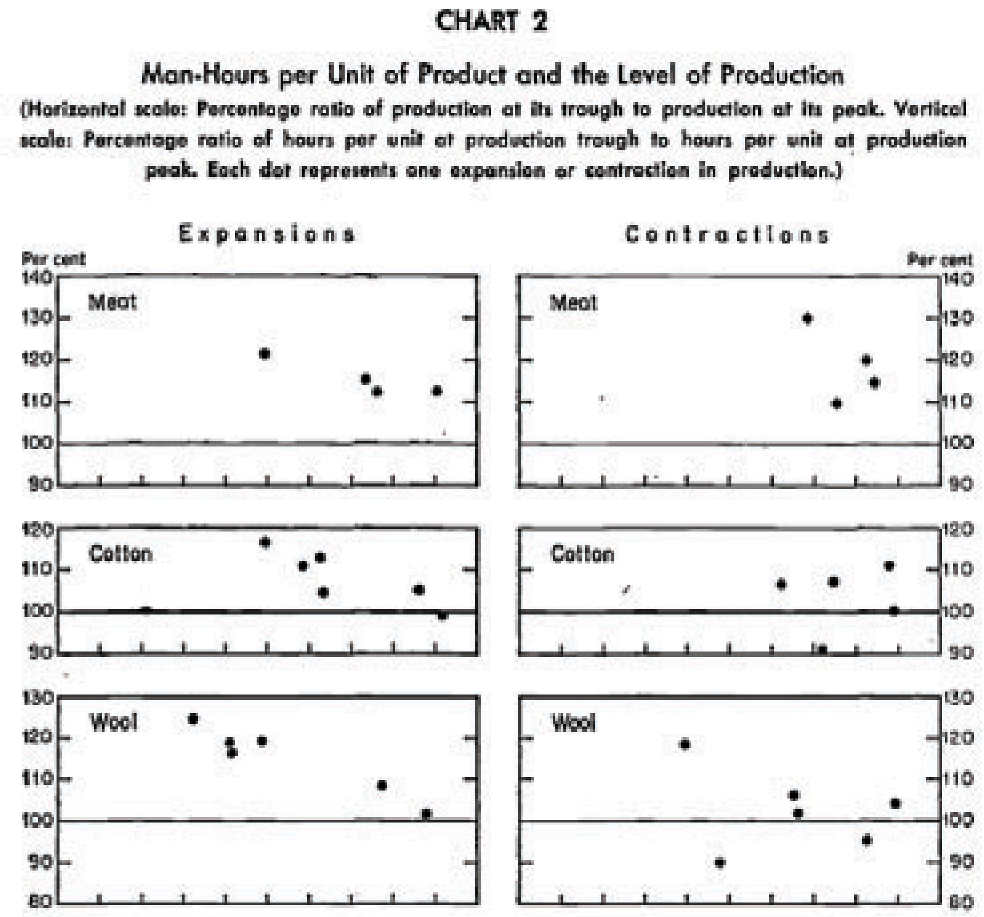
The NBER's official announcement of the reality of procyclical labor productivity came in 1959, when NBER Director of Research Solomon Fabricant (1959, p. 10) declared: "Over the whole period since 1889, productivity fluctuated with the state of business. Year-to-year rises in productivity were greater than the long-term rate when business was generally expanding, and less (or often, falling), when business was generally contracting." Fabricant based his conclusion mainly on further research by Hultgren (1960). Using monthly output data from several industries, Hultgren had identified the peaks and troughs of each industry's "specific cycles" and calculated the change in labor hours per unit of output for each peak-to-trough period (contraction) and each trough-to-peak period (expansion). Pooling data across industries, Hultgren (p. 8) reported: "In one industry or another, at one time or another, we have data on ninety expansions of production and ninety-nine contractions. In eighty-three, or 92 percent, of the ninety expansions, there was a net decline in hours per unit. In seventy, or 71 per cent, of the ninety-nine contractions, there was a net rise in hours per unit. The pooled data suggest a strong tendency toward an inverse relation between hours per unit and total output." Hultgren's data also revealed a strong positive relationship between productivity at the industry level and movements of aggregate output.

Hultgren represented his evidence of procyclical labor productivity graphically with industry-level scatter plots, some of which are shown in Figure 1. Each point on a plot represents the ratio of hours per unit of output at the trough of a cycle phase to hours per unit at the peak of the phase, from cycles covering the years 1921–32. Since Hultgren measured productivity as hours per unit of output rather than the more conventional output per hour, points above the horizontal 100 percent line represent cases in which productivity moved in the same direction as output.

At the time Hultgren's (1960) book was published, Edwin Kuh (1960) had just completed a report for the Joint Economic Committee of Congress that demonstrated the procyclical behavior of labor productivity using data and methods quite different from Hultgren's. Kuh's output series was value product generated by the corporate sector, quarterly, from 1947 to 1957, and with some estimation and interpolation, Kuh constructed a corresponding man-hour series. He then regressed output per man-hour on the level of output and a trend. The coefficient on output was positive and significant: that is, labor productivity rose with output.

Kuh was one of several economists of the time working on macroeconometric models based on a Keynesian theory, a central purpose of which was to provide a more reliable foundation for economic stabilization policies, and a number of these macroeconometricians soon followed Kuh in demonstrating the existence of procyclical labor productivity in both the United States and the United Kingdom, including Soligo (1963), Neild (1963), and Brechling (1965). (These papers circulated for some time prior to publication.) For the most part, these economists did not articulate their motivation in terms of understanding the behavior of productivity over the cycle, but in terms of better modeling the short-run relationship between employment and output, often for the purposes of producing more accurate forecasts. It became common to summarize this relationship using the

Figure 1
Hultgren's Graphical Portrayal of Procyclical Labor Productivity at the Industry Level



Source: Hultgren (1960).
Notes: Each point on a plot represents the ratio of hours per unit of output at the trough of a cycle phase to hours per unit at the peak of the phase, from cycles covering the years 1921–32. Since Hultgren measured productivity as hours per unit of output rather than the more conventional output per hour, points above the horizontal 100 percent line represent cases in which productivity moved in the same direction as output.

estimated short-run elasticity of employment with respect to output (Soligo 1963). These authors understood that a value below 1 for this elasticity implied procyclical labor productivity; but the behavior of labor productivity was not the central focus of their work. Conversely, Hultgren (1960) understood that the patterns he was uncovering in his labor productivity series implied that the cyclical fluctuations of output were greater than the cyclical fluctuations of employment when expressed in percentage terms; for him, it was simply not the more pertinent way to express the finding.

The Labor Hoarding Concept Emerges

By the early 1960s, then, procyclical labor productivity was a well-accepted fact. As Arthur Okun (1963, p. 6) informed an audience at the American Statistical Association meetings: "The record clearly shows that man-hour productivity is depressed by low levels of utilization, and that periods of movement towards full employment yield considerably above average productivity gains." Likewise, Robert Solow (1964, p. 6), in his presidential address to the Econometric Society, explained: "Generally speaking, productivity rises most rapidly when output is recovering toward capacity and falls or rises least rapidly during recessions."

Okun's (1963) presentation to the American Statistical Association represents the earliest articulation of the labor hoarding concept that I have found involving all three of the components identified in the introduction: an acknowledgement of the fact of procyclical labor productivity, an assertion of the contradiction between that fact and basic neoclassical theory, and a possible explanation of the contradiction rooted in an analysis of how costs of hiring, firing, and training workers affect a firm's optimal employment strategy. Regarding this third point, Okun noted (p. 6):

> [T]he positive relationship between output and labor productivity suggests that much of labor input is essentially a fixed cost for fairly substantial periods. Thus high output levels permit the spread of labor overheads, and low production levels raise unit fixed costs of labor. At times we may take too seriously our textbook examples which view labor as a variable factor, with only capital costs as fixed. Even the most casual empiricism points to an overhead component in labor costs.

Okun's (p. 7) reasons why employment "may not be easily variable" included "Transaction costs associated with laying off labor and then, in the future, doing new hiring . . ." and "Acquired skills that existing employees have learned on the job may make them particularly valuable to the firm, so that it pays to stockpile underemployed labor . . ."

Solow's (1964) acknowledgement of the procyclicality of labor productivity was also accompanied by the two other components of the labor hoarding concept. Solow (pp. 5–6) linked the cyclical movements of productivity to earlier demonstrations that real wages moved procyclically (Dunlop 1938; Tarshis 1939), noting that both phenomena contradicted the neoclassical view of labor markets, in which applying more labor to a fixed stock of capital led to the prediction that "output per man must fall—or rise slower than trend—during business cycle upswings and rise—or rise faster than trend—during downswings or decelerations." He described three mechanisms that had been proposed to explain what he called the "perverse behavior of productivity in the short run," explicitly giving one of them the "labor hoarding" label (pp. 7–8, emphasis in the original):

> The labor hoarding mechanism operates on the assumption that important costs are associated with *changes* in the level of employment and with the risk

that trained workers laid off as output falls may not be available as output rises again. It can be shown that a long run cost-minimization policy may require that even if *labor-requirements* per unit of output fall with output, *employment* per unit of output may well increase as recession sets in.[2]

Reconciling Procyclical Labor Productivity and Neoclassical Theory

Solow (1964) labeled procyclical labor productivity "perverse," because it contradicted the basic neoclassical theory of the firm. Solow's perception of the relationship between the cyclical behavior of labor productivity and economic theory, however, was not shared by most of the economists doing research during the 1950s and early 1960s on the cyclical behavior of productivity. Neither Fabricant nor Hultgren had pointed out the contradiction between the NBER's findings regarding labor productivity and the implications of neoclassical theory, but this is not surprising. Fabricant had no particular commitment to neoclassical theory, and in this he was following Wesley Mitchell, for whom the neoclassical tradition represented just one of many potentially useful types of economic theory. Indeed, Mitchell's analysis of the business cycle, which had provided the framework for Hultgren's empirical study of productivity, eclectically combined elements derived from several theoretical traditions.

The macroeconometric model builders who helped establish the procyclicality of labor productivity displayed a range of attitudes regarding the theoretical implications of their results. For example, Neild's (1963) first priority was to produce better forecasts for policymakers. Both "traditional" and "imperfect competition" theories, he argued, assumed that employment would fall proportionately with output; but that was not what happened. "We do not want to prove or disprove any theory," Neild (p. 1) explained, "but to find relationships which work." In a similar vein, Kuh (1965) estimated short-run and long-run elasticities of employment with respect to output. In summarizing his results, Kuh commented (p. 9) that although many of his estimates seemed to contradict neoclassical reasoning, this fact "should be a cause of neither congratulation nor concern since the business cycle has been excluded by choice from that domain of analysis." From this perspective, procyclical labor productivity was not to be regarded as a challenge to or contradiction of neoclassical theory, but as irrelevant to it.

