The Journal of Economic Perspectives

A journal of the
American Economic Association

Celebrating 30 Years

Summer 2016

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.
No responsibility for the views expressed by the authors in this journal is assumed by the editors or by the American Economic Association.
THE JOURNAL OF ECONOMIC PERSPECTIVES (ISSN 0895-3309), Summer 2016, Vol. 30, No. 3. The *JEP* is published quarterly (February, May, August, November) by the American Economic Association, 2014 Broadway, Suite 305, Nashville, TN 37203-2418. Annual dues for regular membership are $20.00, $30.00, or $40.00 depending on income; for an additional $15.00, you can receive this journal in print. E-reader versions are free. For details and further information on the AEA go to https://www.aeaweb.org/.
Periodicals postage paid at Nashville, TN, and at additional mailing offices.
POSTMASTER: Send address changes to the *Journal of Economic Perspectives*, 2014 Broadway, Suite 305, Nashville, TN 37203. Printed in the U.S.A.

The Journal of
Economic Perspectives

Contents

Volume 30 • Number 3 • Summer 2016

Symposia

Schools and Accountability

Motivated Beliefs

NSF Funding for Economists

Articles

Features

Statement of Purpose

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

Policy on Data Availability

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

Policy on Disclosure

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

Journal of Economic Perspectives

Journal of Economic Perspectives—Volume 30, Number 3—Summer 2016—Pages 3–32

The Importance of School Systems: Evidence from International Differences in Student Achievement

Ludger Woessmann

Average achievement levels of students differ markedly across countries. On the most recent international achievement tests in math and science, the average 15 year-old student in Singapore, Hong Kong, Korea, Japan, and Taiwan is more than half a standard deviation ahead of the average student of the same age in the United States (Hanushek and Woessmann 2015b). Following the rule of thumb that average student learning in a year is equal to about one-quarter to one-third of a standard deviation, these differences are roughly equivalent to what students learn during 1.5–2 years of schooling. Similarly, the average student in Finland and Estonia is 40 percent of a standard deviation ahead of the United States, and the average Canadian student is about one-third of a standard deviation ahead. On the other hand, the average student in Peru and Indonesia is more than 1.1 standard deviations behind the United States, and achievement in Ghana, South Africa, and Honduras lags more than 1.5 standard deviations behind the United States. Overall, average achievement levels among 15 year-olds between the top- and bottom-performing countries easily differ by more than two standard deviations, or the equivalent of 6–8 years of learning.

We will present evidence that the considerable differences in student achievement across countries are systematically related to differences in the organization and governance of school systems. For example, students in many high-performing countries such as Korea and Finland, as well as in some provinces of Canada, face external exit exams at the end of high school. Most schools in Hong Kong and the United Kingdom

■ *Ludger Woessmann is Director of the Center for the Economics of Education, ifo Institute for Economic Research, and Professor of Economics, University of Munich, both in Munich, Germany. His email address is woessmann@ifo.de.*

† For supplementary materials such as appendices, datasets, and author disclosure statements, see the article page at
http://dx.doi.org/10.1257/jep.30.3.3

doi=10.1257/jep.30.3.3

have considerable autonomy in deciding which courses to offer and which teachers to hire, whereas virtually no schools have this autonomy in Greece. More than half the students in the Netherlands, Belgium, Ireland, and Korea attend privately operated schools, while hardly any students in Norway and Poland do so. Students in Austria and Germany are tracked into different-ability schools at age 10, while two-thirds of OECD countries have comprehensive school systems at least until age 15.

Educational institutions such as national accountability systems or tracking regimes often vary only slightly or not at all within countries, but an international perspective provides an opportunity for comparisons. As the director of the first pilot project comparing student achievement across countries remarked, "If custom and law define what is educationally allowable within a nation, the educational systems beyond one's national boundaries suggest what is educationally possible" (Foshay 1962). Studies done within or across schools of a particular country can also suffer from the risk of selection bias, if students with specific backgrounds are more likely to attend certain schools or to become involved in certain programs, while studying national-level variation circumvents some of these selection biases.

However, these advantages of the cross-country comparative approach come with some built-in limitations. Identification of causal effects raises particular challenges in an international setting. Countries may differ in a variety of hard-to-observe ways such as cultural traits, valuation of achievement, and other preferences that are associated with both institutional choices and achievement levels. Such unobserved country heterogeneity gives rise to omitted variable bias in cross-country analyses. Moreover, only a limited number of country-level observations are available in the test data.

This essay first describes the size and cross-test consistency of international differences in student achievement. It then uses the framework of an education production function to describe how different factors of the school system, as well as factors beyond the school system, are associated with cross-country achievement differences. The discussion then focuses on research that attempts to go beyond conditional correlations by addressing some sources of potential bias in cross-country analysis. This discussion suggests that the role of resource inputs seems limited; indeed, all of the above-mentioned high-achieving countries spend considerably *less* on schools per student than the United States (OECD 2013). But differences in instruction time and teacher quality do matter. In addition, institutional features including external exams, school autonomy, private competition, and tracking affect the level and distribution of student achievement across countries and account for a substantial part of the cross-country achievement variation. The conclusion points out some major implications of educational achievement for the prosperity of individuals and nations.

How Large and Consistent Are International Differences in Student Achievement?

Large-scale international testing of student achievement has more than half a century of history, and many studies provide evidence on international differences in student achievement and how they have evolved over time.

International Rankings and the Size of Cross-Country Differences

A crucial role in the emergence and continuation of comparative testing has been played by the International Association for the Evaluation of Educational Achievement (IEA), an independent cooperative of national research institutions and government agencies (IEA 2016; Mullis, Martin, Foy, and Arora 2012). Following a pilot project in 1959–61, the IEA conducted its first international math study of eleven countries in 1964. The first science and reading studies occurred with 12–16 countries in the early 1970s, and a second round in each subject was performed in the 1980s and early 1990s. Since 1995, the Trends in International Mathematics and Science Study (TIMSS) has tested math and science achievement mostly in fourth and eighth grade every four years in between 38 and 52 voluntarily participating countries. In addition, the Progress in International Reading Literacy Study (PIRLS) has tested fourth-grade reading achievement every five years since 2001, with 48 countries participating in the most recent wave.

In 2000, the Organisation for Economic Co-operation and Development (OECD) entered international testing as a second major player. Since then, its Programme for International Student Assessment (PISA) tests representative samples of 15 year-old students in math, science, and reading every three years. In both 2009 and 2012, 65 countries participated, and 71 countries have signed up to participate in the most recent PISA installment in 2015.[1]

All these tests draw random samples of students to ensure representativeness for the national target populations. In particular, the three ongoing studies have a two-stage sampling design. At a first stage, they draw a random sample of schools in each country. Within those schools, they then randomly draw one classroom per grade (TIMSS, PIRLS) or a random sample of 15 year-old students (PISA), respectively. Each of these tests uses a common set of questions in all participating countries based on a particular effort to achieve cross-country comparability. PISA, TIMSS, and PIRLS each link their own tests over time, too. But there is no direct link between the scales of the three testing regimes or across time with the older tests.

Table 1 shows the performance of the 81 countries that have participated in the most recent installments of the PISA (2012) and TIMSS (2011) international math and science tests. Achievement is expressed on the PISA scale, which is standardized to have a mean of 500 and a standard deviation of 100 among all students in OECD countries. This standardization was done in 2003 in math and in 2006 in science. On average across OECD countries, the actual within-country standard deviation is 92 in math and 93 in science (OECD 2013). The transformation of the TIMSS 2011 data to the PISA scale follows the method suggested in Hanushek and Woessmann (2015b, Annex B).

[1] In addition, there are a couple of separate international tests whose items are aligned to the US school curriculum (which may limit international comparability), a number of regional tests in Latin America and sub-Saharan Africa, and adult literacy tests (for discussion, see Hanushek and Woessmann 2011a, table 2; Hanushek and Woessmann 2015b, chapter 4; Hanushek, Schwerdt, Wiederhold, and Woessmann 2015). The International Association for the Evaluation of Educational Achievement has also conducted studies in other subjects such as foreign languages, civic education, and computer literacy.

Table 1
Performance on Recent International Student Achievement Tests, 2011–2012

Country	*Score*	*Country*	*Score*	*Country*	*Score*
Shanghai–China	596	Norway	492	Malaysia	420
Singapore	562	Luxembourg	491	Costa Rica	418
Hong Kong–China	558	Spain	490	Mexico	414
Korea	546	Italy	489	Uruguay	413
Japan	542	**United States**	**489**	Montenegro	410
Chinese Taipei	542	Portugal	488	Bahrain	408
Finland	532	Lithuania	487	Lebanon	403
Estonia	531	Hungary	486	Georgia	401
Liechtenstein	530	Iceland	485	Brazil	398
Macao–China	529	Russian Federation	484	Jordan	397
Switzerland	523	Sweden	482	Argentina	397
Netherlands	523	Croatia	481	Albania	396
Canada	522	Slovak Republic	476	Tunisia	393
Poland	522	Ukraine	468	Macedonia	392
Vietnam	520	Israel	468	Saudi Arabia	391
Germany	519	Greece	460	Palestine	388
Australia	513	Turkey	456	Colombia	388
Ireland	512	Serbia	447	Qatar	380
Belgium	510	Bulgaria	443	Syria	379
New Zealand	508	Romania	442	Indonesia	379
Slovenia	508	United Arab Emirates	441	Botswana	376
Austria	506	Cyprus	439	Peru	371
United Kingdom	504	Thailand	435	Oman	369
Czech Republic	504	Chile	434	Morocco	348
Denmark	499	Kazakhstan	428	Honduras	328
France	497	Armenia	428	South Africa	315
Latvia	496	Iran	422	Ghana	291

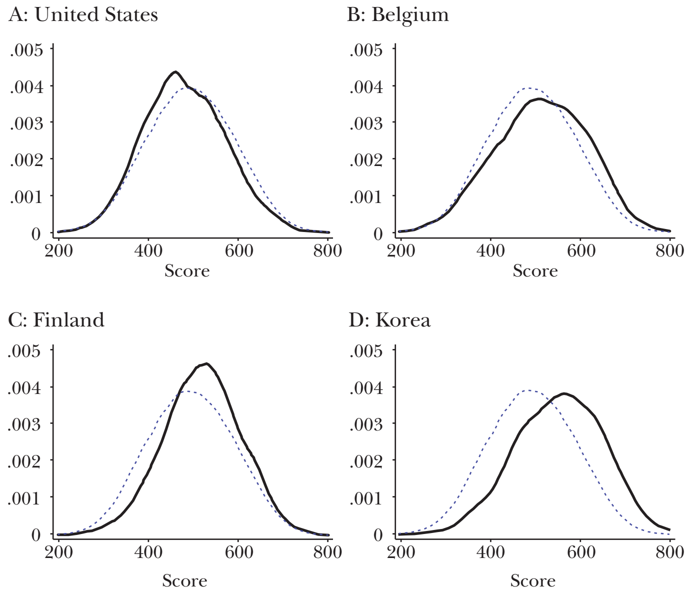
Notes: The table gives the average of the scores on international math and science tests. Black: PISA 2012, 15-year-olds. Grey: TIMSS 2011, 8th grade, transformed to PISA scale as in Hanushek and Woessmann (2015b).

The cross-country differences in knowledge among same-aged students are in some cases extremely large. Remember, as a rule of thumb, learning gains on most national and international tests during one year are equal to between one-quarter and one-third of a standard deviation, which is 25–30 points on the PISA scale. Thus, the achievement difference between the average 15 year-old in the United States and in the PISA top performers—Singapore, Hong Kong, Korea, Japan, Taiwan, Finland, and Estonia—is roughly twice what students usually learn during one year. At the other end, the average difference of US achievement to the PISA bottom performers (Peru and Indonesia) amounts to the equivalent of three to four years of learning, and to five to six years to the TIMSS bottom performers (Ghana and South Africa).

In looking at lists like Table 1, it is important to focus on scores, not just ranks. For example, in the PISA 2012 math test, the achievement levels of most countries are not statistically significantly different from their closest 1–3 neighbors above and below. Where the scores are closely bunched, Portugal's achievement at rank 31 does not differ significantly from ranks 25–37 in the PISA 2012 math test (OECD 2013).

Figure 1

Distribution of Student Achievement in Selected Countries on PISA Math Test, Compared to All OECD Countries



Notes: Kernel densities of student achievement on the PISA 2012 math test. Bold solid line: specified country. Dotted line: OECD countries.

The means in Table 1 may hide important differences in the shape of the overall distribution of achievement in a country. Figure 1 displays the achievement distribution on the PISA 2012 math test for the United States and three selected countries with relatively high performance, comparing each to the overall distribution in OECD countries. The US distribution is shifted to the left and slightly more left-steep compared to the OCED distribution, but it does not have a particularly strong left or right tail. As the three example countries show, it is possible to achieve above-average mean performance with a relatively equitable distribution (Finland), with a distribution that is mostly just shifted to the right of the OECD distribution (Korea), or with a relatively unequal distribution (Belgium).

The relatively low performance of the United States compared to many OECD countries cannot be attributed to the particularly poor performance of a small group of students or of students from disadvantaged backgrounds. For example, the 25th, 50th, and 75th percentiles of the US distribution on the PISA 2012 math test are all between 13 and 15 points below the OECD average of the respective percentiles. In Hanushek, Peterson, and Woessmann (2013), my coathors and I document that both the proportion of students who achieve at a basic proficient level and the proportion of students who achieve at an advanced level in the United States are comparatively low in an international perspective. In addition, in Hanushek, Peterson, and Woessmann (2014), we show that the ranking of US students from better-educated families when compared to students from better-educated families in other countries is not much different from the ranking of US students from less-well-educated families when compared to students from less-well-educated families in other countries.

Consistency across Different Tests

The measurement of educational achievement is subject to many psychometric and measurement choices. For example, the target population of the TIMSS test is eighth graders. Also, TIMSS has a strong curricular focus and is based on an assessment framework developed in a collaborative process with participating countries, with a test-curriculum matching analysis describing how the test matches each participating country's school curriculum. On the other hand, the target population of the PISA test is 15 year-olds, and PISA aims to assess the knowledge and skills essential for full participation in modern society, including the extrapolation and application of learned knowledge to new real-life situations.

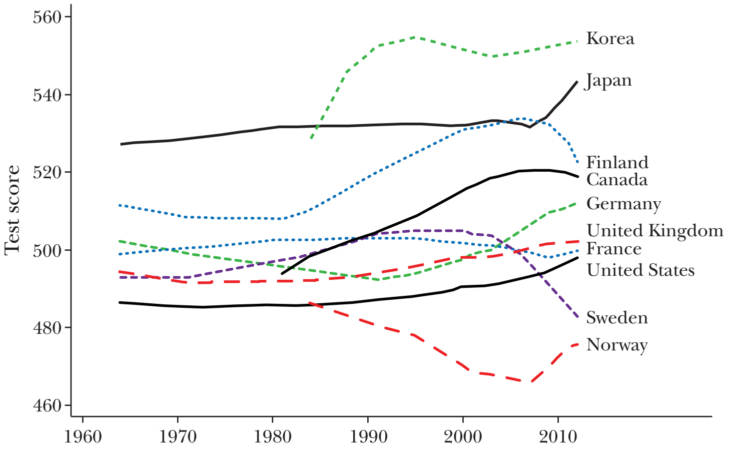
How sensitive are international comparisons to specific measurement choices? We can compare the achievement of the 28 countries that participated in the most recent installments of both tests: PISA 2012 and TIMSS 2011. Despite the differences in timing, target populations, and conceptual approaches, the correlation across the 28 countries participating in both tests is 0.944 in math and 0.930 in science (Hanushek and Woessmann 2015b). This high correlation suggests that when it comes to international comparisons, specific test designs are of secondary importance.

Another potential issue with international achievement tests is cross-country differences in sample selectivity due to different rates of enrollment, exclusion, and nonresponse. While sampling was devised to be representative of the student population in each participating country, some countries do not have universal enrollment at age 15, when students are tested in PISA. In addition, nonrandom differences in patterns of sample exclusions (for example, for handicapped children) and nonresponse can compromise comparability across countries. However, the working paper version of Hanushek and Woessmann (2011c) shows that although these factors are related to average country scores, controlling for these rates does not affect the qualitative results on institutional effects in international education production functions presented later in this paper. The variation in the extent to which countries adequately sample their entire student populations appears orthogonal to the associations analyzed here.

Changes over Time

While an assessment of countries at a point in time is reasonably straightforward, assessing changes in country performance over time is harder. The early international tests, in particular, constitute separate testing incidents without links across different tests. In Hanushek and Woessmann (2012, 2015a), we use an empirical calibration method to put all international tests from 1964 to 2003 on a common standardized scale. Our analysis shows that 73 percent of the variance across the 693 separate test observations in 50 countries occurs between countries. The remaining 27 percent combines true changes over time in countries' scores and any measurement error in the testing. That is, most of the variation in the available panel data of countries over time is across rather than within countries, implying that a large share of the country differences are consistent over time.

Figure 2
Long-Run Test Score Trends in Selected Countries, 1964–2012



Source: Extended from Hanushek and Woessmann (2015a).
Notes: Stylized depiction of standardized data from international tests 1964–2012. The figure is based on age-group- and subject-specific standardized scores from all international tests in 1964–2003 extended with the subsequently available TIMSS, PIRLS, and PISA data to 2012. It takes out age-group- and subject-specific trends in each country, smooths available test observations with locally weighted regressions, and linearly interpolates between available test observations; see Hanushek and Woessmann (2015a) for details.

Still, several countries do show either significant improvements or declines over time. Figure 2 depicts achievement trends observed in selected example countries from 1964 to 2012. The more limited variation in early decades likely reflects the lower frequency of testing before 2000. The figure shows substantial cross-sectional differences across countries. But some countries do show noteworthy changes over time. Ripley (2013) acknowledges that a previous version of this figure motivated her work on the widely acclaimed New York Times bestseller *The Smartest Kids In The World—And How They Got That Way*. While the United States was rather typical compared to most other countries, she wrote there were a few countries where "virtually *all* kids were learning critical thinking skills in math, science, and reading" (p. 4). While some countries such as Canada and Finland over the 1980s and 1990s and Germany and Japan more recently did improve substantially over time, other well-off countries deteriorated, such as Norway during the 1990s, Sweden during the 2000s, and Finland in recent years. Educational achievement levels of countries seem generally consistent over time, but they are not set in stone and can be mutable.

Descriptive Patterns Using an Education Production Function

This section uses the framework of an international education production function to document the extent to which, on a purely descriptive basis, differences in family background, school resources, and institutions can account for cross-country differences in student achievement. These inputs are probably not exogenous to student achievement, so correlations between the inputs and test scores are very likely to be biased by omitted variables, selection, and reverse causation. While these descriptive patterns must be interpreted cautiously, they can serve as a useful guide to the more explicit discussions of causality that follow.

International Education Production Functions

An education production function models the output of education as a function of different inputs (for example, Hanushek 1986, 2002). We combine the input factors into three groups: family background factors, school resources, and institutional structures of school systems. The first group is mostly outside the control of school systems. The other two groups of factors reflect the quantity of resource inputs in the systems and the institutional structures. The basic model can be extended to include interactions between input factors.

A substantial literature has estimated such international education production functions using cross-sectional data (for an extensive review, see Hanushek and Woessmann 2011a). Early studies used aggregate country-level data to study the country-level variation in achievement scores (for example, Bishop 1997; Hanushek and Kimko 2000; Lee and Barro 2001). More recent studies also use country-level data to study, for example, the correlates of gender equality in achievement (Guiso, Monte, Sapienza, and Zingales 2008; Fryer and Levitt 2010).

However, starting with Woessmann (2003b), a number of studies have used the data from international achievement tests at the student level to estimate cross-country education production functions. Examples include Woessmann (2005b), Fuchs and Woessmann (2007), Brunello and Checchi (2007), Woessmann, Luedemann, Schuetz, and West (2009), Schneeweis (2011), and Ammermueller (2013). Because these studies use data on individual students, they can hold constant a large set of observable factors usually unavailable in national datasets. In effect, they can compare "observationally equivalent" students across countries.

For concreteness, Table 2 provides an example of a basic cross-sectional estimation of an international education production function.[2] The table shows the categories of data that are available. The dependent variable is the score from the PISA 2003 math test, with the sample restricted to the 29 participating OECD countries to provide greater comparability. The model includes a large number of explanatory variables in the three groups of input factors: family background, school resources, and institutions. The individual-level measures of family background

[2] This is a simplified version of the model used in Woessmann et al. (2009) and Hanushek and Woessmann (2011a). To allow for a more meaningful accounting analysis below, it drops the GDP per capita of the country (which is correlated with educational spending at 0.93 and yields a counterintuitive negative estimate), class size (which has a counterintuitive positive estimate), and the imputation dummies and their interactions with the main variables contained in those models. Qualitative results are similar with those variables included. Qualitative results are also unaffected when adding the country-average value of the Index of Economic, Social and Cultural Status (ESCS), the average share of students with an immigrant background in a country, or continental fixed effects to the model. Country-average ESCS in fact enters marginally significantly negatively and the migrant share insignificantly. Reported standard errors are clustered at the country level, which may be overly conservative for variables that vary at the school or student level.

are taken from student background questionnaires that students complete in the PISA study; the measures of school resources and institutions are mostly taken from school background questionnaires that the principals of participating schools complete; these measures are combined with country-level data on expenditure per student and external exit exams that come from outside sources (for details, see Appendix A of Woessmann et al. 2009). Descriptively, this model accounts for 34 percent of the achievement variance at the individual student level.

Factors beyond the School System: Family, Socioeconomic, and Cultural Background

Some of the personal characteristics that have meaningful and statistically significant magnitudes in Table 2 include student characteristics such as age, gender, and participation in early childhood education, along with indicators for family status, parental education, parental work status and occupation, the number of books at home, immigration background, and the language spoken at home. For example, the achievement difference between students in the highest category of more than 200 books at home versus the lowest category of fewer than 10 books at home—a proxy for aspects of educational, social, and economic background—amounts to more than half a standard deviation in the PISA test score.

There are two main types of analysis in the literature analyzing socioeconomic backgrounds in the international tests. The first type looks at how much socioeconomic background contributes to country-level differences in educational outcomes. The second type of analysis compares the within-country association of socioeconomic factors with student achievement, sometimes referred to as socioeconomic gradients, across countries. For example, in Schütz, Ursprung, and Woessmann (2008), we estimate the associations of family background with student achievement—interpreted as measures of the inequality of educational opportunity—in different countries using TIMSS data and relate them to measures of institutions of the school systems. We show that family background effects are systematically larger in countries with early tracking and less-extensive pre-primary education systems.[3] Jerrim and Micklewright (2014) use PISA and PIRLS data to analyze the extent to which cross-country comparisons of socioconomic gradients are affected by differences in reporting errors.

Several studies have focused on the achievement of children with an immigration background, looking at both socioeconomic and institutional characteristics. For example, Dustmann, Frattini, and Lanzara (2012) show that in many countries, observed differences in parental background (including parental education and occupation and the language spoken at home) can account for much of the lower PISA achievement of children of immigrants compared to native children. They also find that children of Turkish immigrants perform better in most host countries than Turkish children in Turkey. Also using PISA data, Cobb-Clark, Sinning, and Stillman (2012) show that the migrant–native achievement gap is significantly associated with institutional features of the host countries such as school starting age, ability tracking, private

[3] Applying a similar approach to outcomes beyond school age, Brunello and Checchi (2007) find that early tracking is related to larger effects of family background on educational attainment and earnings in the labor market, but not on on-the-job training and adult literacy.

Table 2

A Simple International Education Production Function: A Least-Squares Regression

(dependent variable is student's mathematics test score)

	Coefficient	*Standard error*
Family Background		
Age (years)	17.825***	(3.160)
Female	–14.733***	(1.639)
Preprimary education (more than 1 year)	6.832***	(2.428)
School starting age	–3.869*	(2.030)
Grade repetition in primary school	–54.579***	(4.734)
Grade repetition in secondary school	–33.726***	(6.702)
Grade		
7th grade	–47.003***	(10.051)
8th grade	–19.213*	(10.242)
9th grade	–6.772	(6.896)
11th grade	–3.275	(5.236)
12th grade	11.949*	(6.398)
Living with		
Single mother or father	20.045***	(3.949)
Patchwork family	22.678***	(4.286)
Both parents	29.524***	(3.956)
Parents' working status		
Both full-time	–2.071	(2.911)
One full-time, one half-time	8.820***	(2.327)
At least one full time	15.926***	(2.891)
At least one half time	10.531***	(2.278)
Parents' job		
Blue collar, high skilled	1.481	(2.365)
White collar, low skilled	3.743*	(1.870)
White collar, high skilled	8.189**	(3.144)
Books at home		
11–25 books	6.760***	(2.290)
26–100 books	24.749***	(2.789)
101–200 books	34.232***	(3.161)
201–500 books	54.400***	(3.238)
More than 500 books	54.166***	(3.703)
Immigration background		
First-generation student	–11.447**	(4.442)
Nonnative student	–13.776**	(5.375)
Language spoken at home		
Other national dialect or language	–17.689**	(7.064)
Foreign language	–7.887***	(2.677)
Index of Economic, Social and Cultural Status (ESCS)	19.926***	(2.153)
Community location[s]		
Town (3,000–100,000)	9.101**	(3.323)
City (100,000–1,000,000)	16.951***	(3.989)
Large city with > 1 million people	13.939***	(4.929)

Continued on next page

Table 2 (continued)

	Coefficient	*Standard error*
School Resources		
Cumulative educational expenditure per student (1,000 $)[c]	0.270**	(0.103)
Shortage of instructional materials[s]		
Large shortage	–8.737**	(3.514)
No shortage	8.678***	(2.015)
Instruction time (minutes per week)	0.044***	(0.015)
Teacher education (share at school)[s]		
Fully certified teachers	7.699	(8.588)
Tertiary degree in pedagogy	10.211	(6.547)
Institutions		
Competition[c]		
Private operation (country share)	56.941***	(9.758)
Government funding (country share)	57.847***	(19.486)
Accountability		
External exit exams[c]	9.433	(9.055)
Assessments used for student retention/promotion[s]	11.744**	(4.320)
Monitoring of teacher lessons by principal[s]	6.785*	(3.442)
Monitoring of teacher lessons by external inspectors[s]	4.842*	(2.816)
Assessments used to compare school to district/nation[s]	4.188	(2.870)
Assessments used to group students[s]	–8.261**	(3.021)
Autonomy and its interaction with external exit exams[s]		
Autonomy in establishing starting salaries	–15.769***	(5.229)
External exit exams × Autonomy in establishing starting salaries	14.550*	(8.104)
Autonomy in formulating budget	–9.624	(6.901)
External exit exams × Autonomy in formulating budget	7.882	(8.478)
Autonomy in determining course content	–2.053	(5.435)
External exit exams × Autonomy in determining course content	11.504	(7.262)
Autonomy in hiring teachers	18.349*	(10.436)
External exit exams × Autonomy in hiring teachers	–24.723**	(11.796)
Constant	116.126**	(51.774)
Students	219,794	
Schools	8,245	
Countries	29	
R^2 (at student level)	0.340	

Source: Own calculations on the basis of Woessmann et al. (2009) using data from the Programme for International Student Assessment (PISA) 2003; the sample is OECD countries.

Notes: The table presents results from a least-squares regression weighted by students' sampling probability. The dependent variable is student's mathematics test score. Measures vary at the student level unless noted otherwise. Robust standard errors adjusted for clustering at the country level in parentheses.

[s] Observed at school level.

[c] Observed at country level.

***, **, and * represent significance levels of 1, 5, and 10 percent, respectively.

schools, and teacher evaluation in a cross-sectional model. In a country-level analysis of the PISA data, Brunello and Rocco (2013) find that an increased share of immigrant students has a small negative effect on the achievement level of native students.