Indeed, many economists in the 1960s took the view that the domain of neoclassical analysis was properly limited to some subset of economic activity, with

[2] Solow (1964, p. 8) also cited what he called the "decreasing cost" mechanism, which "rests on the hypothesis that much labor in modern industry is 'overhead' in character, and can be thought of as a fixed factor in the short run." But he argued that since "many of the characteristics which lead employers to hoard labor are precisely those which give some labor its quality as overhead," he would "use the term labor-hoarding as a catch-phrase to stand for all the frictions involved in meeting transitory variations in output with variations in employment."

alternative theoretical frameworks required to understand phenomena outside that subset.[3] Brechling (1965, p. 188), for instance, began constructing his model of the short-run employment–output relationship by assuming "imperfect competition and administered prices" and commented that he did not believe that cost minimization "or any other motivation" could be assumed over the short periods he was studying.

Solow's (1964) address, however, clearly framed procyclical labor productivity as a challenge to economic theory, explicitly rejecting both the attitude that "finding relationships that work" for forecasting was more important than resolving such challenges and the arguments that sidestepped the apparent challenge with appeals to theoretical pluralism in economics. Solow (pp. 29–30) wrote:

> I know it will be as obvious to you as it is to me that I have not solved the problem of giving a good theoretical explanation of short-run productivity behavior. I hope it will be as obvious to you as it is to me that this is a problem worth solving. Its importance goes far beyond the desirability of being able to predict how output per man-hour will change from this quarter to the next. . . . What I am looking for is a way to unify the economics we teach our students and the economics we use when we advise governments and analyze passing economic events, and do it in a way amenable to econometric treatment. This patching up of theory to explain experimental or statistical uniformities is the way science usually proceeds.

That Solow considered a "good theoretical explanation" of the behavior of labor productivity to be one consistent with the neoclassical framework is apparent from the fact that all the models he proposed in his address involved the assumption of optimization and a long-run production function with standard neoclassical properties.

Costs of Hiring, Firing, and Training Workers

By 1964, speculation was prevalent that the recently discovered procyclical labor productivity was a manifestation of various "fixed" or "overhead" costs associated with labor.[4] As one example, Okun's (1963, pp. 6–7) list of such costs included

[3] For example, some believed that neoclassical models were the best tools for understanding the behavior of market- and industry-level variables, while a Keynesian approach was better for the analysis of aggregate variables. Others limited the domain of neoclassical theory to the analysis of long-run trends in those markets and industries that closely approached the competitive ideal.

[4] The analysis of the costs associated with the labor input is an important part of the development of this element of the labor hoarding concept. Over the period examined, authors suggested various categorization schemes for labor costs to facilitate such analyses. Only in the late 1980s did a standard terminology for labor costs begin to emerge, with "variable" labor costs referring to costs that change with work time, holding constant the number of employees (for example, the hourly wage rate); "fixed" costs being those that depend on the number of employees on the payroll; and "adjustment costs" being those associated with changing the size of the labor force but not included in these

"contractual commitments" (contractual terms of employment, severance pay); "technological factors" (indivisibilities that prevented certain types of labor input from being varied proportionately with the variation in output); "transaction costs" associated with laying off labor, and then doing new hiring in the future; and "acquired skills" that might be lost if the worker could not be rehired after layoff. But discussions of such costs, and assertions that they would or should influence the response of the profit-maximizing employer to variations in product demand, actually had a history in economics going back at least to the scientific management movement of the early 20th century.

The rhetoric of scientific management envisioned the amelioration of many of the problems facing the industrial worker via the voluntary adoption by enlightened business owners of personnel management policies that were rational, fair, and also beneficial to the bottom line. In the early 1920s, then-Commerce Secretary Herbert Hoover convened a "Conference on Unemployment," one component of which was a study by the NBER of the facts regarding business cycles and unemployment. It included research into the policies being adopted by firms to "stabilize" employment in the face of seasonal and cyclical fluctuations, as it was hoped that wider knowledge of such policies would spur wider adoption. One chapter in the NBER study was written by N. I. Stone, a Columbia-trained economist and statistician then serving as an executive for clothing manufacturer Hickey Freeman. Stone (1923, p. 117) wrote:

> Apart from the social injury which intermittent production causes, a broad view of the ultimate interests of the individual manufacturing concern discloses the disadvantages of intermittent production and the gains that would flow from continuous operation . . . Business men now recognize the wastefulness of a large labor turnover, the expensiveness of training new help, and their inefficiency and resultant high cost . . . Added to these is the loss of the more capable and ambitious workers who drift away during periods of idleness to more steady occupations unless held by the inducement of higher rates of wages . . .

Stone also described methods employed by various firms to stabilize employment, which included labor hoarding strategies during economic downturns like "manufacturing to stock" and transferring "surplus help from one operation or department to another."

These ideas were discussed in a neoclassical framework in J. M. Clark's (1923) book *Studies in the Economics of Overhead Cost*. A central theme of Clark's book was

---

first two categories—for example, the cost of hiring a worker (search costs, paperwork) or the cost of training a new worker. These adjustment costs are also associated with labor turnover, as they are incurred each time an employee leaves and is replaced. Rather than employing technical terms that are/were specific to a particular author or time period, the exposition here uses the phrases "costs of hiring, firing, and training workers" or "costs of adjusting the labor force" to speak of the sorts of labor costs that were central to all discussions of this aspect of the labor hoarding concept.

that overhead costs often led to situations in which private costs and benefits of various business decisions deviated from their social costs and benefits. A chapter on "labor as an overhead cost" developed the concept of a social overhead cost to maintaining a healthy and properly trained labor force, highlighting the social waste associated with unemployment. Clark also explained, however, that some labor costs were overhead costs from the employer's point of view as well. For example, it was often desirable to "keep a nucleus of the working force together through a period of depression" so as to avoid "the cost of building up the force again when business revived" because "when new men are taken on, there is waste in teaching them." Clark (pp. 43–44, 50–51, 92, 94, 160, 184–186) used numerical examples and marginal analysis to show how these considerations affected the profit maximizing businessman's responses to fluctuations in demand. Clark's book went through seven printings over the next 25 years, helping to keep alive among economists the idea that employment adjustment costs mattered for hiring policies.

Formalization and Acceptance of the Labor Hoarding Concept

The acceptance by economists of the proposition that costs associated with adjusting the labor input were a likely reason for procyclical labor productivity owed much to the appearance in the early 1960s of a pair of formal models characterizing the firm's demand for labor over a multiperiod time horizon. One was in the book *Planning Production, Inventories, and Work Force* by Charles Holt, Franco Modigliani, John Muth, and Herbert Simon (1960); the other in Walter Oi's 1962 article "Labor as a Quasi-Fixed Factor," based on his 1961 doctoral dissertation. Each model represented the firm's employment decision as a mathematical optimization problem, and both implied that changes in product demand would not be met by proportional changes in employment.

The Holt–Modigliani–Muth–Simon (1960) model was presented as a tool for managers trying to make production plans in the presence of uncertainty regarding future demand. By the standards of the economics profession of the time, the model and the regression-based statistical methods recommended for implementing it were quite technically demanding. In deriving the cost function to be minimized when selecting employment and production levels, and describing the sorts of information needed by managers to estimate that function, the authors mentioned several items that had become common in discussions of overhead or fixed costs of labor, listing them under the broad headings Hiring and Training Costs, Layoff Costs, Overtime Costs, and Idle Time costs, and argued that the aggregate impact of these items on labor cost could be approximated by a quadratic cost function.[5] They also showed that the implementation of the model's optimal decisions rules would lead to a reduction in the fluctuations of both output and employment, but made

[5] The authors indicated that they were summarizing a discussion found in a 1956 issue of the *Journal of Industrial Engineering*, thus linking their work to the scientific management tradition discussed above.

no mention of the relationship between this fact and the cyclical behavior of labor productivity (Holt et al. 1960, pp. 18–20).

The Holt–Modigliani–Muth–Simon (1960) model was picked up quickly by the leading contributors to the macroeconometric literature dealing with the short-run behavior of employment, in which it was established that the assumption that individual firms faced costs of adjusting the labor force led to an econometrically tractable partial adjustment equation describing movements of aggregate employment (Soligo 1963; Kuh 1965; Solow 1964). Thus, Holt, Modigliani, Muth, and Simon provided the macroeconometric research community with a model expressed in a mathematical form that portrayed optimizing behavior, and that, when combined with a neoclassical production function, would lead to a short-run elasticity of aggregate employment with respect to output of less than unity.

Walter Oi (1962) developed his model of the demand for labor in the presence of fixed costs of employment in order to explain certain anomalies in the cyclical behavior of labor markets for which there were "no truly satisfying explanations," by which he meant explanations that assumed optimizing behavior by firms. Among the puzzles that motivated Oi were occupational differences in the stability of employment and earnings and certain discriminatory hiring and firing practices, but not procyclical labor productivity. In Oi's model, a firm with a standard neoclassical production function maximized profit per worker over a multiperiod time horizon. There were fixed costs of hiring and training each type of worker. The firm also had expectations regarding future wages and product demand. Solving the firm's optimization problem led to Oi's first important result: "Even under perfect competition wages would be equated to marginal products if and only if labor is a completely variable factor"—that is, if there were no fixed costs of hiring or training. In a long-run competitive equilibrium, each type of worker's value of marginal product would equal the worker's wage plus the amortized value of the cost of hiring and training that worker.

Oi then considered a firm in long-run equilibrium that learned of a drop in future product demand. Since, for current employees, the costs of hiring and training were sunk, the firm would retain a worker as long as the per-period marginal product was greater than the per-period wage. This led to the result that could be used to explain procyclical labor productivity: The employment of a labor type would only be reduced if its value marginal product, set above its wage in long-run equilibrium, fell below its wage. Thus, the employment of some labor types might not fall at all with a decline in product demand.

As a general matter, Oi's (1962) approach to using his model for analysis was less formal than that of Holt, Modigliani, Muth, and Simon (1960), as were his empirical methods. Rather than attempting to estimate structural parameters of a stochastic version of his model, Oi's approach to empirical testing involved identifying the correlations among observable variables implied by his model, then looking for those correlations in the data using chi-square tests and simple regressions.