Overall, socioeconomic factors contribute substantially to the cross-country variation in test scores.[4] These factors, however, are largely outside the influence of school systems—although not necessarily beyond the effects of other family, social, and redistributive policies.

Factors of the School System: Inputs and Institutions

Measures of school resources often fail to achieve economic and statistical significance in international education production functions, and sometimes even show counterintuitive coefficients. In Table 2, the point estimate on school spending is very small: An increase in cumulative educational expenditure per student until age 15 by $25,000, or one standard deviation, is associated with an increase in student achievement of less than 7 percent of a standard deviation. If class size as observed at the individual student level is added to the model, it has a counterintuitive positive coefficient—purportedly indicating that students achieve at higher levels in larger classes. Other variables have a more intuitive interpretation: for example, students perform worse in schools whose principal reports that the school's capacity to provide instruction is hindered by a shortage or inadequacy of instructional materials such as textbooks. Both weekly instruction time and measures of teacher education are positively associated with student achievement. Evidence from TIMSS, which provides more detailed teacher information from individual teacher background questionnaires, shows similar results (Woessmann 2003b). To the extent that schools with more resources in the tested grade also tended to have more resources in earlier grades, the coefficient estimates on resources capture not just the contemporaneous effect of resources in the specific grade, but the cumulative effect of resources over the previous grades.

In contrast, institutional features of school systems are strongly associated with student achievement in studies of this sort. Table 2 offers some examples, as do Woessmann (2003b, 2005b), Fuchs and Woessmann (2007), and Woessmann et al. (2009). In particular, measures of the extent of private school operation, government funding of schools, and different features of school accountability such as external exit exams, the use of assessments, and monitoring of lessons are positively related to student outcomes.[5] In addition, there is a tendency for school autonomy in different decision-making areas to be negatively related to student achievement in systems without external exit exams but to be unrelated or positively related in systems where external exit exams promote accountability (Woessmann 2005b). In

[4] Additional factors analyzed with international achievement data include gender differences (for example, Guiso et al. 2008; Fryer and Levitt 2010), relative age at school entry (for example, Bedard and Dhuey 2006), and peer effects (for example, Ammermueller and Pischke 2009).

[5] External exit exams reach statistical significance in a specification of the model of Table 2 that excludes the interactions with school autonomy. Results on the country-level variables in Table 2 are qualitatively the same in a two-step specification that first estimates Table 2 with country fixed effects and then regresses the coefficients captured on these fixed effects on the country-level variables.

a study of a variable not included in Table 2, Edwards and Garcia Marin (2015) find no significant association of country-aggregate student achievement in the PISA test with whether the right to education is included in a country's political constitution.

The results on instruction time, teacher education, and institutional effects provide a *prima facie* case for the relevance of school systems. Another piece of evidence for this relevance arises from adding school fixed effects to the estimation of an international education production function. Using PISA data, Freeman and Viarengo (2014) show that estimated school fixed effects are associated with observable school policies and teaching practices as well as with socioeconomic gradients. While they do not rule out nonrandom selection into schools as playing a role here, they interpret these results as indications of the potential importance of what schools do, as opposed to national or individual traits.

While most of the international achievement datasets are cross-sectional, Singh (2015) uses a longitudinal dataset that observes individual students at ages 5 and 8 in four developing countries. The findings show that the large cross-country learning gaps between low-performing Peru and high-performing Vietnam (apparent earlier in Table 1) are virtually nonexistent at school-entry age. They emerge over the first few school years in a way that is most consistent with large cross-country differences in the productivity of a school year (estimated from discontinuities in completed grades emerging from birth months in combination with enrollment thresholds), rather than with observed differences in socioeconomic background and time use. Again, these findings suggest that school systems have important effects.

Accounting for the Cross-Country Variation in Test Scores

As indicated, the model in Table 2 accounts for about one-third of the total student-level variation in the international model. This variation includes within-country variation as well as cross-country variation. The former is likely to include a component of random measurement error because of idiosyncrasies in individual performance on the testing day, a component that would cancel out at the national level.

So to what extent can family background factors, school resources, and institutions account for differences in student achievement across countries? To answer this question, we have to combine the large number of explanatory variables into a smaller number of factors. The student-level estimation of Table 2 provides one coefficient per variable: that is, it effectively forces the between-country associations of student achievement with the input factors to be the same as the within-country associations. We use these coefficient estimates on the individual variables in the model of Table 2 to combine the family background variables into one factor. That is, we simply calculate a linear combination that is the sum of the products of the individual variables times their respective coefficient estimates. We do the same for the school resource variables and the institutional variables. We then collapse the three combined input factors to the level of the 29 OECD country observations to obtain three aggregate country-level variables.

For descriptive purposes, we regress aggregate academic achievement on these three composite inputs for the 29 country-level observations. The share of the cross-country variance in achievement accounted for by the three input factors is 83 percent. That is, using the student-level model to additively and linearly combine

Table 3
Accounting for the Achievement Variance at the Country Level

	Family background	*School resources*	*Institutions*	*All three factors*
Accounted variance when only this factor is included in the model	0.504	0.181	0.533	0.834
Change in accounted variance when this factor is added to a model that already includes the other two factors	0.208	0.045	0.259	

Source: Author using data from the PISA 2003.
Notes: The table shows the share of the country-level variance in PISA 2003 mathematics test scores accounted for by the respective factor. Each factor represents a linear combination of individual variables using coefficient estimates from the student-level regression shown in Table 2, collapsed to the country level.

the input variables into three factors that can be collapsed to the country level, our simple international education production function descriptively accounts for more than four-fifths of the total cross-country variation in student achievement.

Table 3 breaks this explained variance in the country-level model down into components accounted for by the three groups of input factors. As in any regression analysis, the contribution of each factor depends on the other variables in the model. However, the role of family background factors appears substantial, contributing between 21 and 50 percent to the total cross-country variance in student achievement. By contrast, the contribution of school resources is much smaller, at 4 to 18 percent. Institutional differences again contribute importantly to the cross-country achievement variation, at 26 to 53 percent.[6]

Details of the extent to which the simple model accounts for the achievement of individual countries are shown in Table 4. For each country, the table shows how much of the country's difference from the international mean can be accounted for by each set of input factors.[7] For 14 of the 29 countries, the unaccounted-for residual achievement is less than 10 percent of a standard deviation. But for some examples, the model does not perform very well: in top-performing Finland, only 12.9 of the 44.5 percentage points of superior achievement (in standard deviations) are accounted for by the model. Differences from the international mean in family

[6] Compared to the models in Woessmann et al. (2009) and Hanushek and Woessmann (2011a), the model here excludes GDP per capita and class size, whose counterintuitive coefficients would hamper the interpretation of the accounting analysis. Including them would, in fact, reduce the separate contributions accounted for by the family background and school resource factors at the country level. Results are similar when including the imputation dummies contained in those models. It is debatable whether the model should include grade levels, individual grade repetition, and school starting age; however, results are similar when excluding these variables. The family background factor includes both individual student characteristics and genuine family factors; when separating the two, most of the country-level contribution goes to the genuine family factors and little to the student characteristics.

[7] To estimate the contribution of each input factor, we first run the country-level model on demeaned variables and then multiply the respective coefficient estimates with each country's value of the respective input factor. The contributions of the three input factors then sum to the predicted value (shown as "accounted difference" in Table 4) in this model.

Table 4

Accounting for Each Country's Difference from the International Mean

				Of which: accounted for by		
	Observed difference (1)	*Unaccounted difference* (2)	*Accounted difference* (3)	*Family background* (4)	*School resources* (5)	*Institutions* (6)
Finland	44.5	31.7	12.9	2.7	–1.3	11.5
Korea	42.0	14.3	27.7	13.0	5.6	9.1
Netherlands	38.4	–8.0	46.4	–3.4	–0.3	50.1
Japan	34.0	4.4	29.6	17.5	2.9	9.2
Canada	33.0	17.4	15.6	15.9	3.2	–3.5
Belgium	29.5	–11.8	41.3	–1.2	1.4	41.0
Switzerland	26.5	27.3	–0.8	–13.2	9.5	2.9
Australia	24.5	2.1	22.4	14.0	6.6	1.7
New Zealand	24.5	17.8	6.7	16.2	–3.0	–6.4
Czech Republic	16.4	2.1	14.3	16.1	–9.0	7.2
Iceland	15.1	–11.6	26.7	29.7	4.9	–7.9
Denmark	14.1	6.0	8.1	0.4	6.5	1.2
Sweden	10.0	5.5	4.5	5.9	–1.0	–0.4
United Kingdom	8.4	–9.1	17.5	13.0	2.7	1.8
Austria	5.5	5.7	–0.2	2.1	6.1	–8.5
Ireland	3.9	–15.0	18.8	–3.3	1.6	20.5
Germany	3.5	5.4	–1.9	–4.0	–0.8	2.8
Slovak Republic	–1.0	6.3	–7.3	4.2	–18.0	6.5
Norway	–4.3	–26.4	22.1	22.1	2.1	–2.1
Luxembourg	–6.3	–10.7	4.4	–25.5	19.3	10.6
Hungary	–9.3	–18.7	9.4	4.5	–5.4	10.4
Poland	–9.5	2.5	–12.0	–11.5	–8.1	7.6
Spain	–14.1	–2.7	–11.4	–4.8	–5.4	–1.2
United States	–16.1	–14.7	–1.4	2.3	9.1	–12.9
Portugal	–33.5	23.0	–56.5	–27.0	–2.8	–26.7
Italy	–33.9	–5.5	–28.3	2.7	3.6	–34.7
Greece	–55.1	–22.1	–33.0	–4.1	–3.0	–26.0
Turkey	–75.8	–4.4	–71.5	–31.7	–17.5	–22.3
Mexico	–114.8	–10.6	–104.2	–52.7	–9.9	–41.6

Notes: Each entry shows the country's test score difference from the international mean on the PISA 2003 mathematics test, expressed in student-level standard deviations. Column 1: actual difference. Column 2: difference not accounted for by a country-level regression of the actual test score difference on the three combined input factors (family background, school resources, institutions), each of which is measured as a linear combination of individual variables using coefficient estimates from the student-level regression of Table 2, collapsed to the country level. Column 3: difference accounted for by this country-level regression. Columns 4–6: difference accounted for by family background, school resources, and institutions, respectively. By constructions, columns 2 and 3 sum to column 1, and columns 4–6 sum to column 3.

background and school resources hardly contribute to this, but 11.5 percentage points are contributed by differences in the institutional setting, which in Finland include the existence of external exams, almost universal use of assessments for student retention, and widespread school autonomy over course content. For Korea, about two-thirds of the high relative achievement is accounted for by the model, and all three groups of input factors contribute to this, including a large share of privately operated schools, external exams, widespread monitoring of teacher lessons, and universal course-content autonomy. For third-achieving Netherlands, the model in fact over-predicts its high achievement, and all of this is due to superior institutions—in particular, the largest share of privately operated schools, external exams, widespread course-content autonomy, and use of assessments for retention. At the lower end, most of the poor performance of Mexico and Turkey is accounted for by the model, in particular detrimental family background and institutions. The model does not do well at predicting US performance; institutions such as salary autonomy without external exit exams would predict the lower-than-average achievement level, but better family background and, in particular, abundant school resources would point the other way.

Inputs to the School System: Explorations into Causal Effects

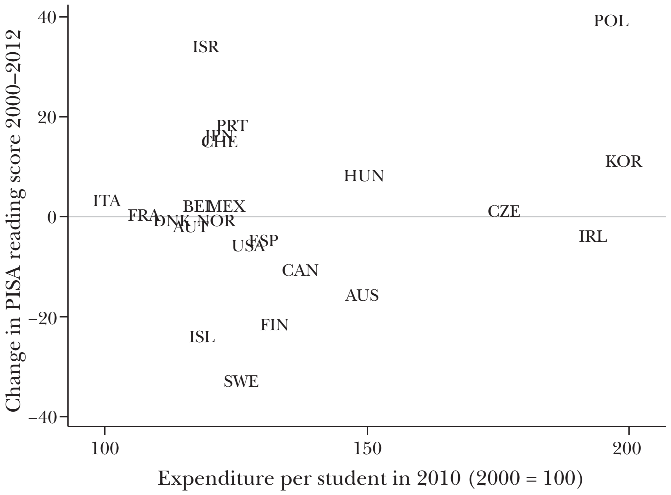
Inputs are clearly not exogenous to the education process. There may be reverse causation, for example, if educational systems assign additional resources to schools that serve low-achieving students, or if schools with poor student outcomes are induced to implement specific reforms. There may be bias from selection in that parents from low-achieving (or high-achieving) students tend to select into schools that offer specific resources for their children, or if high-performing schools have some ability to select high-achieving students. There may be omitted variables correlated with both inputs and outcomes, including country-level factors such as culture and valuation of education that may drive both inputs and learning effort, and also differences in preferences for high-quality education among parents or differences in motivation or ability of students. The direction of the net bias from such factors is not always obvious.

As a straightforward first step to exclude certain sources of bias when analyzing the possible effect of expenditure per student, one can ignore differences in the *levels* of expenditure and only use *changes* in average country expenditure over time as an explanatory variable in first-difference or differences-in-differences panel-type models. To the extent that sources of bias such as countries' cultures and parental background do not change significantly over time, they will no longer bias estimates based on changes in expenditure. In this spirit, in Gundlach, Woessmann, and Gmelin (2001), we calculated changes in expenditure and changes in test performance in several OECD countries over a 25-year period (1970–1994), finding that even substantial increases in real expenditure per student did not go hand in hand with improvements in student achievement.

More recently, the linking of the PISA tests over time allows for a direct comparison of spending changes to changes in achievement. As is directly obvious from Figure 3, changes in PISA performance from 2000 to 2012 are not systematically related to concurrent changes in expenditure per student. Countries with large spending increases do not show different achievement trends from countries that spend only little more.[8] While this analysis may be attenuated by the fact that changes

[8] The coefficient estimate on expenditure in the simple underlying first-differenced regression is insignificant. Similarly, using data from the first three PISA waves, the working-paper version of Hanushek

Figure 3
Changes in Educational Spending and in Student Achievement across Countries



Source: Hanushek and Woessmann (2015a) based on OECD data.
Notes: Scatter plot of the expenditure per student in 2010 relative to 2000 (constant prices, 2000 = 100) against change in PISA reading score, 2000–2012.

in expenditure may take some time to translate into actual inputs and then to affect student outcomes, the 25-year time horizon of the previous analysis should be able to reflect major effects. Of course, if other factors changed in a way correlated with both spending and test outcomes, looking for correlations between them—whether in levels or in differences—would still suffer from bias in these aggregate analyses.