While Oi (1962) did not discuss procyclical labor productivity, the relevance of his model to that topic was quickly recognized (for example, in Kuh 1965). The major impact of Oi's model, however, was to serve as a component in neoclassical models developed to explain seemingly anomalous labor market phenomena. Also, many economists whose theoretical priors led them to view procyclical labor productivity as a "puzzle" readily accepted Oi's model as a solution to the puzzle: procyclical labor productivity arose from labor hoarding, which arose from optimizing behavior by firms operating in a competitive environment and facing fixed costs of hiring, firing, and training workers (for example, Rosen 1969, p. 257).

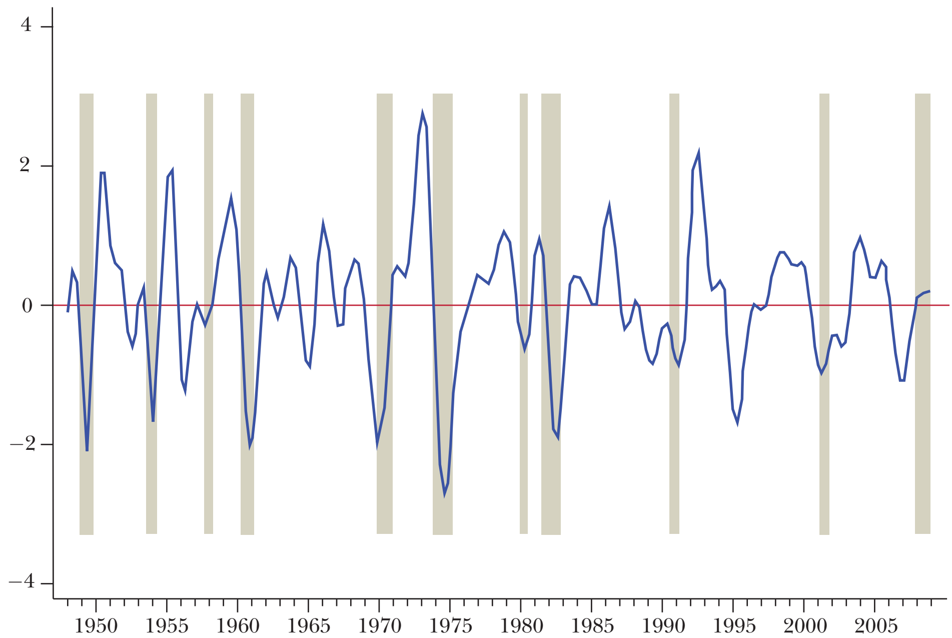
While a number of the macroeconometric models of the early 1960s incorporated the partial employment adjustment mechanism implied by the labor hoarding models, others incorporated an additional mechanism to explain procyclical labor productivity that was rooted in the technology of production rather than the costs associated with hiring labor (Soligo 1963, pp. 19, 30–31; Wilson and Eckstein 1964). This second type of explanation was related to the idea, mentioned by Mitchell in 1951, that modern plants were built to operate optimally at a particular capacity level of output, and that deviations from this level of output in either direction would lead to inefficiency.

Ray Fair, in his doctoral dissertation on *The Short Run Demand for Workers and Hours*, surveyed the relevant literature as of 1968 and expressed his unease with the technology-based explanation, arguing that an econometric model that attributed the positive correlation between productivity and output to increasing returns to labor was probably misspecified (Fair 1969, p. 31). Fair instead assumed a fixed-proportions short-run production function that led to constant returns to labor, and then, to account for procyclical labor productivity, argued that because of the fixed costs associated with the labor input, employers usually paid for more labor than they utilized. He also developed a procedure for constructing an empirical index of the implied "excess labor."

In 1985, Fay and Medoff published a paper that further bolstered the credibility of the modern labor hoarding concept. They sent questionnaires to managers of manufacturing plants asking about the size of the workforce maintained during the most recent recession. The authors' chief conclusion from their survey data was that at the trough of the business cycle, the typical plant employed 8 percent more labor than was technically necessary to produce its measured output, but that half of that labor was employed in other useful tasks, leaving 4 percent of the labor to be classified as truly "hoarded." Fair reacted to Fay and Medoff's data by applying his indirect method of measuring excess labor to aggregate data, "to see if the quantitative estimates of Medoff and Fay are consistent with the aggregate evidence," and found a similar answer. Fair (1985, p. 239) wrote: "If this is the case, which the results in this paper show, it provides a strong argument in favor of the excess labor hypothesis. Essentially the same conclusion has been reached using two very different data sets. This is one of the few examples in macroeconomics where a hypothesis has been so strongly confirmed using detailed micro data."

The argument that increasing returns in production play a role along with labor hoarding in generating procyclical labor productivity remained alive in the 1980s and 1990s (for example, Hall 1986). Also, discussions of labor hoarding began to include the suggestion that hoarded workers would be employed in tasks, such as

Figure 2
The Vanishing Procyclicality of Labor Productivity
(output per hour)



Source: Gali and van Rens (2010).
Notes: Figure shows output per hour in the US private sector, rendered stationary with a bandpass filter. Shaded areas are NBER-dated recessions.

maintenance, that did not show up in the firm's measured output, so that at least some of the observed positive correlation between labor productivity and output was due to measurement error (McConnell, Brue, and McPherson 1999; p. 545; Fay and Medoff 1985, p. 639). Still, as of the 1990s, the existence of labor hoarding as a response to costs of hiring, firing, and training workers was widely accepted and believed to be a likely explanation of the observed procyclicality of labor productivity.

Afterword: Labor Productivity Turns Countercyclical?

Although this fact has only recently become apparent, the procyclical tendency of labor productivity began to weaken in the US economy in the 1980s, culminating in the trend-adjusted increase in labor productivity that occurred during the Great Recession. Various authors have demonstrated this development econometrically and graphically, including Barnichon (2010), Mulligan (2011), and Gordon (2011). Figure 2 shows a recent visual representation from Gali and van Rens (2010) in which the upward movement of productivity during the last two recessions can be seen.

The change in the cyclical behavior of US labor productivity poses a challenge for macroeconomists. It is not necessarily inconsistent with the model of employer behavior underlying the labor hoarding hypothesis: for example, if the procyclical labor productivity that prevailed during most of the 20th century resulted from costs of adjusting employment levels, then the disappearance of procyclical labor productivity may be due to reductions in the costs of adjusting employment levels. This line of attack is suggested by Ohanian (2010) and developed by Gali and van Rens (2010) and Gordon (2011). The countercyclical move of labor productivity is perhaps a greater puzzle from the perspective of real business cycle models, which, since their introduction in the 1980s, have accounted for procyclical labor productivity by means of their fundamental assumption that business cycles are driven by shocks to productivity (Plosser 1989). However, those working in this tradition, like McGratten and Prescott (2012), are now exploring mechanisms by which exogenous shocks to productivity might generate countercyclical movements of measured labor productivity. Taking yet another approach, Mulligan (2011) argues that productivity increase during the Great Recession arose from a dampening effect on labor supply of rising marginal tax rates.

A half-century ago, the results of Hultgren (1960) and Kuh (1960) convinced economists that labor productivity was procyclical. This fact was perceived to be a puzzle and led to a revised account of the operation of labor markets, neoclassical in spirit but synthesized out of disparate ideas, and then built into macroeconomic models. The recent countercyclical behavior of US labor productivity is likewise a puzzle, and seems to be stimulating a rethinking of how labor markets and the macroeconomy function during cyclical downturns.

References

Barnichon, Regis. 2010. "Productivity and Unemployment over the Business Cycle." *Journal of Monetary Economics* 57(8): 1013–25

Brechling, F. P. R. 1965. "The Relationship Between Output and Employment in British Manufacturing Industries." *Review of Economic Studies* 32(3): 187–216.

Burns, Arthur F. 1952. "Introductory Sketch." In *Wesley Clair Mitchell: The Economic Scientist*, edited by Arthur Burns, 3–54. New York: National Bureau of Economic Research.

Clark, J. M. 1923. *Studies in the Economics of Overhead Costs.* University of Chicago Press.

Dornbusch, Rudiger, and Stanley Fischer. 1981. *Macroeconomics*, 2nd edition. New York: McGraw Hill.

Douglas, Paul. 1922. "Personnel Problems and the Business Cycle." *Administration* 4(1): 15–27.

Dunlop, John T. 1938. "The Movement of Real and Money Wage Rates." *Economic Journal* 48(191): 413–34.

Fabricant, Solomon. 1959. "Basic Facts on Productivity Change." Occasional Paper 63. New York: National Bureau of Economic Research.

Fair, Ray C. 1969. *The Short Run Demand for Workers and Hours.* Amsterdam: North Holland, 1969.

Fair, Ray C. 1985. "Excess Labor and the Business Cycle." *American Economic Review* 75(1): 239–45.

Fay, Jon A., and James Medoff. 1985. "Labor and Output over the Business Cycle: Some Direct Evidence." *American Economic Review* 75(4): 638–55.

Flanagan, Robert J., Robert S. Smith, and Ronald G. Ehrenberg. 1984. *Labor Economics and Labor Relations.* Glenview, IL: Scott, Foresman, and Co.

Gali, Jordi, and Thijs van Rens. 2010. "The Vanishing Procyclicality of Labor Productivity." Unpublished manuscript. Accessed at http://www.thijsvanrens.com/VPLP/.

Gordon, Robert J. 2011. "The Evolution of Okun's Law and of Cyclical Productivity Fluctuations in the United States and in the EU-15." Paper prepared for Presentation at EES/IAB Workshop, "Labor Market Institutions and the Macroeconomy," Nuremberg, June 17–18, 2011. http://faculty-web.at.northwestern.edu/economics/gordon/NUE_EU_vs_20_US_Combined_110609.pdf.

Hall, Robert E. 1986. "Market Structure and Macroeconomic Fluctuations." *Brookings Papers on Economic Activity*, no. 2, pp. 285–338.

Hamermesh, Daniel, and Albert Rees. 1984. *The Economics of Work and Pay*, 3rd ed. New York: Harper and Rowe.

Holt, Charles C., Franco Modigliani, John F. Muth, and Herbert A. Simon. 1960. *Planning Production, Inventories, and Work Force.* Englewood Cliffs, NJ: Prentice-Hall.

Hultgren, Thor. 1948. *American Transportation in Prosperity and Depression.* New York: National Bureau of Economic Research.

Hultgren, Thor. 1960. "Changes in Labor Costs during Cycles of Production and Business." New York: National Bureau of Economic Research.

Kuh, Edwin. 1960. "Profits, Profit Markups, and Productivity." Study Paper No. 15 of the Joint Economic Committee Study of Employment, Growth, and Price Levels. Washington: USGPO.

Kuh, Edwin. 1965. "Cyclical and Secular Labor Productivity in United States Manufacturing." *Review of Economics and Statistics* 47(1): 1–12.

McConnell, Campbell, and Stanley Brue. 1986, 1989. *Contemporary Labor Economics,* 1st and 2nd editions. New York: McGraw Hill. (Fifth edition in 1999 also includes David A. MacPhereson as an author).

McGrattan, Ellen R., and Edward C. Prescott. 2012. "The Labor Productivity Puzzle." Working Paper 694, Federal Reserve Bank of Minneapolis.

Mills, Frederick C. 1933. "Changes in Physical Production, Industrial Productivity and Manufacturing Costs, 1927–1932." NBER Bulletin 45, February 23.

Mitchell, Wesley. 1913. *Business Cycles.* Berkeley: University of California Press.

Mitchell, Wesley Clair. 1951. *What Happens During Business Cycles: A Progress Report.* New York: National Bureau of Economic Research.

Mulligan, Casey B. 2011. "Rising Labor Productivity During the 2008–09 Recession." NBER Working 17984, November.

National Bureau of Economic Research. 1938. *Report of the Executive Director for 1937.* New York: National Bureau of Economic Research. Accessed at http://papers.nber.org/books/unkn38-3.

Neild, R. R. 1963. "Pricing and Employment in the Trade Cycle." Occasional Paper 21 of the National Institute of Economic and Social Research. Cambridge University Press.

Ohanian, Lee E. 2010. "The Economic Crisis from a Neoclassical Perspective." *Journal of Economic Perspectives* 24(4): 45–66

Oi, Walter Y. 1962. "Labor as a Quasi-Fixed Factor." *Journal of Political Economy* 70(6): 538–55.

Okun, Arthur. 1963. "Potential GNP: Its Measurement and Significance." Cowles Foundation Paper 190 (reprinted from the 1962 Proceedings of the Business and Economics Statistics Section of the American Statistical Association). Available at http://cowles.econ.yale.edu/P/cp/py1963.htm.

Plosser, Charles I. 1989. "Understanding Real Business Cycles." *Journal of Economic Perspectives* 3(3): 51–77.

Rosen, Sherwin. 1969. "On the Interindustry Wage and Hours Structure." *Journal of Political Economy* 77(2): 249–73.

Soligo, Ronald. 1963. *The Short Run Relationship between Employment and Output.* Dissertation. Yale University. Available from ProQuest.

Solow, Robert. 1964. Draft of Presidential Address on the Short Run Relation of Employment and Output. Robert Solow papers, David M. Rubenstein Rare Book & Manuscript Library, Duke University. Box 70, Folder 1964, dated 1964.

Stone, N. I. 1923. "Methods of Stabilizing Production of Textiles, Clothing, and Novelties." In *Business Cycles and Unemployment*, edited by Committee of the President's Conference on Unemployment, and a Special Staff of the National Bureau, pp. 118–35. New York: National Bureau of Economic Research.

Tarshis, Lorie. 1939. "Changes in Real and Money Wages." *Economic Journal* 49(193): 150–54.

Wilson, Thomas A., and Otto Eckstein. 1964. "Short-Run Productivity Behavior in U.S. Manufacturing." *Review of Economics and Statistics* 46(1): 41–54.

Woirol, Gregory R. 2006. "New Data, New Issues. The Origins of the Technological Unemployment Debate." *History of Political Economy* 38(3): 473–96.

Journal of Economic Perspectives—Volume 28, Number 2—Spring 2014—Pages 213–220

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., Saint Paul, Minnesota, 55105.

Smorgasbord

The IMF Fiscal Affairs Department provides an international overview of "Fiscal Policy and Income Inequality." "Between the late-1990s and the late-2000s, public support for redistribution increased in almost 70 percent of the advanced and developing economies surveyed. For instance, support increased substantially in Finland, Germany, and Sweden, and also in China and India. In the late-1990s, results for only 15 economies out of the 57 in the sample (26 percent) indicate majority support for more redistribution. By the late-2000s, the percentage of countries where a majority supported more redistribution grew to 56 percent. . . . Over recent decades, direct income taxes and transfers have decreased inequality in advanced economies by an average of one-third. For instance, in 2005, the average Gini for disposable income

■ *Timothy Taylor is Managing Editor, Journal of Economic Perspectives, based at Macalester College, Saint Paul, Minnesota. He blogs at http://conversableeconomist.blogspot.com.*

http://dx.doi.org/10.1257/jep.28.2.213. doi=10.1257/jep.28.2.213

was 14 percentage points below that of the average market income Gini. The redistributive impact of transfers accounts for about two-thirds of the decrease in the Gini. Within transfers, non-means-tested transfers (including public pensions and family benefits) account for the bulk of the redistribution. On the tax side, personal income taxes make an important contribution to reducing inequality in a number of economies—in fact, in most economies, the redistribution achieved through income taxes is even higher than for means-tested transfers." IMF Policy Paper, January 23, 2014. http://www.imf.org/external/np/pp/eng/2014/012314.pdf.

Lawrence H. Summers delivered the 2013 Martin Feldstein lecture on the topic, "Economic Possibilities for our Children." "When I was an MIT undergraduate in the early 1970s, a young economics student was exposed to the debate about automation. There were two factions in those debates. There were the stupid Luddite people, who mostly were outside of economics departments, and there were the smart progressive people, who at that time were personified by Bob Solow. The stupid people thought that automation was going to make all the jobs go away and there wasn't going to be any work to do. And the smart people understood that when more was produced, there would be more income and therefore there would be more demand. It wasn't possible that all the jobs would go away, so automation was a blessing. I was taught that the smart people were right. Until a few years ago, I didn't think this was a very complicated subject; the Luddites were wrong and the believers in technology and technological progress were right. I'm not so completely certain now. . . . In the United States today a higher fraction of the workforce receives disability insurance than does *production work* in manufacturing. (Many workers in the manufacturing sector are not production workers.) . . . I think it is also fair to say that the evolution and growth of disability insurance is substantially driven also by the technological and social changes that are leading to a smaller fraction of the workforce working. At the same time, as has famously and repeatedly been noted, the share of income going to the top one percent of our population has steadily increased." *NBER Reporter,* 2013, Number 4, http://www.nber.org/reporter/2013number4/2013no4.pdf.

David Roodman discusses "Armageddon or Adolescence? Making Sense of Microfinance's Recent Travails." "Microfinance has been growing for 35 years and now reaches upwards of 100 million people, who cannot all be wrong in their judgments about the utility of microfinance. Moreover, most of them are served by institutions that are nearly or completely self-sufficient in financial terms . . . Because of the vicissitudes of poverty, poor people need financial services more than the rich. Their financial options will always be inferior—that's part of being poor—and microfinance offers additional options with distinctive strengths and weaknesses." "Nevertheless, the recent travails are signs that something is wrong in the industry. What is wrong is, ironically, what was once so right about the industry: it largely bypassed governments in favor of an experimental, bottom-up approach to institution building. The industry got so good at building institutions and injecting funds into them that it often forgot that a durable financial system consists of more than retail institutions and their investors. . . . To mature, the industry and its supporters should recognize the imbalance it has created. Where possible, they should work to strengthen institutions of moderation such as credit bureaus and regulators. Accepting that such institutions will often be weak, they should err on the side of investing less. In microfinance funding, less is sometimes more." Center for Global Development Policy Paper 35, January 2014. At http://www.cgdev.org/sites/default/files/armaggedon-adolescence-microfinance-recent-travails_final_1.pdf.

In "Whither the Euro?" Kevin Hjortshøj O'Rourke contends: "For years economists have argued that Europe must make up its mind: move in a more federal direction, as seems required by the logic of a single currency, or move backward? It is now 2014: at what stage do we conclude that Europe has indeed made up its mind, and that a deeper union is off the table? The longer this crisis continues, the greater the anti-European political backlash will be, and understandably so: waiting will not help the federalists. We should give the new German government a few months to surprise us all, and when it doesn't, draw the logical conclusion. With forward movement excluded, retreat from the EMU may become both inevitable and desirable. . . . The demise of the euro would be a major crisis, no doubt about it. We shouldn't wish for it. But if a crisis is inevitable then it is best to get on with it, while centrists and Europhiles are still in charge. Whichever way we jump, we have to do so democratically, and there is no sense in waiting forever. If the euro is eventually abandoned, my prediction is that historians 50 years from now will wonder how it ever came to be introduced in the first place." *Finance and Development*, March 2014, pp. 14–16. This issue also includes several other articles on the euro and the European economy. Reza Moghadam writes on "Europe's Road to Integration"; Nicolas Véron discusses prospects for a banking union in "Tectonic Shifts"; and Helge Berger and Martin Schindler consider policies for reducing unemployment and spurring growth in "A Long Shadow over Growth." At http://www.imf.org/external/pubs/ft/fandd/2014/03/index.htm.

The World Bank has published "Social Gains in the Balance: A Fiscal Policy Challenge for Latin America and the Caribbean." "In 2012, the Latin America and the Caribbean (LAC) region continued its successful drive to reduce poverty and build the middle class. The proportion of the region's 600 million people living in extreme poverty, defined in the region as life on less than $2.50 a day, was cut in half between 2003 and 2012 to 12.3 percent. . . . The middle class, currently 34.3 percent of the population, is growing rapidly and is projected to replace the vulnerable as the largest economic group in LAC by 2016. . . . About 68 percent of poverty reduction between 2003 and 2012 was driven by economic growth, with the remaining 32 percent arising from decline in inequality. Poverty reduction was accompanied by strong income growth of the bottom 40 percent of the population, the World Bank's indicator of shared prosperity. Between 2003 and 2012, the real per capita income of the bottom 40 percent grew by more than five percent annually, while overall income in LAC rose by about 3.3 percent. However . . . [t]he region suffered an economic slowdown from an annual GDP per capita growth rate of about 4.3 percent in 2010 to an estimated 1.3 percent in 2013 and is projected to grow at only 1.7 percent in 2014. Also, after falling steadily between 2001 and 2010, progress in reducing inequality in LAC has stagnated with the Gini coefficient remaining fairly constant at 0.52." February 2014. Available at: http://documents.worldbank.org/curated/en/2014/02/19120905/social-gains-balance-fiscal-policy-challenge-latin-america-caribbean.