Several studies have sought to use arguably exogenous variation in particular inputs by applying more elaborate identification methods. Here, we will discuss some of the evidence suggesting that smaller class sizes do not make much difference to educational outcomes but that more instruction time and higher teacher quality do make a difference.

Class Size

Most of the efforts that seek to uncover a causal effect of class size on test outcomes using international data turn to within-country variation. For example, in each school, natural cohort fluctuations in enrollment give rise to random class-size variation between adjacent grades (Hoxby 2000). In Woessmann and West (2006), we combine school fixed effects—which seek to eliminate between-school variation—with an approach that uses average class size in the school's grade as an instrumental variable, thus eliminating bias from sorting within a grade in a school. Applying this identification strategy to TIMSS data in 18 countries, we find significant beneficial

and Woessmann (2011b) reports insignificant negative coefficient estimates on expenditure per student in first-differenced and fixed-effects models.

effects of smaller classes in only two countries and can rule out large class-size effects in the majority of countries. Our estimates using this approach suggest that conventional cross-sectional estimates of class-size effects are substantially biased.[9]

These results are in line with results from a second quasi-experimental identification strategy suggested by Angrist and Lavy (1999) that exploits the existence of maximum class-size rules in many countries. Say that the maximum class size is 40, and that a certain grade has 120 students divided into three classes of 40 students each. If the grade rises to 121 students, the group is then divided into four classes—three of 30 students and one of 31 students. In this way, the rules give rise to discontinuous jumps in average class sizes whenever the enrollment in a grade in a school passes multiples of the maximum class size. Exploiting the induced class-size variation for ten European countries in a regression discontinuity design using TIMSS data, the results in Woessmann (2005a) rule out large causal class-size effects in all countries, with statistically significant but small effects in only two countries. Furthermore, the cross-country variation in estimated class-size effects in both studies is consistent with an interpretation that smaller classes have beneficial effects only in countries with relatively low teacher quality, as measured by relative teacher salary and teacher education.

The latter result is also confirmed in Altinok and Kingdon (2012), who apply yet another identification strategy to estimating class-size effects. To avoid bias from nonrandom sorting of students into schools and from unobserved student and family characteristics, they exploit the fact that the same students are tested in different subjects in TIMSS—math and science (sometimes in several specific domains). Using student fixed effects, they identify class-size effects from variation in class size between the two subjects for the same students (in countries where such variation exists). They find significant class-size effects in only 14 of 47 countries and even these are mostly small, confirming that class sizes play a limited role at best in understanding achievement differences in the international data.

Instruction Time

The length of school instruction time seems to play a more important role in educational achievement. For example, in an attempt to address omitted variable bias from unobserved individual subject-invariant characteristics such as underlying ability, motivation, or parental support, Lavy (2015) applies the within-student between-subject identification approach to estimate the effect of instruction time in the PISA 2006 data. The approach exploits the fact that different students have different instruction times in math, language, and science. He finds that instruction time has a significant positive effect on student achievement that is modest to large, suggesting that increasing instruction time by one hour per week would increase achievement by 6 percent of a standard deviation in OECD and Eastern European

[9] Applying the same instrumental variable strategy combined with school fixed effects—as well as an identification strategy based on restrictions placed on higher moments of the error distribution—to the PISA math data for the United States and the United Kingdom, Denny and Oppedisano (2013) find positive effects of larger classes, significant in the United Kingdom.

countries. This effect is only about half as large in developing countries. Furthermore, the effect of instruction time is larger in schools that have accountability measures such as using achievement data for evaluation, as well as in schools that have budgetary and personnel autonomy.

Rivkin and Schiman (2015) replicate the main finding of positive effects of instruction time in the within-student between-subject approach using the PISA 2009 data and confirm it in a model that uses within-subject variation in instructional time across grades within schools for identification. Furthermore, their results indicate that there are diminishing returns to instruction time and its effect is larger in classrooms with better environments as indicated by survey responses on questions about disruption, bullying, attendance, and other indicators of the quality of classroom environments.

Positive effects of instruction time are also confirmed in the setting of a specific education reform in Germany. The reform, which was implemented across German states at different times in the 2000s, reduced the length of the academic-track high school from nine to eight years. The reform did not change the curriculum requirements or the minimum required instruction time, so that the weekly instruction time increased in each grade. Pooling the 9th-grade samples of the extended PISA test in Germany from 2000 to 2009, Andrietti (2015) estimates the effects of the reform in a differences-in-differences framework that exploits the differing implementation years across states. Results suggest that an increase in weekly instruction time by one hour in both 8th and 9th grade increases achievement in the different subjects by between 2 and 3 percent of a standard deviation. Results are also confirmed in a "triple-difference" model that includes students in school types not affected by the reform as an additional control group.

A couple of studies have also shown that additional instruction time is related to smaller achievement gaps between different socioeconomic groups. Pooling several waves of TIMSS and PISA data, Schneeweis (2011) finds that instruction time is positively associated with the integration of immigrant students, with some models including country fixed effects so that effects are effectively identified from within-country changes over time. Pooling data from PISA and PIRLS for a differences-in-differences estimation with country fixed effects, Ammermueller (2013) finds that the achievement difference between students with different numbers of books at home is lower when instruction time is longer. There is also descriptive evidence that enrollment in early childhood education—that is, additional time before school—is related to reduced socioeconomic gradients and to better integration of migrant children (Schütz, Ursprung, and Woessmann 2008; Schneeweis 2011). Taken together, the results indicate that school instruction time can increase educational opportunities for students from disadvantaged backgrounds.

Teacher Quality

Teacher quality can be measured in a variety of ways. For example, Hanushek, Piopiunik, and Wiederhold (2014) use occupation-specific data on adult skills from the Programme for the International Assessment of Adult Competencies (PIAAC) to measure teacher skills in numeracy and literacy in 23 countries. Combining these aggregate measures of teacher skills with student-level PISA data, they estimate the effect of teacher cognitive skills on international differences in student achievement, controlling among other factors for PIAAC-based estimates of parents' cognitive skills. Models with student fixed effects that exploit within-country variation between subjects suggest that teacher skills increase student achievement. Constructing a pseudo-panel from the PIAAC data using teachers' year of birth, they also exploit cross-country differences in how alternative job opportunities for women over time have attracted people with different skills into teaching. Bietenbeck, Piopiunik, and Wiederhold (2015) apply a within-student between-subject approach to a regional achievement test of 13 sub-Saharan African countries that includes subject-specific tests of teachers. They find a significant positive effect of teacher subject knowledge on student achievement that is complementary to access to subject-specific textbooks.

Measuring teacher quality by both absolute teacher salary and teachers' relative salary position in a country's income distribution, Dolton and Marcenaro-Gutierrez (2011) find that higher teacher quality is related to better student achievement using data from several TIMSS and PISA waves. The results are consistent with positive effects of recruiting higher ability individuals into teaching. Results are confirmed when adding country fixed effects, so that estimates are identified from (relatively short-term) fluctuations in teacher pay within countries.

Apart from studies of direct measures of teacher quality, recent evidence also indicates the relevance of teaching practices. Again applying within-student between-subject identification to circumvent bias from unobserved student characteristics, Schwerdt and Wuppermann (2011) show in the US TIMSS sample that for given levels of teaching methods, traditional lecture-style teaching is related to better student achievement compared to classroom problem-solving. Using the same estimation strategy on TIMSS data for the United States and nine advanced countries, Bietenbeck (2014) finds that traditional teaching practices are related to better overall skills, factual knowledge, and solving of routine problems, whereas modern teaching practices are related to better reasoning skills. After showing cross-country correlations of teaching practices with measures of social capital, Algan, Cahuc, and Shleifer (2013) apply a cross-sectional model with school fixed effects to TIMSS and PIRLS data to show that progressive practices of having students work in groups are positively related to student beliefs about cooperation and to student self-confidence.

Despite the result that resource inputs overall play a limited role, instruction time and certain dimensions of teacher quality do seem to matter for student achievement. More broadly, these findings suggest ways in which what school systems do are relevant for educational achievement. Moreover, looking into determinants of instruction time and teacher quality leads naturally to questions about the institutional framework of school systems that may frame how resources are used.

Institutional Structures of School Systems: Explorations into Causal Effects

An international comparative approach promises to be fruitful in studying the effects of educational institutions because institutional structures often do not vary nearly as much within countries as they do across countries. Specific institutional features that have been found to matter for cross-country differences in student achievement include external exams, school autonomy, private competition, and tracking.

External Exams

In some countries, learning outcomes are assessed by curriculum-based external exit exams that have real consequences for students (Bishop 1997). A large literature has shown consistent positive associations between external exams and student achievement (Hanushek and Woessmann 2011a). However, such cross-country associations may be biased by unobserved country characteristics such as specific cultures. For example, a society that favors high educational achievement might both introduce external exams and also make efforts to induce students to study, and a positive correlation between external exams and student achievement does not show that the former has a causal effect on the latter.

There are several ways to explore whether these cultural effects are important in explaining the connection from exit exams to test scores. One approach is to look at variation in test scores and exams only within continents. If the international variation in test scores would have been biased by features more relevant in some continents than in others—say, if countries in Asia place a higher value on educational success than countries in other regions—then the coefficient on external exams will decline in such a model. However, in Woessmann (2003a), I find that the association between external exams and student achievement in the first two TIMSS waves is robust to the inclusion of continental fixed effects. Another approach looks at evidence across states within Germany and compares this with other OECD countries. German states differ in whether they have external exams or not, but are otherwise much more similar than OECD countries. However, in this mixture of PISA data on German states and other countries, students in systems with external exams have around 20 percent of a standard deviation higher achievement, and this association is statistically indistinguishable between the OECD country sample and the German state sample (Woessmann 2010). This result corroborates that the international association is unlikely to be driven by fundamental differences in culture, language, or other institutional settings that do not vary within Germany.

In yet another approach, Jürges, Schneider, and Büchel (2005) use the German TIMSS 1995 data in a differences-in-differences approach that exploits variation across subjects: specifically, in the relevant school tracks, most German states that have external exams have them in math but not in science. The identifying assumption of this model is that cross-state achievement differences would not differ between subjects in the absence of the external exam treatment. While smaller than their cross-sectional estimates, their differences-in-differences estimates are significant and substantial at between 13 and 26 percent of a standard deviation. If there are spillovers between subjects—for example, improved math knowledge due to external exams also facilitates students' learning in science—these estimates provide a lower bound for the full effect of external exams. Until the early 2000s, only seven of the 16 German states had external exams, but all but one have introduced them over the course of the 2000s. Lüdemann (2011) exploits the different timing of the introduction of external exams across states and school types in a differences-in-differences approach using the German extended PISA waves from 2000 to 2006. The identifying assumption is that there would have been common trends in the absence of the external exam treatment. Results indicate significant positive effects of the introduction of central exit exams even in the short run.

While external exams direct incentives particularly on students, a way to incentivize teachers to focus on student outcomes is performance-related pay. Apart from showing a positive association of teacher pay with student achievement in PISA data, in Woessmann (2011), I find that teacher salary adjustments for outstanding performance are positively associated with student achievement across countries. The use of a country-level measure of teacher performance pay avoids bias from within-country selection, and results are robust to including continental fixed effects and to controlling for other forms of teacher salary adjustments that are not based on performance. An advantage of the cross-country approach is that it captures general-equilibrium effects such as sorting into the teaching profession and other long-run incentive effects, whereas short-term merit pay experiments capture only incentive effects, not selection effects.

School Autonomy

On the one hand, school autonomy may be conducive to student achievement in school systems with strong surrounding structures that ensure high common standards; on the other hand, school-based decision-making could hurt student achievement in low-performing systems that lack basic standards and local capacity. Cross-sectional evidence from international achievement tests concerning school autonomy has been mixed (Hanushek and Woessmann 2011a), but these studies may also be particularly plagued by identification issues.

To avoid bias from unobserved cross-country differences such as those arising from culture and other government institutions, in Hanushek, Link, and Woessmann (2013), we introduce the analytical approach of country panel analysis with country fixed effects. Because many countries have reformed their school systems to become more or less autonomous over time, we can exploit country-level variation over time by including country fixed effects that control for systematic, time-invariant differences across countries. While such panel analysis does not necessarily identify random variation, we show that prior achievement and prior GDP do not predict autonomy reforms. To avoid bias from within-country selection of students into autonomous schools and of schools to become autonomous, we aggregate the school autonomy measure to the country level, reflecting the average share of autonomous schools in a country.

Pooling the individual data of over one million students in 42 countries in the four PISA waves from 2000 to 2009, we find that school autonomy has a significant effect on student achievement, but this effect varies systematically with the level of economic and educational development: The effect is strongly positive in developed and high-performing countries but strongly negative in developing and low-performing countries. The estimates suggest that going from no to full autonomy over academic content would increase student achievement by 53 percent of a standard deviation in the highest-income country (Norway) and reduce student achievement by 55 percent of a standard deviation in the lowest-income country (Indonesia).

If part of the negative effect of school autonomy stems from a lack of accountability, these negative aspects should be eased in school systems where external exams provide comparative information on ultimate performance. Indeed, in Hanushek, Link, and Woessmann (2013), we find a significant positive interaction between changes in school autonomy and (initial) external exit exams—that is, introducing autonomy is more beneficial in school systems that have accountability through external exams.