The *Journal of Medical Ethics* has a symposium on the issue of whether people should be allowed to sell a kidney. The lead article is by Simon Rippon: "Imposing options on people in poverty: the harm of a live donor organ market." Because selling a kidney is not a legal option, Rippon argues, "This means that even if you have no possessions to sell and cannot find a job, nobody can reasonably criticise you for, say, failing to sell a kidney to pay your rent. If a free market in organs was permitted and became widespread, then it is reasonable to assume that your organs would soon enough become economic resources like any other, in the context of the market. Selling your organs would become something that is simply expected of you as and when financial need arises. . . . We should ask questions such as the following: Would those in poverty be eligible for bankruptcy protection, or for public assistance, if they have an organ that they choose not to sell? Could they be legally forced to sell an organ to pay taxes, paternity bills, or rent? How would society view someone who asks for charitable assistance to meet her basic needs, if she could easily sell a healthy 'excess' organ to meet them? . . . Wherever there is great value in not being put under social or legal pressure to sell something as a result of economic forces, we should think carefully about whether it is right to permit a market and to thereby impose the option on everyone to sell it." Comments follow from Gerald Dworkin, Janet Radcliffe-Richards, and Adrian Walsh, along with a response from Rippon (although the comments are not freely accessible online). March 2014, pp. 145–50. Available at: http://jme.bmj.com/content/40/3.toc#Featurearticle. A Symposium on Organ Transplants appeared in the Summer 2007 issue of this journal.

Jonathan Huntley summarizes estimates of "The Long-Run Effects of Federal Budget Deficits on National Saving and Private Domestic Investment." His central estimate about the long-run effects of more government borrowing: for each additional dollar of government budget deficit, private saving rises by 43 cents, and the inflow of foreign capital rises by 24 cents and domestic investment declines by 33 cents. Congressional Budget Office Working Paper 2014-2, February 2014. http://www.cbo.gov/sites/default/files/cbofiles/attachments/45140-NSPDI_workingPaper.pdf.

Tony Atkinson and Salvatore Morelli have produced an intriguing "Chartbook of Economic Inequality." The chartbook features a set of figures showing long-run trends in inequality as measured by a variety of statistics for 25 different countries, with all the statistics appearing on a single chart for each country. The charts appear in two forms: there's a colorful online version and then a black-and-white version that can be printed out from a PDF file. March 2014, http://www.chartbookofeconomicinequality.com.

Smoking

In 1964, the US Surgeon General issued its first report finding that smoking was hazardous to your health. Now in 2014, the current Surgeon General has published

The Health Consequences of Smoking—50 Years of Progress. Most of the nearly 1,000-page report focuses on health effects of tobacco use, but several chapters near the end focus on policies for reducing tobacco use: “Despite declines in the prevalence of current smoking, the annual burden of smoking-attributable mortality in the United States has remained above 400,000 for more than a decade and currently is estimated to be about 480,000, with millions more living with smoking-related diseases. . . . Annual smoking-attributable economic costs in the United States estimated for the years 2009–2012 were between $289–332.5 billion, including $132.5–175.9 billion for direct medical care of adults, $151 billion for lost productivity due to premature death estimated from 2005–2009, and $5.6 billion (in 2006) for lost productivity due to exposure to secondhand smoke.” http://www.surgeongeneral.gov/library/reports/50-years-of-progress/full-report.pdf.

The January 8, 2014, issue of the *Journal of the American Medical Association* (JAMA) has a useful set of articles reviewing the evidence and arguments. In “Tobacco Control and the Reduction in Smoking-Related Premature Deaths in the United States, 1964–2012,” Theodore R. Holford, Rafael Meza, Kenneth E. Warner, Clare Meernik, Jihyoun Jeon, Suresh H. Moolgavkar, and David T. Levy take on the task of estimating how much smoking in the United States has been reduced as a result of the antismoking efforts. They write (bracketed information about the statistical confidence intervals deleted): “In 1964–2012, an estimated 17.7 million deaths were related to smoking, an estimated 8.0 million fewer premature smoking-related deaths than what would have occurred under the alternatives and thus associated with tobacco control (5.3 million men and 2.7 million women). This resulted in an estimated 157 million years of life saved, a mean of 19.6 years for each beneficiary (111 million for men, 46 million for women). During this time, estimated life expectancy at age 40 years increased 7.8 years for men and 5.4 years for women, of which tobacco control is associated with 2.3 years (30%) of the increase for men and 1.6 years (29%) for women.” Table of contents for the issue available at: http://jama.jamanetwork.com/Issue.aspx?journalid=67&issueID=929635&direction=P.

About Economists

The newsletter of the Committee on the Status of Women in the Economics Profession (CSWEP) has published “A Celebration of the Life of Anna Jacobson Schwartz,” with remembrances from eight contributors. Michael Bordo, who co-authored 30 articles and two books with her, said: “What I remember most about Anna is how much she loved her work. Her whole life was organized around going to the office. She officially retired from the Bureau when she was 65, but she didn’t stop working until she was 94. She went into the Bureau every day when she was in her eighties and nineties, and she still put in a full eight-hour day. . . . Yet she was a balanced person. She had a great family—Isaac, a caring husband with a great sense of humor, who died in 1999, four children, and many grandchildren and great grandchildren, and they used to come into New York to see her often.

She had season tickets to the Metropolitan Opera, which she loved, and she rarely missed a performance. She was a very active person in other dimensions as well. She always had a few novels going, and especially liked Anthony Trollope. She was on top of what was going on in politics and economic policy everywhere in the world. She read the *Wall Street Journal* and the *New York Times* each day, picking up every little detail. She never missed a beat." Eloise Pasachoff, one of Schwartz's grandchildren, wrote: "I think I have a bigger lesson, and it's about what they call 'work-life balance.' Except when I think about my grandmother's example, I want to call it 'work-life joy.'" Fall 2013. http://www.aeaweb.org/committees/cswep/newsletters/CSWEP_nsltr_Fall_2013.pdf.

Aaron Steelman has an "Interview" with John Cochrane. On Dodd–Frank: "I think Dodd–Frank repeats the same things we've been trying over and over again that have failed, in bigger and bigger ways. . . . The deeper problem is the idea that we just need more regulation—as if regulation is something you pour into a glass like water—not smarter and better designed regulation. Dodd-Frank is pretty bad in that department. It is a long and vague law that spawns a mountain of vague rules, which give regulators huge discretion to tell banks what to do. It's a recipe for cronyism and for banks to game the system to limit competition." On how to stop bailing out large financial institutions: "You have to set up the system ahead of time so that you either can't or won't need to conduct bailouts. Ideally, both. . . . The worst possible system is one in which everyone thinks bailouts are coming, but the government in fact does not have the legal authority to bail out." On time-varying risk premiums: "One big unresolved issue in finance is why risk premiums are so big and why they vary so much over time. You can look at the spread between what you have to pay to borrow and what the U.S. government pays in order to see that risk premiums are big and varying. . . . For macroeconomics, the fact of time-varying risk premiums has to change how we think about the fundamental nature of recessions. Time-varying risk premiums say business cycles are about changes in people's ability and willingness to bear risk." *Econ Focus*, Federal Reserve Bank of Richmond, Third Quarter 2013, pp. 34–38. https://www.richmondfed.org/publications/research/econ_focus/2013/q3/pdf/interview.pdf.

Douglas Clement interviews Neil Wallace, with much of the discussion focused on the underpinnings and functions of money. Here's Wallace on the idea of "money is memory": "'Money is memory' is a better idea. It leads you to think about various kinds of payment instruments in terms of the kind of informational structure that supports them. The money that is the best current counterpart to the 'money is memory' idea is currency. You don't need much of an informational network for currency; in fact, you probably don't need any, except for worrying about counterfeiting. When you use a credit card, you're issued a loan. Why are you able to receive one? Because there's an informational network behind your card. Your bank is actually guaranteeing your credit payment up to probably some large amount, as large as you mostly use. And they're doing that because they know something about you." On how money and banking doesn't fit easily into a standard economics framework: "This goes back to economists' feelings that the general competitive model, often labeled the Arrow–Debreu model, is the main model in economics. It's very general. We don't need to have a special theory of production for bookcases and a special theory for bottled water. But when people try to shove banking into this model, it's hugely unsuccessful. Why? Because anything that banks might be viewed as doing is redundant in that model. According to the Arrow–Debreu model, you face prices at which you can costlessly trade anything for anything. More generally, no activity that we see in the economy that has to do with transacting fits comfortably within that model. In particular, nothing in the GDP accounts that falls under the FIRE heading—finance, insurance, real estate—fits into that model." *The Region,* Federal Reserve Bank of Minneapolis, December 2013, pp. 12–24. https://www.minneapolisfed.org/pubs/region/13-12/region_december_2013_interview_with_neil_wallace.pdf.

Discussion Starters

Melissa S. Kearney and Phillip B. Levine ask "Teen Births Are Falling: What's Going On?" "We speculate that there are two likely candidate explanations: (1) access to improved contraceptive technologies, most notably long-acting reversible contraception (LARCs) such as implants and intrauterine devices (IUDs) and (2) increased educational attainment along with better labor market prospects for young women. . . . The policy challenge that we believe offers the greatest potential is to address the needs of those young women who are not committed to avoiding a pregnancy. These are teens whose views are characterized by ambivalence. For them, the issue is more about finding ways to make them want to avoid a teen birth. . . . Simply put, increased aspirations and expanded opportunities for young women have the potential to extend the downward trend in teen childbearing." Brookings Institution, Economic Studies Group Policy Brief, March 2014 Policy Brief for the Economics Studies group, http://www.brookings.edu/~/media/research/files/reports/2014/03/teen-births-falling-whats-going-on-kearney-levine/teen_births_falling_whats_going_on_kearney_levine.pdf. This essay complements the article by these two authors "Why Is the Teen Birth Rate in the United States So High and Why Does It Matter?" in the Spring 2012 issue of this journal.

Harald Bauer, Jan Veira, and Florian Weig consider "Moore's Law: Repeal or Renewal?" "Moore's law states that the number of transistors on integrated circuits doubles every two years, and for the past four decades it has set the pace for progress in the semiconductor industry. . . . Adherence to Moore's law has led to continuously falling semiconductor prices. Per-bit prices of dynamic random-access memory chips, for example, have fallen by as much as 30 to 35 percent a year for several decades. . . . Some estimates ascribe up to 40 percent of the global productivity growth achieved during the last two decades to the expansion of information and communication technologies made possible by semiconductor performance and cost improvements." But this continued technological progress comes at an ever-higher price. "A McKinsey analysis shows that moving from 32nm to 22nm nodes on 300-millimeter (mm) wafers causes typical fabrication costs to grow by roughly 40 percent. It also boosts the costs associated with process development by about 45 percent and with chip design by up to 50 percent. These dramatic increases will lead to process-development costs that exceed $1 billion for nodes below 20nm. In addition, the state-of-the art fabs needed to produce them will likely cost $10 billion or more. As a result, the number of companies capable of financing next-generation nodes and fabs will likely dwindle." McKinsey Global Institute, December 2013, http://www.mckinsey.com/insights/high_tech_telecoms_internet/moores_law_repeal_or_renewal.

The Office of Inspector General of the US Postal Service has published a report considering if the Post Office might be a useful mechanism for "Providing Non-Bank Financial Services for the Underserved." "The Postal Service has played a longstanding role in providing domestic and international money orders. The Postal Service is actually the leader in the U.S. domestic paper money order market, with an approximately 70 percent market share. This is a lucrative business line and demonstrates that the Postal Service already has a direct connection to the underserved, who purchased 109 million money orders in fiscal year (FY) 2012. . . . While its domestic and international money orders are currently paper-based, the Postal Service does offer electronic money transfers to nine Latin American countries through the Dinero Seguro® (Spanish for 'sure money') service. For several years now, the Post Office has been selling debit cards, both for American Express and for specific retailers like Amazon, Barnes & Noble, Subway, and Macy's." January 27, 2014. http://www.uspsoig.gov/sites/default/files/document-library-files/2014/rarc-wp-14-007.pdf. The Universal Postal Union published a report in March 2013 by Alexandre Berthaud and Gisela Davico, *Global Panorama on Postal Financial Inclusion: Key Issues and Business Models*, which notes that 1 billion people around the world in 50 countries do at least some of their banking through postal banking systems. http://www.upu.int/fileadmin/documentsFiles/activities/financialInclusion/publicationGlobalPanoramaFinancialInclusionEn.pdf.

Journal of Economic Perspectives—Volume 28, Number 2—Spring 2014—Pages 221–226

Correction and Update: The Economic Effects of Climate Change†

Richard S. J. Tol

Gremlins intervened in the preparation of my paper "The Economic Effects of Climate Change" published in the Spring 2009 issue of this journal. In Table 1 of that paper, titled "Estimates of the Welfare Impact of Climate Change," minus signs were dropped from the two impact estimates, one by Plambeck and Hope (1996) and one by Hope (2006). In Figure 1 of that paper, titled "Fourteen Estimates of the Global Economic Impact of Climate Change," and in the various analyses that support that figure, the minus sign was dropped from only one of the two estimates. A coding error affected the upper bound of one of confidence intervals (the wider one) shown in the original Figure 1.

The corresponding Table 1 and Figure 1 below correct these errors. Figure 2 titled, "Twenty-One Estimates of the Global Economic Impact of Climate Change" adds two overlooked estimates from before the time of the original 2009 paper and five more recent ones. The confidence interval of the original erroneous impact curve overlaps with the corrected and updated one (compare Figure 1 to Figure 2). The parameters of the impact curves are not statistically significantly different from one another—neither between the original and the corrected impacts nor between the original and corrected and updated impacts, as you can see in Table 2. Estimates are few and the future is uncertain so that confidence intervals are wide.

■ *Richard S. J. Tol is Professor of Economics, University of Sussex, Falmer, United Kingdom. He is also Professor of the Economics of Climate Change, Institute for Environmental Studies and Department of Spatial Economics, Vrije Universiteit, Amsterdam, Netherlands; Research Fellow, Tinbergen Institute, Amsterdam, Netherlands; and Research Fellow, CESifo, Munich, Germany. His email address is r.tol@sussex.ac.uk.*

† To access the dataset, visit
http://dx.doi.org/10.1257/jep.28.2.221 doi=10.1257/jep.28.2.221

Table 1
Corrected and Updated Estimates of the Welfare Impact of Climate Change
(changed estimates in bold; previously omitted estimates in italics)

		Impact (% GDP)			
Study	*Warming (°C)*	*Central estimate*	*SD*	*Min*	*Max*
Estimates from papers summarized in Tol (2009)					
(Nordhaus 1994b)	3.0	−1.3			
(Nordhaus 1994a)	3.0	−4.8		−30.0	0.0
(Fankhauser 1995)	2.5	−1.4			
(Tol 1995)	2.5	−1.9			
(Nordhaus and Yang 1996)	2.5	−1.7			
(Plambeck and Hope 1996)	2.5	**−2.5**[a]		**−11.4**	**−0.5**
(Mendelsohn et al. 2000)	2.5	0.0			
	2.5	0.1			
(Nordhaus and Boyer 2000)	2.5	−1.5			
(Tol 2002)	1.0	2.3	1.0		
(Maddison 2003)	2.5	−0.1			
(Rehdanz and Maddison 2005)	1.0	−0.4			
(Hope 2006)	2.5	**−0.9**[b]		**−2.7**[b]	**0.2**[b]
(Nordhaus 2006)	2.5	−0.9	0.1		
	3.0[c]	*−1.1*[c]	*0.1*[c]		
(Nordhaus 2008)[c]	*3.0*[c]	*−2.5*[c]			
New estimates that appeared after Tol (2009)					
(Maddison and Rehdanz 2011)	3.2	−11.5			
(Bosello et al. 2012)	1.9	−0.5			
(Roson and van der Mensbrugghe 2012)	2.9	−1.8			
	5.4	−4.6			
(Nordhaus 2013)	2.9	−2.0			

Notes: The welfare impact of climate change is expressed as an equivalent income gain or loss in percent GDP. SD is standard deviation.
[a] This estimate was reported incorrectly as 2.5 percent in Table 1 but correctly as −2.5 percent in Figure 1 of Tol (2009).
[b] This estimate was reported incorrectly as 0.9 percent with confidence interval as [0.02, 2.7] in Table 1 and Figure 1 of Tol (2009).
[c] These estimates, an additional one from Nordhaus (2006) and one from Nordhaus (2008), are overlooked estimates that did not make it into Tol (2009).

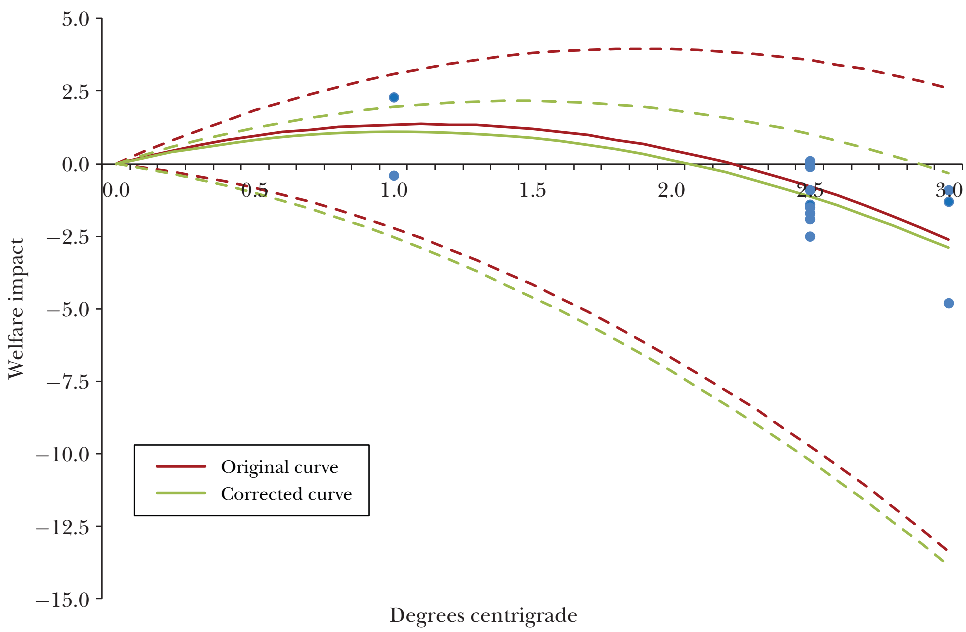
I nonetheless highlight two differences between the old and the new results. First, unlike the original curve (Tol 2009, Figure 1) in which there were net benefits of climate change associated with warming below about 2°C, in the corrected and updated curve (Figure 2), impacts are always negative, at least in expectation. This is irrelevant for policy because, as I discussed in that paper, the net benefits reported for earlier stages of climate change were sunk benefits; these benefits would have been reaped regardless of mitigation policy. Second, the corrected and updated damages do not accelerate as fast for more pronounced warming. For instance, the

Table 2
Estimates of the Parameters of the Climate Change Impact Curve

Parameter	*Unit*	*Original*		*Corrected*		*Corrected and updated*	
Expectation							
Linear	$\%GDP\Delta°C^{-1}$	2.46	(1.25)	2.13	(1.05)	−0.28	(0.60)
Quadratic	$\%GDP\Delta°C^{-2}$	−1.11	(0.48)	−1.03	(0.40)	−0.16	(0.18)
Standard deviation							
Optimistic	$\%GDP\Delta°C^{-1}$	0.87	(0.28)	0.43	(0.16)	0.36	(0.14)
Pessimistic	$\%GDP\Delta°C^{-1}$	1.79	(0.87)	1.81	(0.86)	1.51	(0.76)

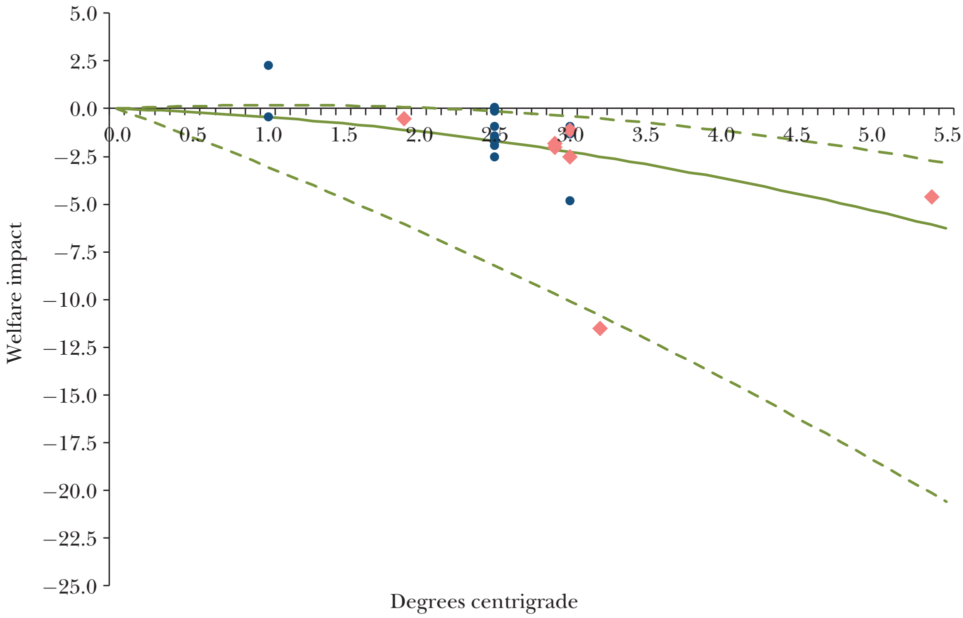
Notes: Table 2 gives the parameters of the curves in Figures 1 and 2. Impacts are measured in percent GDP for temperatures of degrees Celsius; the units of the parameters of the impact curve follow. Standard errors are shown in parentheses.