The effects of school autonomy may also be interrelated with the management capacity of schools. Collecting data on school management practices in operations, monitoring, target setting, and people management in eight countries, Bloom, Lemos, Sadun, and Van Reenan (2015) find higher management quality to be related to better student achievement. While mostly focusing on specific national achievement datasets, they also report a positive correlation with average PISA scores across Italian and German regions. Furthermore, autonomous public schools score highly in terms of management quality. Interestingly, while their previous work suggested that most of the variation in management quality in other sectors is within-country, about half of the variance in management quality in the school sector is between countries, underlining the importance of cross-country analysis of institutional environments in school systems.

Private Competition

The extent to which schools are operated by public or private entities differs markedly across countries. For example, more than three-quarters of 15 year-old students in the Netherlands attend privately operated schools and more than 60 percent in Belgium and Ireland, but this share is below 10 percent in many other countries. Private school operation is largely independent of the funding of schools; for example, the average share of government funding of Dutch privately operated schools is the same (at 95 percent) as in public schools, a feature going back to constitutional provisions. Private school operation may be related to the extent of school autonomy, but again these are conceptually different issues: public schools can have substantial autonomy, and private schools can have limited autonomy. A key point is that competition from private alternatives may improve the performance of public schools as well, which may lift the achievement level systemwide.

Cross-country evidence indeed suggests a strong association of achievement levels with the share of privately operated schools (for example, Woessmann 2009), but identification issues are again obvious in cross-country analyses: low quality of the public school system may induce a political system to encourage private alternatives or parents to choose private alternatives, and other country features related to the supply of or demand for private schools may introduce omitted variable bias.

To identify exogenous variation in the share of private schools across countries, in West and Woessmann (2010), my coauthor and I argue that historical differences in Catholic versus Protestant denominations provide a natural experiment. In late 19th century, Catholic doctrine resisted the emerging nondenominational public school systems and spurred efforts to establish private schools in many countries. These efforts were most successful in countries with substantial shares of Catholic populations but without a Catholic state religion. Therefore, the share of Catholics in a country's population in 1900 (interacted with an indicator as to whether Catholicism was the state religion) can be used as an instrumental variable for the share of privately operated schools in the 2003 PISA data.[10] The results suggest that a 10 percentage point increase in private school shares, induced by historical Catholic resistance to state schooling, leads to an increase in math achievement by at least 9 percent of a standard deviation. Much of this effect accrues to students in public schools, suggesting that most of the overall effect reflects benefits of private competition and parental choice. In addition to increasing achievement, private competition is also estimated to reduce total educational expenditure per student.

Tracking

Countries vary in the extent to which students are tracked into different school types by ability. No country has differing-ability schools in the early grades of primary school. Some countries such as Austria and Germany track students into different-ability schools as early as age 10. Many other countries have a comprehensive school system (although perhaps with some streaming within schools) through the end of high school. A common concern is that early tracking may increase inequality as lower-achieving groups are tracked into lower-ability schools, perhaps because of peer effects.

In Hanushek and Woessmann (2006), we suggest an identification strategy that compares achievement changes from primary to later schooling across tracked and untracked countries. Using country-level data for several pairs of PIRLS, TIMSS, and PISA achievement tests administered at the primary and secondary school levels in the context of a differences-in-differences model, we find that early tracking significantly increases the inequality in countries' achievement outcomes. We do not find a consistent effect of early tracking on the level of achievement, although most estimates tend to be negative. Interestingly, simple cross-sectional estimations do not indicate an association of tracking with educational inequality.

A variety of other results suggest that earlier tracking tends to raise the inequality of educational outcomes. Applying the same estimation strategy across grades to student-level PIRLS and PISA data, Ammermueller (2013) finds that early tracking and the number of tracked school types increase the effect of parental education on student achievement. Again using the same identification strategy to estimate the effect of tracking on the migrant–native achievement gap in a pooled micro dataset of all PIRLS, TIMSS, and PISA waves from 1995 to 2012, Ruhose and Schwerdt (2016) do not find that early tracking affects native and migrant students differently in general. However, they find a detrimental effect of early tracking on the relative achievement of first-generation migrants and the presumably less-integrated subgroup of second-generation migrant students who do not speak the host-country language at home. Piopiunik (2014) exploits a school reform in Bavaria that

[10] There is ample evidence that historically, Catholics have placed less emphasis on education than Protestants (for example, Becker and Woessmann 2009), which would bias the instrumental-variable model against finding beneficial effects of competition. Indeed, the current share of Catholics enters negatively in the second-stage model.

lowered the age of tracking between the two lowest-ability school types to estimate a "triple-difference" model using variation across three German PISA waves that allow a comparison of outcomes in the reformed system to pre-reform outcomes, to other German states, and to the non-treated highest-ability school type. Results suggest that earlier tracking reduced achievement in both low- and middle-track schools.

Conclusions

What explains the large international differences in student achievement? On a descriptive basis, a simple model of three combined factors of family background, school resources, and institutions is able to account for more than four-fifths of the total cross-country variation in student achievement. Family background and institutions contribute roughly equally to this exercise, whereas the contribution of school resources is quite limited—although the predictive power of the model varies across countries. Beyond these descriptive patterns, a growing literature uses quasi-experimental methods in an attempt to identify causal effects of school systems in the international test data, as well as different types of fixed effects models that aim to avoid certain sources of bias. Some patterns emerge from this literature. First, this work tends to confirm that resource inputs such as expenditure per student or class size appear to have limited effects on student achievement. Second, instruction time and measures of teacher quality do play a role. Third, a number of institutional features of school systems seem to contribute to the cross-country differences in student achievement. For example, external exit exams and competition from privately operated schools positively affect achievement levels. School autonomy has positive effects in developed countries and where external exit exams introduce accountability, but negative effects in developing countries. Early tracking into differing-ability schools seems to increase inequality in achievement without increasing achievement levels.

Clearly, the exploitation of the potential of international differences in student achievement to improve our understanding of educational processes is work in progress. In the future, increasing numbers of participating countries and an expanding number of waves of available international achievement tests will raise the scope of possible investigations. A useful direction for international testing efforts would be to conduct studies in many countries that are longitudinal at the student level. Existing causal identification strategies will be sharpened and new approaches developed. It may be especially useful to focus on interactions between the kinds of factors examined here: for example, little is known about the particular institutional settings that may strengthen the effectiveness of resource use.

As this work proceeds, it is perhaps useful to remember what is at stake. Levels and changes in educational achievement are a powerful determinant of output levels and economic growth. It has long been common to use average years of schooling in regressions that seek to explain economic growth. But average years of schooling may be a very noisy measure of actual educational achievement as measured by test scores. Thus, in Hanushek and Woessmann (2012, 2015a), we show that a model that includes only countries' average years of schooling and their initial level of GDP per capita as predictors accounts for one-quarter of the total cross-country variation in growth rates in GDP per capita from 1960 to 2000 (or 2009). However, adding average scores on the international achievement tests between 1964 and 2003 to the model accounts for more than three-quarters of the variation in long-term growth rates of per-capita GDP—indeed, it renders the commonly used quantitative measure of years of schooling insignificant. Differences in math and science achievement can fully account both for the stunning growth performance of the East Asian miracle countries and for the disheartening growth performance of Latin American countries (Hanushek and Woessmann 2016).

In Hanushek and Woessmann (2012, 2015a), we report several econometric analyses that provide a *prima facie* case that the close and robust association of educational achievement with countries' long-run economic growth reflects a causal effect of population skills. To preclude simple reverse causation, we show that achievement tests before 1985 predict subsequent growth. To address potential bias from omitted factors such as differing economic institutions or cultures, we present instrumental-variable models that use only part of the skill variation that can be predicted from institutional differences in school systems; show that changes in test scores predict changes in growth; perform development accounting analyses that take parameter values from the micro literature; and report differences-in-differences models showing that immigrants educated in their home countries receive returns to their home-country cognitive skills on the US labor market, whereas immigrants from the same home countries schooled in the United States do not. Additional recent work on student achievement and international income differences is given in Kaarsen (2014), and we provide reviews in Hanushek and Woessmann (2008, 2011a). Within the United States, in Hanushek, Ruhose, and Woessmann (2015), we confirm an important role for educational achievement in explaining differences in GDP per capita across US states. At the individual level, performance on adult achievement tests is strongly associated with employment and earnings in each of the 23 countries analyzed in Hanushek, Schwerdt, Wiederhold, and Woessmann (2015). Murnane, Willett, and Levy (1995) and Chetty et al. (2011), among others, provide additional evidence on individual returns to educational achievement in the United States.

Of course, the implications of improved educational achievement go well beyond individual earnings and macroeconomic growth rates. Education is important for economic inequality and the transmission of inequality across generations (for example, Black and Devereux 2011). Education affects the education and health of children, own health, crime, and citizenship (for example, Lochner 2011). More broadly, a "capabilities approach" to welfare analysis in the style of Sen and Nussbaum (for example, Nussbaum and Sen 1993) emphasizes that education is an important determinant of the ability of people to develop their own capacities and in that sense to be able to exercise autonomy and choice in all aspects of life.

■ *Helpful comments from Mark Gertler, Gordon Hanson, Eric Hanushek, Enrico Moretti, Marc Piopiunik, Jens Ruhose, Timothy Taylor, Martin West, and Simon Wiederhold are gratefully acknowledged.*

References

Algan, Yann, Pierre Cahuc, and Andrei Shleifer. 2013. "Teaching Practices and Social Capital." *American Economic Journal: Applied Economics* 5(3): 189–210.

Altinok, Nadir, and Geeta Kingdon. 2012. "New Evidence on Class Size Effects: A Pupil Fixed Effects Approach." *Oxford Bulletin of Economics and Statistics* 74(2): 203–34.

Ammermueller, Andreas. 2013. "Institutional Features of Schooling Systems and Educational Inequality: Cross-country Evidence from PIRLS and PISA." *German Economic Review* 14(2): 190–213.

Ammermueller, Andreas, and Jörn-Steffen Pischke. 2009. "Peer Effects in European Primary Schools: Evidence from the Progress in International Reading Literacy Study." *Journal of Labor Economics* 27(3): 315–48.

Andrietti, Vincenzo. 2015. "The Causal Effects of Increased Learning Intensity on Student Achievement: Evidence from a Natural Experiment." Universidad Carlos III de Madrid, Working Paper, Economic Series 15-06. Madrid: Universidad Carlos III.

Angrist, Joshua D., and Victor Lavy. 1999. "Using Maimonides' Rule to Estimate the Effect of Class Size on Scholastic Achievement." *Quarterly Journal of Economics* 114(2): 533–75.

Becker, Sascha O., and Ludger Woessmann. 2009. "Was Weber Wrong? A Human Capital Theory of Protestant Economic History." *Quarterly Journal of Economics* 124(2): 531–96.

Bedard, Kelly, and Elizabeth Dhuey. 2006. "The Persistence of Early Childhood Maturity: International Evidence of Long-Run Age Effects." *Quarterly Journal of Economics* 121(4): 1437–72.

Bietenbeck, Jan. 2014. "Teaching Practices and Cognitive Skills." *Labour Economics* 30: 143–53.

Bietenbeck, Jan, Marc Piopiunik, and Simon Wiederhold. 2015. "Africa's Skill Tragedy: Does Teachers' Lack of Knowledge Lead to Low Student Performance?" CESifo Working Paper 5470.

Bishop, John H. 1997. "The Effect of National Standards and Curriculum-based Examinations on Achievement." *American Economic Review* 87(2): 260–64.

Black, Sandra E., and Paul J. Devereux. 2011. "Recent Developments in Intergenerational Mobility." In *Handbook of Labor Economics*, Vol. 4B, edited by Orley Ashenfelter and David Card, 1487–1541. Amsterdam: North Holland.

Bloom, Nicholas, Renata Lemos, Raffaella Sadun, and John Van Reenen. 2015. "Does Management Matter in Schools?" *Economic Journal* 125(584): 647–74.

Brunello, Giorgio, and Daniele Checchi. 2007. "Does School Tracking Affect Equality of Opportunity? New International Evidence." *Economic Policy* 22(52): 781–861.

Brunello, Giorgio, and Lorenzo Rocco. 2013. "The Effect of Immigration on the School Performance of Natives: Cross Country Evidence Using PISA Test Scores." *Economics of Education Review* 32(1): 234–46.

Chetty, Raj, John N. Friedman, Nathaniel Hilger, Emmanuel Saez, Diane Whitmore Schanzenbach, and Danny Yagan. 2011. "How Does Your Kindergarten Classroom Affect Your Earnings? Evidence from Project STAR." *Quarterly Journal of Economics* 126(4): 1593–1660.

Cobb-Clark, Deborah A., Mathias Sinning, and Steven Stillman. 2012. "Migrant Youths' Educational Achievement: The Role of Institutions." *Annals of the American Academy of Political and Social Science* 643(1): 18–45.

Denny, Kevin, and Veruska Oppedisano. 2013. "The Surprising Effect of Larger Class Sizes: Evidence Using Two Identification Strategies." *Labour Economics* 23: 57–65.

Dolton, Peter, and Oscar D. Marcenaro-Gutierrez. 2011. "If You Pay Peanuts Do You Get Monkeys? A Cross-country Analysis of Teacher Pay and Pupil Performance." *Economic Policy* 26(65): 5–55.

Dustmann, Christian, Tommaso Frattini, and Gianandrea Lanzara. 2012. "Educational Achievement of Second-Generation Immigrants: An International Comparison." *Economic Policy* 27(69): 143–85.

Edwards, Sebastian, and Alvaro Garcia Marin. 2015. "Constitutional Rights and Education: An International Comparative Study." *Journal of Comparative Economics* 43(4): 938–55.

Foshay, Arthur W. 1962. "The Background and the Procedures of the Twelve-Country Study." In *Educational Achievement of Thirteen-year-olds in Twelve Countries: Results of an International Research Project, 1959–61*, edited by Arthur W. Foshay, Robert L. Thorndike, Fernand Hotyat, Douglas A. Pidgeon, and David A. Walker. Hamburg: Unesco Institute for Education.

Freeman, Richard B., and Martina Viarengo. 2014. "School and Family Effects on Educational Outcomes across Countries." *Economic Policy* 29(79): 395–446.

Fryer, Roland G., Jr., and Steven D. Levitt. 2010. "An Empirical Analysis of the Gender Gap in Mathematics." *American Economic Journal: Applied Economics* 2(2): 210–240.