Figure 1
Fourteen Estimates of the Global Economic Impact of Climate Change



Notes: Figure 1 shows 14 estimates of the global economic impact of climate change, expressed as the welfare-equivalent income gain or loss, as a function of the increase in the annual global mean surface air temperature relative to preindustrial times. The dots represent the estimates (from Table 1). The central lines are the original and corrected least squares fits. The dashed lines are the boundaries of the original and corrected 95 percent confidence intervals. In all cases, the corrected curves are below the corresponding original curves.

Figure 2
Twenty-One Estimates of the Global Economic Impact of Climate Change



Notes: Figure 2 shows 21 estimates of the global economic impact of climate change, expressed as the welfare-equivalent income gain or loss, as a function of the increase in the annual global mean surface air temperature relative to preindustrial times. The figure includes two overlooked estimates from before the time of the original 2009 paper and five more recent ones. The dots and diamonds represent the estimates (from Table 1); dots were included in Tol (2009); diamonds are additional estimates. The central line is the least squares fit. The dashed lines are the boundaries of the 95 percent confidence interval.

original impact curve projects an impact of −15 (−7 to −33) percent of income for a 5°C warming, whereas the corrected and updated curve has −6 (−3 to −21) percent. This is relevant because the benefits of climate policy are correspondingly revised downwards.

The data for this erratum and update are now available at the JEP website along with the paper at http://e-jep.org.

■ *I am grateful to Bob Ward for finding a small error, to Mike Mastandrea for finding a bigger one, to Doug Arent for checking things again and again, to David Autor and Tim Taylor for their understanding, and to Ann Norman for superb editorial support. All remaining errors are, of course, mine and mine only.*

References

Bosello, Francesco, Fabio Eboli, and Roberta Pierfederici. 2012. "Assessing the Economic Impacts of Climate Change." *Review of Environment Energy and Economics*, February 10.

Fankhauser, Samuel. 1995. *Valuing Climate Change—The Economics of the Greenhouse*, 1st edition. London: EarthScan.

Hope, Chris. 2006. "The Marginal Impact of CO_2 from PAGE2002: An Integrated Assessment Model Incorporating the IPCC's Five Reasons for Concern." *Integrated Assessment Journal* 6(1): 19–56.

Maddison, David. 2003. "The Amenity Value of the Climate: The Household Production Function Approach." *Resource and Energy Economics* 25(2): 155–75.

Maddison, David, and Katrin Rehdanz. 2011. "The Impact of Climate on Life Satisfaction." *Ecological Economics* 70(12): 2437–45.

Mendelsohn, Robert, Wendy Morrison, Michael E. Schlesinger, and Natalia G. Andronova. 2000. "Country-Specific Market Impacts of Climate Change." *Climatic Change* 45(3–4): 553–69.

Nordhaus, William D. 1994a. "Expert Opinion on Climate Change." *American Scientist* 82(1): 45–51.

Nordhaus, William D. 1994b. *Managing the Global Commons: The Economics of Climate Change*. Cambridge. The MIT Press.

Nordhaus, William D. 2006. "Geography and Macroeconomics: New Data and New Findings." *Proceedings of the National Academy of Science* 103(10): 3510–17.

Nordhaus, William D. 2008. *A Question of Balance—Weighing the Options on Global Warming Policies*. New Haven: Yale University Press.

Nordhaus, William D. 2013. *The Climate Casino—Risk, Uncertainty, and Economics for a Warming World*. New Haven: Yale University Press.

Nordhaus, William D., and Joseph Boyer. 2000. *Warming the World: Economic Models of Global Warming*. The MIT Press.

Nordhaus, Willam D., and Zili Yang. 1996. "A Regional Dynamic General-Equilibrium Model of Alternative Climate-Change Policy" *American Economic Review* 86(4): 741–65.

Plambeck, Erica L., and Chris Hope. 1996. "PAGE95—An Updated Valuation of the Impacts of Global Warming." *Energy Policy* 24(9): 783–93.

Rehdanz, Katrin, and David J. Maddison. 2005. "Climate and Happiness." *Ecological Economics* 52(1): 111–25.

Roson, Roberto, and Dominique Van der Mensbrugghe. 2012. "Climate Change and Economic Growth: Impacts and Interactions." *International Journal of Sustainable Economy* 4(3): 270–85.

Tol, Richard S. J. 1995. "The Damage Costs of Climate Change toward More Comprehensive Calculations." *Environmental and Resource Economics* 5(4): 353–74.

Tol, Richard S. J. 2002. "Estimates of the Damage Costs of Climate Change—Part I: Benchmark Estimates." *Environmental and Resource Economics* 21(1): 47–73.

Tol, Richard S. J. 2009. "The Economic Effects of Climate Change." *Journal of Economic Perspectives* 23(2): 29–51.

Journal of Economic Perspectives—Volume 28, Number 2—Spring 2014—Pages 227–230

Farewell to Notes

Timothy Taylor

The great composer Johannes Brahms once remarked: "It is not difficult to compose; but it is incredibly difficult to let the superfluous notes drop under the table" (as quoted in Musgrave and Pascall 1987, p. 138). Here at the *Journal of Economic Perspectives*, the challenges of composing each issue remain, but the "Notes" have become superfluous, at least in their paper version.

The "Notes," as those who lurk in these back pages of JEP know well, announce forthcoming conferences, calls for papers, awards, and the like. However, the Internet has made it obsolete to deliver such information on paper in a quarterly journal. "News and Notes" from the American Economic Association is at http://www.aeaweb.org/news.php; outside announcements are on the "Bulletin Board" website at http://www.aeaweb.org/bulletinboard.php; and for a complete list of economics conferences by field area, see http://www.aeaweb.org/RFE/conferences.php.

But as we say farewell to the print version of the "Notes," a moment of remembrance seems appropriate. The first issue of the *American Economic Review*, published in 1911, found it worthwhile to devote 13 out of 219 total pages to "Notes." The first "Notes" (freely available online at http://www.jstor.org) was a mixed bag. It started with news on the growth of AEA membership, followed with an announcement more than a page long about the opening of a Bureau of Railway Economics in Washington, DC. The US Bureau of Labor announced a number of jobs for "trained economists." There was progress on the formation of Economic Clubs in various cities under the guidance of the National Economic League, and the Tulane Society of Economics continued to hold monthly meetings.

■ *Timothy Taylor is Managing Editor of the Journal of Economic Perspectives. His email address is taylort@macalester.edu.*

http://dx.doi.org/10.1257/jep.28.2.227. doi=10.1257/jep.28.2.227

The first "Notes" included mentions of conferences that had already happened: a New York State Conference on Taxation that included many government officials and a few economists; the first National Conference on Vocational Guidance; the Fourth Annual Convention of the National Society for the Promotion of Industrial Education; the eleventh annual meeting of the National Civic Federation; and James Mackaye who had given eight lectures on the "Philosophy of Socialism," including his plan for the transition from capitalism to socialism. There were mentions of conferences still coming: the Thirtieth Annual conference of the American Federation of Labor; the Universal Races Congress, where "one session will be devoted to special problems in inter-racial economics"; a second conference in Chicago on the Teaching of Economics.

There were updates on a number of research efforts: the formation of a Bureau of Social Research in Providence, Rhode Island; the work of the Research Department by the Women's Educational and Industrial Union of Boston; the work of the New York Commission on Employer's Liability and Causes of Industrial Accidents, Unemployment, and Lack of Farm Labor; and the efforts of the American Association for Labor Legislation to prohibit the use of phosphorus in the manufacture of matches.

The inaugural "Notes" included soon-to-be-published books: including *The Principles of Economics* from Frank Taussig; *The Navigable Rhine* by Edwin J. Clapp; the *Principles of Industrial Management* by J. C. Duncan; and a number of others. Library acquisitions were included. The New York State Library had received the Rensselaerswyck manuscripts, all 200 volumes and 25,000 papers, covering 200 years of the Dutch colony in what had become New York. Harry Wagner donated to Yale University Library 7,000 items from his collection of writing on the history of precious metals and currency. Mr. Simons, no first name given but former editor of the *Daily Socialist*, had donated his collection of books and 1,200 pamphlets to the University of Chicago library. The Hart, Schaffner, and Marx prizes for 1910 were announced, and Clapp's *The Navigable Rhine* won first prize of $600.

It included several pages of news about professors changing jobs or going on leave. For example, Richard T. Ely, after whom the annual Ely lecture would be named, was taking a leave from Wisconsin to study land problems in Germany and England. Harrison S. Smalley was leaving the University of Michigan to become Associate Professor of Economics in Leland Stanford Junior University. "Professor W. Z. Ripley of Harvard will be absent during the second half of the current academic year; a part of his vacation will be spent in Egypt."

This array of notes made considerable sense for a small organization, seeking to point out opportunities for its members at a time when the mechanisms for spreading specialized news were the postal service, the telegraph, and word-of-mouth. In 1911, the American Economic Association had 2,190 members, which was already a big jump from the 572 members the AEA had in 1893, when reliable membership data are first available (Siegfried 1998).

When the first issue of the JEP was published in 1987, we argued that the "Notes" should leave the AER and reside instead in the new journal. We wanted the "Notes" in

JEP, because we all knew (then) that the "Notes" mattered. It was a way of showing that the new journal was here to stay and (potentially) attracting some readers. The descriptions of already-occurred conferences and soon-to-be-published books had already been dropped from the "Notes" years before. The Fall 1987 issue of JEP included seven pages of notes; about two-thirds were announcements of conferences, calls for papers, and announcements of grants and awards, while the rest was deaths, retirements, promotions, and appointments.