Fuchs, Thomas, and Ludger Woessmann. 2007.

"What Accounts for International Differences in Student Performance? A Re-examination Using PISA Data." *Empirical Economics* 32(2–3): 433–64.

Guiso, Luigi, Ferdinando Monte, Paola Sapienza, and Luigi Zingales. 2008. "Culture, Math, and Gender." *Science* 320(5880): 1164–65.

Gundlach, Erich, Ludger Woessmann, and Jens Gmelin. 2001. "The Decline of Schooling Productivity in OECD Countries." *Economic Journal* 111(471): C135–C47.

Hanushek, Eric A. 1986. "The Economics of Schooling: Production and Efficiency in Public Schools." *Journal of Economic Literature* 24(3): 1141–77.

Hanushek, Eric A. 2002. "Publicly Provided Education." In *Handbook of Public Economics*, Vol. 4, edited by Alan J. Auerbach and Martin Feldstein, 2045–2141. Amsterdam: North Holland.

Hanushek, Eric A., and Dennis D. Kimko. 2000. "Schooling, Labor Force Quality, and the Growth of Nations." *American Economic Review* 90(5): 1184–1208.

Hanushek, Eric A., Susanne Link, and Ludger Woessmann. 2013. "Does School Autonomy Make Sense Everywhere? Panel Estimates from PISA." *Journal of Development Economics* 104: 212–32.

Hanushek, Eric A., Paul E. Peterson, and Ludger Woessmann. 2013. *Endangering Prosperity: A Global View of the American School.* Washington, DC: Brookings Institution Press.

Hanushek, Eric A., Paul E. Peterson, and Ludger Woessmann. 2014. "U.S. Students from Educated Families Lag in International Tests." *Education Next* 14(4): 8–18.

Hanushek, Eric A., Marc Piopiunik, and Simon Wiederhold. 2014. "The Value of Smarter Teachers: International Evidence on Teacher Cognitive Skills and Student Performance." NBER Working Paper 20727.

Hanushek, Eric A., Jens Ruhose, and Ludger Woessmann. 2015. "Human Capital Quality and Aggregate Income Differences: Development Accounting for U.S. States." NBER Working Paper 21295.

Hanushek, Eric A., Guido Schwerdt, Simon Wiederhold, and Ludger Woessmann. 2015. "Returns to Skills around the World: Evidence from PIAAC." *European Economic Review* 73: 103–130.

Hanushek, Eric A., and Ludger Woessmann. 2006. "Does Educational Tracking Affect Performance and Inequality? Differences-in-differences Evidence across Countries." *Economic Journal* 116(510): C63–C76.

Hanushek, Eric A., and Ludger Woessmann. 2008. "The Role of Cognitive Skills in Economic Development." *Journal of Economic Literature* 46(3): 607–68.

Hanushek, Eric A., and Ludger Woessmann. 2011a. "The Economics of International Differences in Educational Achievement." In *Handbook of the Economics of Education*, Vol. 3, edited by Eric A. Hanushek, Stephen Machin, and Ludger Woessmann, 89–200. Amsterdam: North Holland.

Hanushek, Eric A., and Ludger Woessmann. 2011b. "How Much Do Educational Outcomes Matter in OECD Countries?" *Economic Policy* 26(67): 427–91.

Hanushek, Eric A., and Ludger Woessmann. 2011c. "Sample Selectivity and the Validity of International Student Achievement Tests in Economic Research." *Economics Letters* 110(2): 79–82.

Hanushek, Eric A., and Ludger Woessmann. 2012. "Do Better Schools Lead to More Growth? Cognitive Skills, Economic Outcomes, and Causation." *Journal of Economic Growth* 17(4): 267–321.

Hanushek, Eric A., and Ludger Woessmann. 2015a. *The Knowledge Capital of Nations: Education and the Economics of Growth.* Cambridge, MA: MIT Press.

Hanushek, Eric A., and Ludger Woessmann. 2015b. *Universal Basic Skills: What Countries Stand to Gain.* Paris: Organisation for Economic Co-operation and Development.

Hanushek, Eric A., and Ludger Woessmann. 2016. "Knowledge Capital, Growth, and the East Asian Miracle." *Science* 351(6271): 344–45.

Hoxby, Caroline M. 2000. "The Effects of Class Size on Student Achievement: New Evidence from Population Variation." *Quarterly Journal of Economics* 115(3): 1239–85.

IEA. 2016. "Brief History of IEA: 55 Years of Educational Research." Amsterdam: International Association for the Evaluation of Educational Achievement. http://www.iea.nl/brief_history.html.

Jerrim, John, and John Micklewright. 2014. "Socio-economic Gradients in Children's Cognitive Skills: Are Cross-country Comparisons Robust to Who Reports Family Background?" *European Sociological Review* 30(6): 766–81.

Jürges, Hendrik, Kerstin Schneider, and Felix Büchel. 2005. "The Effect of Central Exit Examinations on Student Achievement: Quasi-experimental Evidence from TIMSS Germany." *Journal of the European Economic Association* 3(5): 1134–55.

Kaarsen, Nicolai. 2014. "Cross-Country Differences in the Quality of Schooling." *Journal of Development Economics* 107: 215–24.

Lavy, Victor. 2015. "Do Differences in Schools' Instruction Time Explain International Achievement Gaps? Evidence from Developed and Developing Countries." *Economic Journal* 125(588): F397–F424.

Lee, Jong-Wha, and Robert J. Barro. 2001. "Schooling Quality in a Cross-section of Countries." *Economica* 68(272): 465–88.

Lochner, Lance. 2011. "Nonproduction Benefits of Education: Crime, Health, and Good Citizenship." In *Handbook of the Economics of Education*, Vol. 4, edited by Eric A. Hanushek, Stephen Machin, and Ludger Woessmann, 183–282. Amsterdam: North Holland.

Lüdemann, Elke. 2011. "Intended and Unintended Short-run Effects of the Introduction of Central Exit Exams: Evidence from Germany." In Elke Lüdemann, *Schooling and the Formation of Cognitive and Non-cognitive Outcomes*. ifo Beiträge zur Wirtschaftsforschung 39. Munich: ifo Institut.

Mullis, Ina V. S., Michael O. Martin, Pierre Foy, and Alka Arora. 2012. *TIMSS 2011 International Results in Mathematics*. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.

Murnane, Richard J., John B. Willett, and Frank Levy. 1995. "The Growing Importance of Cognitive Skills in Wage Determination." *Review of Economics and Statistics* 77(2): 251–66.

Nussbaum, Martha C., and Amartya Sen, eds. 1993. *The Quality of Life*. Oxford University Press.

OECD. 2013. *PISA 2012 Results: What Students Know and Can Do—Student Performance in Mathematics, Reading and Science*, Vol 1. Paris: Organisation for Economic Co-operation and Development.

Piopiunik, Marc. 2014. "The Effects of Early Tracking on Student Performance: Evidence from a School Reform in Bavaria." *Economics of Education Review* 42: 12–33.

Ripley, Amanda. 2013. *The Smartest Kids in the World—And How They Got That Way*. New York: Simon & Schuster.

Rivkin, Steven G., and Jeffrey C. Schiman. 2015. "Instruction Time, Classroom Quality, and Academic Achievement." *Economic Journal* 125(588): F425–F448.

Ruhose, Jens, and Guido Schwerdt. 2016. "Does Early Educational Tracking Increase Migrant–Native Achievement Gaps? Differences-in-Differences Evidence across Countries." *Economics of Education Review* 52: 134–54.

Schneeweis, Nicole. 2011. "Educational Institutions and the Integration of Migrants." *Journal of Population Economics* 24(4): 1281–1308.

Schütz, Gabriela, Heinrich W. Ursprung, and Ludger Woessmann. 2008. "Education Policy and Equality of Opportunity." *Kyklos* 61(2): 279–308.

Schwerdt, Guido, and Amelie C. Wuppermann. 2011. "Is Traditional Teaching Really All That Bad? A Within-Student Between-Subject Approach." *Economics of Education Review* 30(2): 365–79.

Singh, Abhijeet. 2015. "Learning More with Every Year: School Year Productivity and International Learning Divergence." Presented at the CESifo Area Conference on the Economics of Education, September 11–12, 2015. Available at: http://www.cesifo-group.de/de/ifoHome/events/Archive/conferences/2015/09/2015-09-11-ee15-Hanushek/Programme.html.

West, Martin R., and Ludger Woessmann. 2010. "'Every Catholic Child in a Catholic School': Historical Resistance to State Schooling, Contemporary Private Competition and Student Achievement across Countries." *Economic Journal* 120 (546): F229–F255.

Woessmann, Ludger. 2003a. "Central Exit Exams and Student Achievement: International Evidence." In *No Child Left Behind? The Politics and Practice of School Accountability*, edited by Paul E. Peterson and Martin R. West, 292–323. Washington, D.C.: Brookings Institution Press.

Woessmann, Ludger. 2003b. "Schooling Resources, Educational Institutions, and Student Performance: The International Evidence." *Oxford Bulletin of Economics and Statistics* 65(2): 117–70.

Woessmann, Ludger. 2005a. "Educational Production in Europe." *Economic Policy* 20(43): 446–504.

Woessmann, Ludger. 2005b. "The Effect Heterogeneity of Central Exams: Evidence from TIMSS, TIMSS-Repeat and PISA." *Education Economics* 13(2): 143–169.

Woessmann, Ludger. 2009. "Public–Private Partnerships and Student Achievement: A Cross-Country Analysis." In *School Choice International: Exploring Public–Private Partnerships*, edited by Rajashri Chakrabarti and Paul E. Peterson, 13–45. Cambridge, MA: MIT Press.

Woessmann, Ludger. 2010. "Institutional Determinants of School Efficiency and Equity: German States as a Microcosm for OECD Countries." *Journal of Economics and Statistics* 230(2): 234–70.

Woessmann, Ludger. 2011. "Cross-Country Evidence on Teacher Performance Pay." *Economics of Education Review* 30(3): 404–18.

Woessmann, Ludger, Elke Luedemann, Gabriela Schuetz, and Martin R. West. 2009. *School Accountability, Autonomy, and Choice around the World*. Cheltenham, UK: Edward Elgar.

Woessmann, Ludger, and Martin R. West. 2006. "Class-Size Effects in School Systems around the World: Evidence from Between-Grade Variation in TIMSS." *European Economic Review* 50(3): 695–736.

Journal of Economic Perspectives—Volume 30, Number 3—Summer 2016—Pages 33–56

Accountability in US Education: Applying Lessons from K–12 Experience to Higher Education

David J. Deming and David Figlio

A new push for accountability has become an increasingly important feature of education policy in the United States and throughout the world. Broadly speaking, accountability seeks to hold educational institutions responsible for student outcomes using tools ranging from performance "report cards" to explicit rewards and sanctions.

In the United States, the accountability movement was presaged by the 1983 publication of *A Nation at Risk*, an incendiary report authored by a commission appointed by President Ronald Reagan. The report famously stated that "if an unfriendly foreign power had attempted to impose on America the mediocre educational performance that exists today, we might well have viewed it as an act of war." *A Nation at Risk* called for—among other changes—more rigorous performance measurement, including nationwide standardized testing. Beginning in the late 1980s, some early-adopting US states—along with countries such as Chile and the United Kingdom—began rating and ranking K–12 schools using measures of student performance. The accountability movement in the United States culminated with the passage of the No Child Left Behind (NCLB) Act of 2001, which required all states to test K–12 students regularly in core subjects and to evaluate

■ *David Deming is Professor, Harvard Graduate School of Education, Cambridge, Massachusetts. David Figlio is Orrington Lunt Professor of Education and Social Policy and Director, Institute for Policy Research, both at Northwestern University, Evanston, Illinois. Deming is a Faculty Research Fellow and Figlio is a Research Associate, National Bureau of Economic Research, Cambridge, Massachusetts. Their email addresses are david_deming@gse.harvard.edu and figlio@northwestern.edu.*

† For supplementary materials such as appendices, datasets, and author disclosure statements, see the article page at
http://dx.doi.org/10.1257/jep.30.3.33 doi=10.1257/jep.30.3.33

schools based on whether their students were making adequate progress toward achievement benchmarks, with the goal of 100 percent proficiency by 2014.

Oddly enough, we can now view No Child Left Behind Act of 2001 as the beginning of a process of gradual retreat from accountability. Perhaps unsurprisingly, it became clear soon after the passage of NCLB that states would fail to attain the lofty goal of 100 percent proficiency. Starting in 2011, the Obama administration began waiving NCLB requirements for states that agreed to adopt certain policies such as linking teacher evaluations to student test scores. However, this step proved politically contentious and may have contributed to an anti-testing backlash in states such as New York, Florida, and Texas that had formerly implemented some of the most ambitious accountability policies. In December 2015, No Child Left Behind was replaced by the Every Student Succeeds Act, which scales back testing requirements and returns more implementation power to the states.

This dynamic of initial enthusiasm for accountability followed by gradual retrenchment has also played out in US higher education. President Obama announced in 2013 an ambitious plan to rate colleges based on access, affordability, and student outcomes. However, after a two-year process of soliciting feedback from colleges and higher education experts, the Obama administration elected to scrap the rating system as well as any explicit linkage between federal funding and performance. In its place, the administration released the College Scorecard, a database and interactive website where information about graduation rates, earnings, and annual costs of postsecondary institutions can be compared in a standardized format.

In both cases, accountability began with a period of surging policy interest combined with technocratic exuberance about measuring and tracking educational outcomes, which was then followed by caution about unintended consequences and a sense that certain efforts may have overreached. The status quo represents an uneasy compromise. Even the harshest critics of accountability would probably concede that it is hard to stop measuring and tracking performance once you have started. Accountability in some form is probably here to stay. But important questions remain concerning the specifics of how an accountability system should be designed, and what such a system can reasonably be expected to accomplish.

Our purpose in this article is to provide a framework for understanding educational accountability at the K–12 and higher education levels. We start with a discussion of the context from which this push for greater accountability emerged and discuss some of the theoretical arguments behind approaches to accountability in education. We then turn to the well-developed empirical literature on accountability in K–12 education and consider what lessons we can learn for the design and impact of college ratings.

Our bottom line is that accountability works, but rarely as well as one would hope, and often not entirely in the ways that were intended. Research on K–12 accountability offers some hope but also a number of cautionary tales. Importantly, the benefits of K–12 accountability seem to be concentrated among the most disadvantaged students in the lowest-performing schools, both perhaps because failure is easier to diagnose than success and because lower-performing schools face less scrutiny from stakeholders in the absence of government monitoring.

What lessons can we learn for the design of accountability in US higher education? US colleges and universities vary greatly in selectivity, their broad purposes, and whether they are public or private, making the design challenges even more formidable than in the K–12 setting. In this uncertain environment, the conservative approach of the College Scorecard—standardized performance information, but no explicit stakes—is a sensible choice, though one that still could have significant unintended consequences.

However, we will argue that it is possible to do better by targeting regulatory efforts toward lower-performing institutions of higher education where students have less "skin in the game." This includes many for-profit colleges that depend heavily on federal financial aid for revenue, but it also potentially includes public (and some private) institutions where few students are paying out-of-pocket due to federal and state subsidies. We think "skin in the game" can work as a guiding principle, for two reasons. First, paying customers send a market signal that schools are providing a valuable product—where value is defined by the student herself rather than policymakers. Second, schools rightly deserve increased public scrutiny when they are more heavily subsidized by taxpayer funds. We close with a call for state-level policy experimentation, in the spirit of critical attention to design details and cautious incrementalism.

The Institutions of School Accountability

The push for school accountability arose from well-known concerns about the performance of the US education system, which we briefly summarize here. At the K–12 level, spending per student has risen substantially over time, from $5,984 per student in 1970 to $12,008 in 2000 to $13,142 per student in 2013 (all expressed in 2015 dollars). However, progress in terms of student achievement has been much slower. Between 1971 and 1999, reading scores of nine year-olds on the National Assessment of Educational Progress (NAEP)—a national exam with consistent scoring over time (it is often referred to as "The Nation's Report Card")—rose by only 4 points (from 208 to 212). Gains in mathematics, while stronger, were still relatively modest. Moreover, students from other countries often outperform US students on international tests (as discussed in Woessmann's paper in this symposium). Encouragingly, NAEP scores have risen relatively rapidly between 1999 and 2012, with particularly large gains for younger students and for students of color. An additional piece of good news is that high school graduation rates rose by more than 10 percentage points between 2000 and 2013 after stagnating during the previous three decades (Murnane 2013). There is some evidence—discussed later in the paper—that accountability has had relatively larger impacts at the bottom of the achievement distribution, and thus may have contributed to the narrowing of achievement gaps over this period. However, other candidate explanations such as changing family environments and increases in early education lead us to stress that this conclusion is speculative.

Trends in US higher education are much less positive. While college attendance rates have risen steadily over the last several decades, high dropout rates have led to only modest growth in bachelor's degree attainment. Only about 60 percent of bachelor's degree-seeking students successfully obtain a degree after six years. As a result, the US four-year college degree attainment rate ranks slightly below the average of high-income OECD countries.

Public higher education in the United States is funded primarily by a combination of student tuition and state legislative appropriations. State subsidies allow public colleges and universities to spend more per student than they charge in tuition prices. Yet declining state support means that students are paying for a larger share of their education. Between the 1999–2000 and 2014–2015 school years, inflation-adjusted state funding per full-time equivalent (FTE) student declined by about 25 percent. Tuition revenue per FTE student increased over this period by a similar amount, leaving total per-student spending in public institutions relatively constant. Prices are increasing, but spending is not—students are just footing a higher share of the bill through out-of-pocket spending and student loans, which now total more than $1.3 trillion.

Despite these headwinds, college continues to be a worthwhile investment on average. Avery and Turner (2012) estimate that the present discounted lifetime value of a college degree relative to a high school degree—net of tuition—is positive and large, and has actually grown over time despite rising prices and growing student loan debt. In sum, college appears to be an increasingly risky—yet also increasingly necessary—investment.

What is the role of US educational institutions in producing these mixed outcomes? The governance and funding of the US education system, as in most countries, comes primarily from the public sector. US K–12 public schools are managed by elected school boards, while principals, teachers, and other employees are public sector workers. US higher education has a more diverse institutional structure, but the majority of students attend public institutions that receive considerable public support, both directly through tax revenues and indirectly through provision of student loans and other methods. Private colleges and universities receive considerable government funding as well, be it directly through program support or grants or indirectly through subsidization via the deductibility of charitable contributions that subsidize institutional endowments, and students of public and private institutions alike receive federal and state financial aid. Most institutions of higher education, whether public or private, are managed by a combination of a board of trustees, who hire the president and sometimes other top administrators, and the faculty of those institutions.

Given the importance of the public sector in providing and subsidizing education, many issues in educational accountability can be understood through the lens of a classic principal–agent problem. Policymakers, parents, and students wish to contract with schools to provide education. However, the provision of education requires the system insiders to make an array of decisions and budgetary choices, about hiring, discipline, tenure, curriculum, pedagogy, pay and benefits, grading and exams, and class sizes. It is difficult for interested outsiders to monitor the actions of schools and universities on these and other dimensions. The hope of greater educational accountability is that it will pressure the insiders in schools and universities to alter their production decisions and to improve in some key areas.

When questions arise about improving accountability, an economist's first instinct is often to ask why "the market" cannot provide sufficient accountability among providers. However, as economists have long recognized, education is an industry where the power of consumers to ensure quality by choosing among alternatives is often quite limited: the range of school choices is constrained by political jurisdictions and geography; direct public provision of educational services is widespread, often financed either completely (in the case of K–12 education) or substantially (in the case of higher education) by tax revenues; and possibilities for entry and exit are limited.

In the case of K–12 education, as pointed out by Milton Friedman (1962) more than 50 years ago, the institutional structure does not facilitate competition. Choice among K–12 public schools largely operates indirectly through choice of residential location (for example, Hoxby 2003), although in some places students may have some access to choosing public charter schools or schools in neighboring jurisdictions. Public schools are funded mostly by property taxes.

At first glance, the scope for consumer choice to drive accountability appears much more promising in higher education. Selective colleges and universities compete fiercely for the best students in a nationwide market, and these schools are most often the focus of public discussion. However, the vast majority of US college students attend nonselective and mostly public institutions that are close to home (Hoxby 2009). While public colleges in the United States receive considerable lower levels of state appropriations than once had been the case, many are still heavily subsidized through state legislative appropriations, and most still charge only a fraction of the true per-student cost (Winston 1999). Private colleges are also subsidized through the tax deductibility of charitable contributions, and many of the less-selective private institutions are heavily dependent on federal financial aid subsidies that students bring with them. As a result, the market for higher education—with the possible exception of elite colleges—is probably not very competitive.

The difficulty of fully monitoring actors within the educational system combined with the limited scope for external incentives through consumer choice probably justifies some form of accountability for educational institutions. This can take a variety of forms, from increased information provision and disclosure requirements to more heavy-handed regulation and incentive structures. The hard questions involve figuring out how many institutions are really in need of remediation, as well as the specific design details, including practical and political constraints.

Approaches to Accountability in Education

Educational accountability begins with collecting consistent information on specific outcomes and inputs of interest over time. This information can be used in two broad ways. A first approach, *called report-card accountability*, makes certain information public, but without other explicit stakes. This approach is the norm in many countries. As a consequence of the No Child Left Behind Act of 2001, nearly every US state has developed school report cards with information on test performance and other outcomes by K–12 grade, subject, and student subgroup. The second approach is the use of rewards and sanctions to motivate increased performance—what Hanushek and Raymond (2005) call *consequential accountability*. This means attaching rewards and sanctions to benchmarks, such as the percent of students meeting the proficiency standard on a mathematics test, or the rate of return on investment in a college degree.

The most controversial elements of the No Child Left Behind Act of 2001 were a set of escalating sanctions for repeated failure to meet achievement benchmarks. In the first year a school failed to make "adequate yearly progress," it was required to develop a school improvement plan. Repeated failure led to more severe consequences, beginning with providing students with a transfer option and ending with closure or conversion into a charter or private school.

Accountability systems can be designed with either stronger or weaker linkages between outcomes and incentives or consequences. Some argue that accountability systems with low stakes for educators will not induce them to improve educational practice, and push for strong consequences associated with measured performance. However, the problem with high-stakes accountability is that the objective metrics are typically incomplete descriptions of performance. Schools are trying to accomplish many objectives—higher student achievement on certain tests, but also achievement in those areas that may not be well-captured by performance on standardized tests, performance in other academic areas that do not appear on the accountability test, and more abstract goals such as critical thinking, open-mindedness, maturity, and citizenship. When faced with strong incentives to concentrate on some metrics but not on others, schools might be expected to focus on short-run gains in what is being measured—sometimes obtained through strategic behavior such as "teaching to the test"—at the expense of long-run skill acquisition. Moreover, even low-stakes accountability systems that are based exclusively on information can have high stakes for educators if stakeholders respond to that information (Figlio and Lucas 2004; Figlio and Kenny 2009).

With this conceptual framework in mind, school accountability has often been studied as an application of a multitask moral hazard model where performance measures are used in place of a true objective that cannot be observed directly (Holmstrom and Milgrom 1991; Baker 1992; MacLeod 2003). In the Holmstrom and Milgrom (1991) theoretical analysis of these models, they use the example of teachers teaching basic skills and higher-order thinking, where the latter cannot be measured. The key insight from these models is that the optimal strength of performance incentives is increasing in the correlation between a performance metric (say, high-stakes tests) and the true objective (say, developing broader capacities, or perhaps earnings). Crucially, it is the correlation *at the margin* that matters (Hout and Elliott 2011). When schools face pressure to raise test scores, and they take action, what is the effect of those actions on the long-run outcomes that are the true objective of schooling? When the correlation between test score gains and gains in long-run outcomes is weak, low-powered incentives, or even no incentives at all may be preferable.

Moreover, an inherent tension also arises between using achievement tests both as a diagnostic tool and also as a high-stakes performance measure (Neal 2013). This follows from what is known colloquially as Campbell's law (1976)—"the more any quantitative social science indicator is used for social decision-making, the more subject it will be to corruption pressures and the more apt it will be to distort and corrupt the social processes it is intended to monitor."

This discussion suggests that as long as what is being measured is only a proxy for the truly desired outcome, the effects of accountability efforts are theoretically ambiguous. Raising the stakes magnifies the impact of accountability on behavior, but whether students are helped or harmed overall by the changes in behavior is ultimately an empirical question.

Evidence on Accountability in K–12 Education

Reactions to Accountability

The evidence suggests that families respond strongly to provision of information about the K–12 education system, even in the absence of explicit stakes. For example, differences in test scores are capitalized into housing markets. In an early study of this effect, Black (1999) looked at Massachusetts houses that were located close to the boundary between school districts and found that homeowners were willing to pay 2.5 percent more for a 5 percent increase in test scores. For a literature review of later studies finding a broadly similar result, see Black and Machin (2011). Using a similar approach of looking at house location and school zone boundaries with Florida data, Figlio and Lucas (2004) find that the grades given to schools by the state of Florida also affect real estate values (and the effect of grades is in addition to the effect of any change in test scores). In another study of Florida data, Figlio and Kenny (2009) took advantage of an administrative shift in how Florida graded its schools to show that voluntary contributions to schools (typically made through parent–teacher organizations) rise and fall based on accountability measures. Hastings and Weinstein (2008) show that when parents are making choices between schools in the Charlotte-Mecklenburg school district, when they receive new information about test scores in schools, they are more likely to choose the schools with higher test scores.

Those who work in schools respond to accountability ratings, too. Several case studies have found that principals and teachers perceive that their job security is tied to their school's accountability rating: for example, Toenjes and Garst (2000) discuss this connection for Texas schools and districts; the collection of essays in Evers and Walberg (2002) includes a comparison of accountability systems in Texas, Florida, and California; Lemons, Luschei, and Siskin (2003) conduct case studies of six high schools in two states; and Mintrop and Trujillo (2005) compare three medium-sized states (Kentucky, Maryland, and North Carolina), four larger ones (California, Florida, New York, and Texas), and two large districts (Chicago and Philadelphia).

State and federal accountability policies typically combine information provision with performance incentives in a variety of ways, making it difficult to distinguish one from the other. However, a lot of evidence suggests that K–12 public schools do respond to accountability pressure. Several studies of state and federal accountability systems have found modest positive impacts on low-stakes test scores in reading and math, including in single-state or single-district studies that use differences in program rules for identification (Chiang 2009, Figlio and Rouse 2006, and Greene, Winters, and Forster 2004, all in Florida; Rockoff and Turner 2010, in New York City; Jacob 2005, in Chicago; and Ladd 1999, in Dallas) and overseas (Allen and Burgess 2012, in the United Kingdom). These studies tend to find larger effects for math than for reading, though in some cases (for example, Jacob 2005) the estimated effects are larger for reading. A typical finding suggests that accountability boosts math test scores in the lowest-performing schools by about one-tenth of a standard deviation relative to those in higher-performing schools, though the studies exhibit a wide range of impacts.

Researchers have tended to focus on low-stakes assessments because of concerns that gains on high-stakes tests reflect strategic responses to accountability pressures rather than genuine improvements. Winters, Trivitt, and Greene (2010) demonstrate that Florida's accountability system improved outcomes not just in math and reading, but also had spillover effects into science, even though there were no stakes attached to science performance at the time. Rouse, Hannaway, Goldhaber, and Figlio (2013) find that schools facing accountability pressure do change their practice: for example, such schools reorganize the school day and the learning environment to focus on low-performing students and lengthen the amount of instruction time, while also increasing resources available to teachers. They show that these changes in policies and practices account for a substantial fraction of the test score improvement. Dee and Jacob (2011) find that practices installed as a result of the No Child Left Behind legislation increased student achievement on the low-stakes National Assessment of Educational Progress.