Over time, fewer people sent news about promotions and appointments, so we closed down those categories. Frankly, we were grateful not to be sent this kind of news. AEA membership had risen from 4,154 members in 1945 to 17,835 members in 1968, and was above 20,000 members by 1987 (Siegfried 1998). If every institution sent their promotions, leaves, and appointments, it would have created pages of dull lists. Only a smattering of retirements and deaths were reported to JEP, and so we eventually stopped publishing them, too. More recently, those with conferences or calls for papers have become steadily less likely to send them to JEP. Our 1987 belief that it matters to print the "Notes" in the paper pages of JEP probably hasn't been true for years.

Admittedly, the ending of the "Notes" section as printed within the covers of the *Journal of Economic Perspectives* doesn't rank with some of the other great endings, like the revelation of what Citizen Kane meant by "Rosebud"; or "Forget it, Jake, it's Chinatown"; or "Oh, Auntie Em, there's no place like home!" But in its own small way, the end of the paper version of the "Notes" after its run of 103 years is one more sign of the remarkable changes in information and communication technology that surround us—and thus worth remarking.

References

Musgrave, Michael, and Robert Pascall. 1987. "The String Quartets Op. 51 No. 1 in C minor and No. 2 in A minor: A preface." In *Brahms 2: Biographical, Documentary, and Analytical Studies,* edited by Michael Musgrave, Michael, 137–43. University of Cambridge.

Siegfried, John J. 1998. "Who Is a Member of the AEA?" *Journal of Economic Perspectives* 12(2): 211–22.

2014 Application/Renewal for Membership

AMERICAN ECONOMIC ASSOCIATION
2014 Broadway, Suite 305
Nashville, TN 37203
Ph. 615-322-2595 fax: 615-343-7590
Federal ID No. 36-2166945
www.vanderbilt.edu/AEA

RENEWING MEMBERS, ENTER ACCT. NUMBER & EXP. DATE	IF PAYING BY CREDIT CARD, PLEASE FILL OUT BELOW
ACCOUNT NUMBER:	CARD NUMBER:
EXPIRATION DATE:	EXP DATE: CSC CODE:

FIRST NAME:	MI:	LAST NAME:
ADDRESS:		
CITY:	STATE/PROVINCE:	ZIP:
COUNTRY:	☐ Check here if non-US	
PHONE:	FAX:	
PRIMARY FIELD OF SPECIALIZATION:		
SECONDARY FIELD OF SPECIALIZATION:		
EMAIL:	☐ Check here to exclude your email address from the public directory	

Please include my email address to receive:
☐ Announcements about public policy affecting economists or the economics profession
☐ Surveys of economists for research purposes

MEMBERSHIP DUES — Based on annual income. Please select one below.

☐ Annual income of $70,000 or less	$20	$
☐ Annual income of $70,000 to $105,000	$30	$
☐ Annual income over $105,000	$40	$

The AEA dues above include online access to all seven AEA journals.
For print or CD subscription(s) indicate preference below and add appropriate charge(s).

Journal		Print	Int'l Postage*	CD*	
AER	(12 issues, incl. P&P)	☐ Add $25	☐ Add $35	☐ Add $25	$
AER Papers & Proceedings Only*		☐ Add $10	n/a	n/a	$
JEL	(4 quarterly issues)	☐ Add $15	☐ Add $15	☐ Add $15	$
JEP	(4 quarterly issues)	☐ Add $15	☐ Add $15	☐ Add $15	$
AEJ: Applied	(4 quarterly issues)	☐ Add $15	☐ Add $15	n/a	$
AEJ: Policy	(4 quarterly issues)	☐ Add $15	☐ Add $15	n/a	$
AEJ: Macro	(4 quarterly issues)	☐ Add $15	☐ Add $15	n/a	$
AEJ: Micro	(4 quarterly issues)	☐ Add $15	☐ Add $15	n/a	$

* Int'l postage applies only to print journals mailed outside of the U.S. No additional postage is required for CDs or the AER Papers and Proceedings.

AEA Journals via JSTOR online

JSTOR	☐ Add $16	$

				Sub Total	$
Check One:	☐ 1 Year	☐ 2 Years	☐ 3 Years	TOTAL AMOUNT	$

Make checks payable to: American Economic Association.
Must be drawn on a US bank.
Apply online at http://www.aeaweb.org/membership.php

Payments must be made in advance. We accept checks (in US dollars only, with correct coding for processing in US banks) and credit cards; online or by faxing or mailing the application. Please choose one method; it is the Association's policy NOT TO REFUND dues.

The *Journal of Economic Perspectives*: Proposal Guidelines

Considerations for Those Proposing Topics and Papers for *JEP*

Articles appearing in the journal are primarily solicited by the editors and associate editors. However, we do look at all unsolicited material. Due to the volume of submissions received, proposals that do not meet *JEP*'s editorial criteria will receive only a brief reply. Proposals that appear to have *JEP* potential receive more detailed feedback. Historically, about 10–15 percent of the articles appearing in our pages originate as unsolicited proposals.

Philosophy and Style

The *Journal of Economic Perspectives* attempts to fill part of the gap between refereed economics research journals and the popular press, while falling considerably closer to the former than the latter. **The focus of *JEP* articles should be on understanding the central economic ideas of a question, what is fundamentally at issue, why the question is particularly important, what the latest advances are, and what facets remain to be examined.** In every case, articles should argue for the author's point of view, explain how recent theoretical or empirical work has affected that view, and lay out the points of departure from other views.

We hope that most *JEP* articles will offer a kind of intellectual arbitrage that will be useful for every economist. For many, the articles will present insights and issues from a specialty outside the readers' usual field of work. For specialists, the articles will lead to thoughts about the questions underlying their research, which directions have been most productive, and what the key questions are.

Articles in many other economics journals are addressed to the author's peers in a subspecialty; thus, they use tools and terminology of that specialty and presume that readers know the context and general direction of the inquiry. By contrast, **this journal is aimed at all economists, including those not conversant with recent work in the subspecialty of the author.** The goal is to have articles that can be read by 90 percent or more of the AEA membership, as opposed to articles that can only be mastered with abundant time and energy. Articles should be as complex as they need to be, but not more so. Moreover, the necessary complexity should be explained in terms appropriate to an audience presumed to have an understanding of economics generally, but not a specialized knowledge of the author's methods or previous work in this area.

The *Journal of Economic Perspectives* is intended to be scholarly without relying too heavily on mathematical notation or mathematical insights. In some cases, it will be appropriate for an author to offer a mathematical derivation of an economic relationship, but in most cases it will be more important that an author explain why a key formula makes sense and tie it to economic intuition, while leaving the actual derivation to another publication or to an appendix.

JEP does not publish book reviews or literature reviews. Highly mathematical papers, papers exploring issues specific to one non-U.S. country (like the state of agriculture in Ukraine), and papers that address an economic subspecialty in a manner inaccessible to the general AEA membership are not appropriate for the *Journal of Economic Perspectives*. Our stock in trade is original, opinionated perspectives on economic topics that are grounded in frontier scholarship. If you are not familiar with this journal, it is freely available on-line at <http://e-*JEP*.org>.

Guidelines for Preparing *JEP* Proposals

Almost all *JEP* articles begin life as a two- or three-page proposal crafted by the authors. If there is already an existing paper, that paper can be sent to us as a proposal for *JEP*. However, given the low chances that an unsolicited manuscript will be published in *JEP*, no one should write an unsolicited manuscript intended for the pages of *JEP*. **Indeed, we prefer to receive article proposals rather than completed manuscripts.** The following features of a proposal seek to make the initial review process as productive as possible while minimizing the time burden on prospective authors:

- Outlines should begin with a paragraph or two that precisely states the main thesis of the paper.
- After that overview, an explicit outline structure (I., II., III.) is appreciated.
- The outline should lay out the expository or factual components of the paper and indicate what evidence, models, historical examples, and so on will be used to support the main points of the paper. The more specific this information, the better.
- The outline should provide a conclusion
- Figures or tables that support the article's main points are often extremely helpful.
- The specifics of fonts, formatting, margins, and so forth do not matter at the proposal stage. (This applies for outlines and unsolicited manuscripts).
- Sample proposals for (subsequently) published *JEP* articles are available on request.
- For proposals and manuscripts whose main purpose is to present an original empirical result, please see the specific guidelines for such papers below.

The proposal provides the editors and authors an opportunity to preview the substance and flow of the article. For proposals that appear promising, the editors provide feedback on the substance, focus, and style of the proposed article. After the editors and author(s) have reached agreement on the shape of the article (which may take one or more iterations), the author(s) are given several months to submit a completed first draft by an agreed date. This draft will receive detailed comments from the editors as well as a full set of suggested edits from *JEP*'s Managing Editor. Articles may undergo more than one round of comment and revision prior to publication.

Readers are also welcome to send e-mails suggesting topics for *JEP* articles and symposia and to propose authors for these topics. If the proposed topic is a good fit for *JEP*, the *JEP* editors will work to solicit paper(s) and author(s).

Correspondence regarding possible future articles for *JEP* may be sent (electronically please) to the assistant editor, Ann Norman, at <anorman@JEPjournal.org>. Papers and paper proposals should be sent as Word or pdf e-mail attachments.

Guidelines for Empirical Papers Submitted to *JEP*

The *JEP* is not primarily an outlet for original, frontier empirical contributions; that's what refereed journals are for! Nevertheless, *JEP* occasionally publishes original empirical analyses that appear uniquely suited to the journal. In considering such proposals, the editors apply the following guidelines (in addition to considering the paper's overall suitability):

1) The paper's main topic and question must not already have found fertile soil in refereed journals. *JEP* can serve as a catalyst or incubator for the refereed literature, but it is not a competitor.
2) In addition to being intriguing, the empirical findings must suggest their own explanations. If the hallmark of a weak field journal paper is the juxtaposition of strong claims with weak evidence, a *JEP* paper presenting new empirical findings will combine strong evidence with weak claims. The empirical findings must be robust and thought provoking, but their interpretation should not be portrayed as the definitive word on their subject.
3) The empirical work must meet high standards of transparency. *JEP* strives to only feature new empirical results that are apparent from a scatter plot or a simple table of means. Although *JEP* papers can occasionally include regressions, the main empirical inferences should not be regression-dependent. Findings that are not almost immediately self-evident in tabular or graphic form probably belong in a conventional refereed journal rather than in *JEP*.

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.

The Journal of
Economic Perspectives

Spring 2014, Volume 28, Number 2

Symposia

Big Data

Global Supply Chains

Articles

Features

AMERICAN ECONOMIC ASSOCIATION