We are aware of only one study that investigates the impact of K–12 school accountability on long-run outcomes. Deming, Cohodes, Jennings, and Jencks (forthcoming) find that accountability pressure in Texas high schools led to increases in college attainment and earnings for low-scoring students in low-scoring schools. However, they also found some evidence of *negative* impacts for other students, which they argue arose from schools' strategic responses to the rules around student testing exemptions. As we discuss in the next section, there are many examples of so-called "strategic" responses to school accountability. This foreshadows a key design challenge for accountability in US higher education—schools typically respond strongly to performance incentives, but not always in the ways that the advocates of such incentives would like.

Critiques of Accountability: Unclear Information and Strategic Responses

One criticism of "report card" accountability is that governments are not very good at providing information in an easily digestible format. Yet here, for better or worse, the private market has already stepped in. Websites such as greatschools.org, schoolgrades.org, k12.niche.com, and schooldigger.com make their living by translating publicly available information about school quality into user-friendly formats. However, a more fundamental question is the extent to which this information captures school quality accurately or comprehensively. Public provision of inaccurate or noisy measures of quality cannot be expected to improve student outcomes in any meaningful way.

Economists often focus on wages as an omnibus measure of impact. But drawing causal connections between test scores (or graduation rates) today and wages in the future is extremely difficult. A broader concern, however, is that school quality is subjective and multidimensional. As a result, the benchmark measures chosen for K–12 or postsecondary accountability, such as levels or gains or pass rates on certain tests, are always going to be incomplete proxies for the overall goals of education.

In this situation, one should be concerned that reducing the accounting of school quality to a small number of conveniently measurable outcomes narrows the focus of actors within an educational system. Indeed many studies have found that schools facing accountability pressure narrow their curriculum or instructional practices at the expense of nontested groups or subjects (Stecher, Barron, Chun, and Ross 2000; Diamond and Spillane 2004; Booher-Jennings 2005; Hamilton, Berends, and Stecher 2005; Diamond 2007; Ladd and Lauen 2010; Reback 2008; Neal and Schanzenbach 2010; Özek 2012; Reback, Rockoff, and Schwartz 2014).

Narrowing the curriculum is not necessarily a bad thing. Yet some strategic responses to accountability are harder to justify. Figlio and Winicki (2005) demonstrate that Virginia schools subject to accountability pressure strategically raise the calorie content of meals on test days, and find suggestive evidence that schools that use this approach see a larger rise in high-stakes pass rates. The fact that schools react in this manner, and that their reactions lead to improved measured outcomes, shows one way in which the reporting of measures of school quality can help to undermine their validity. Other studies have suggested responses that are potentially more insidious, such as strategic reclassification of students into disability categories (Deere and Strayer 2001; Cullen and Reback 2006; Figlio and Getzler 2006; Deming et al. forthcoming), using disciplinary procedures to suspend low-performing students from school when the tests are given (Figlio 2006), and outright teacher cheating (Jacob and Levitt 2003).

Lessons about Accountability from the K–12 Experience

Four lessons emerge from the existing literature on performance measurement and accountability in K–12 education.[1]

First, when public reporting and rewards and sanctions are tied to specific measures, as is often the case, organizations will seek to maximize short-run performance on those measures at the potential expense of other outcomes of interest.

Second, design details of accountability metrics strongly influence organizational behavior. For example, consider the difference that arises if an assessment measure is based on the share of students who exceed a certain proficiency target, or if it is based on the value-added gains that students made from their scores in the previous year. When K–12 schools are assessed based on proficiency targets, they have a strong incentive to focus on "bubble" students who are near the threshold, or on other students more likely to "count" for accountability (Neal 2010; Neal and Schanzenbach 2010; Figlio and Loeb 2011; Özek 2012; Figlio and Ladd 2015).

While value-added metrics reduce incentives to focus on particular students (after all, it is much harder to target students with high potential *gains*), they may introduce additional scope for distortions. For example, Macartney (2016) finds that schools and teachers in North Carolina responded to value-added performance targets by *reducing* effort in earlier periods—the so-called "ratchet effect." In addition, yearly fluctuation in test scores may make value-added metrics quite noisy and difficult for families to understand and use (Kane and Staiger 2002; Chay, McEwan, and Urquiola 2005). Indeed, Brehm, Imberman, and Lovenheim (2015) show using data from the Houston school district that student test score gains were so noisy that incentives based on value-added measures failed to elicit any performance response from teachers.

Third, the scope for strategic responses to accountability increases with the number and complexity of high-stakes metrics. A typical K–12 accountability system employs tests in multiple subjects, across multiple (but not all) grades and student subgroups, and includes a frequently byzantine set of exemptions and secondary metrics and assessments. Every additional layer of complexity introduces more opportunities for strategic behavior.

[1] Similar lessons arise from evidence on performance incentives in other settings. Institutions do respond to the information embedded in accountability systems, but not always in socially desired ways. For examples of some positive reactions, see Jin and Leslie (2003) on the reduction in food-related hospitalizations after Los Angeles County posted restaurant hygiene grade cards, and Bennear and Olmstead (2008) on how utilities that were required to disclose customer confidence reports reduced their health violations. For examples of some accountability systems with mixed results, Heckman, Heinrich, and Smith (1997, 2002) analyze performance standards for Job Training Partnership Act (JTPA) centers; Cutler, Huckman, and Landrum (2004) show that cardiac surgery report cards in New York led to patient selection and subsequent improvement of poorly performing hospitals, although Dranove, Kesser, McClellan, and Satterthwaite (2003) find that these report cards (and those in Pennsylvania) also resulted in higher levels of resource use and reduced health outcomes; and Lu (2012) demonstrates that the Nursing Home Quality Initiative led to improvements in reported measures of quality but deterioration in unreported areas.

Fourth, accountability works best at improving results at the bottom of the distribution. The balance of the evidence from studies of accountability, either across states or within them, suggests larger gains for low-income, minority, and low-achieving students (Carnoy and Loeb 2002; Dee and Jacob 2011; Lauen and Gaddis 2012; Deming et al. forthcoming). Additionally, many of the studies that find positive impacts of accountability pressure compare schools on either side of a cutoff that defines a "failing" grade (Figlio and Rouse 2006; Chiang 2009; Allen and Burgess 2012; Rouse et al. 2013; Reback, Rockoff, and Schwartz 2014). These schools are by definition among the lowest-achieving in the state.[2]

Researchers rarely find much of a response to accountability pressure in higher-performing schools. This may be because the lowest performance threshold is the most salient to educators and households, and confers the greatest stigma. Alternatively, families with higher socioeconomic status may monitor schools more closely, leading them to rely less on the public signal sent by an accountability rating. Thus, external accountability works best when institutions would not otherwise face strong internal or community pressures to improve. Families can hold schools accountable by monitoring school performance, and such monitoring may be more intense among affluent households (Ferreyra and Liang 2012). Moreover, affluent families may be more likely to sort across neighborhoods in response to perceived changes in school quality, which also places accountability pressure on schools (Bayer, Ferreira, and McMillan 2007). In contrast, schools serving disadvantaged populations may face less parental pressure or lack the capacity for self-monitoring, making external accountability relatively more important.

Accountability in US Higher Education

In September 2015, the US government released its "College Scorecard," which includes data from the US Department of the Treasury and Internal Revenue Service, as well as US Department of Education records covering over 7,000 colleges and universities (available at http://collegescorecard.ed.gov). The College Scorecard includes information on average college costs, overall and by family income levels; typical student debt loads, fraction of students receiving federal loans, and fraction of students making good progress in paying down their debt; graduation and one-year retention rates; median earnings of students ten years after entering the college; as well as student body characteristics (racial/ethnic breakdown, socioeconomic diversity, and college entry exam scores). These data provide, for the first time, a remarkable wealth of information on not just student body

[2] Sometimes, perhaps due to incentives to boost performance of marginal students, the lowest-performing students in low-achieving schools do not appear to benefit from accountability, even when more marginal students in their low-performing schools do (for example, Deming forthcoming, in Texas), perhaps due to the types of strategic behaviors such as focusing on "bubble kids" (Neal and Schanzenbach 2010) described above.

characteristics—a mainstay of private college ratings—but also some downstream outcomes of colleges.

Making the College Scorecard data available is a public service. However, the variables that are included and excluded send signals to the general public about what is valued and what is not, which in turn raises underlying questions: What are the desired outcomes of postsecondary education? And what are the likely outcomes of making data available and/or constructing rankings based on postsecondary data? While the federal government assiduously avoided constructing explicit college ratings with these data, it is certainly possible to construct ratings using them, and numerous organizations have done so.

An array of studies have found that higher education institutions respond strategically to privately produced institutional rankings, most notably those produced by *U.S. News and World Report*. For example, Monks and Ehrenberg (1999) look at how year-to-year changes in USNWR rankings from the late 1980s through the 1990s influenced the admissions outcomes and pricing policies at selective colleges, while Meredith (2004) finds similar results while studying a broader number of schools. Several studies suggest that the visibility and salience of the rankings are important, even aside from their quality. For example, Luca and Smith (2013) find that a one-rank improvement in the USNWR rankings leads to a one-percentage-point increase in applications. However, this effect only appears when the ranked institutions are listed numerically, and disappears when they are listed alphabetically. Similarly, Bowman and Bastedo (2009), find that being on the front page of the USNWR rankings, or not, has an effect on admissions. Bastedo and Bowman (2010) find that the perceptions of quality expressed by senior administrators at peer institutions are affected by the USNWR ratings, and Espeland and Sauder (2007) document with interview data how the rankings cause laws schools to change their behavior and expectations.

The Case for Information without Rankings

Although students and university administrators undoubtedly pay attention to college rankings, the impact of the rankings on institutional behavior, prices, and student outcomes is much less clear. In some ways, this combination of information without consequential accountability may be a healthy situation. Compared to their K–12 counterparts, US colleges and universities have broader purposes and serve a greater variety of students. This diversity across institutions greatly increases the degree of difficulty in designing an effective accountability system, because the benchmarks are harder to agree upon and the scope for strategic responses is much greater.

While K–12 schools are mostly required to take all comers, postsecondary institutions choose which students to admit—and even open enrollment institutions engage in subtle forms of selection. Public K–12 institutions have clear and well-defined missions and offer a "standard" curriculum. In contrast, higher education institutions decide which programs to offer and differ greatly in their stated institutional missions. Colleges and universities also operate in very different markets, ranging from open-access community colleges with a mandate to serve the local economy to elite institutions that compete for the best students on a global scale.

For these reasons, trying to rank colleges and universities on a few common standards may not make sense. To give just one example, the College Scorecard lists both Boston University and the New England Conservatory of Music as having an average annual cost (defined as the average price net of all financial aid) of around $35,000. Yet the average salary for Boston University graduates ten years later is more than double ($60,600 vs. $29,500). Are we comfortable rating colleges according to a financial benefit–cost calculation that will undoubtedly penalize students who self-select into lower-earning fields of study?

One potential consequence of a college rating system is that selective institutions might become even more stratified. MacLeod and Urquiola (2015) show that reputational incentives lead to stratification even in the absence of direct peer effects. In their model, ability is imperfectly observed and employers use college reputation (defined as the average skill of its graduates) as a signal of worker skill. As a result, students endogenously prefer better peers because of the signal that college reputation sends to the market. Importantly, stratification increases study effort prior to college admission and reduces study effort *in college*. Thus by selecting high-ability students at entry, colleges can have "elite" reputations without necessarily having high value-added to their graduates.

Between the difficulties of ranking institutions of higher education and the expectation that sorting will occur between them, an option is to forgo explicit stakes altogether, focusing instead on providing transparent and easily digestible information about school characteristics and performance, and letting consumers use it as they see fit. This is the College Scorecard approach, and it has much appeal. However, one lesson from K–12 accountability is that information alone can be a powerful driver of decision-making. Thus even if we decide that "report card" accountability for higher education is sufficient, we must think carefully about what information to provide, and in what way.

Targets and Tradeoffs for Accountability: Graduation Rates, Debt, Employment Outcomes, Direct Exams

In thinking about the impact of consequential accountability for US higher education, it is useful to start first with the variables actually used by the College Scorecard.

First, suppose that institutional ratings were based on graduation rates, borrowing rates, default rates, and/or borrowing intensity. These variables can, of course, be affected by institutional quality or policy decisions such as generosity of financial aid. But measured "success" in such a situation is surely also determined by student quality at admission. Institutions that serve primarily disadvantaged or first-generation college students will—all else equal—have lower graduation rates and higher borrowing rates. As a consequence, institutions that have missions to educate larger numbers of first-generation and disadvantaged students will look less attractive by these criteria. Colleges might alter admission criteria in order to improve the likelihood that admitted students will be able to succeed and pay for college without much borrowing.

In principle, one could "risk adjust" performance standards to reflect pre-existing differences in the likelihood of student success. This can reduce some aspects of selection, but as with "value-added" approaches in K–12 education, risk adjustment can also increase measurement error and reduce transparency and public confidence in the rating system. Barnow and Heinrich (2010) discuss the benefits and costs of such risk adjustment in a variety of settings. In higher education, there are large differences in students' prior preparation even within open access institutions (Kurlaender, Carrell, and Jackson 2016). As a result, failure to adjust for differential selection could be quite problematic.

Moreover, the data reported in the College Scorecard are calculated based on different populations for each outcome. For instance, while college graduation rates are calculated for all students, average costs of attendance and subsequent earnings are calculated only among federal financial aid recipients. Thus, outcome data are missing for some types of students and not others, making some type of risk adjustment extremely important for apples-to-apples comparisons across institutions.

Evaluating institutions on the basis of employment and earnings outcomes involves many of the same complications as other performance metrics. However, an additional complication comes from the wide variety in average compensation by field of study. Four-year college graduates with the highest-paying majors earn two-and-a-half times on average what the four-year college graduates with the lowest-paying majors earn (Hershbein and Kearney 2014). Majors that prepare students to work with children (like early childhood education and elementary education) or provide community and counseling services (like family sciences, social work, and theology) have the lowest average earnings. Evaluating institutions on one dimension like earnings could lead to reductions in opportunities to prepare for fields that are socially desirable but not financially lucrative; this is one example of how accountability can exacerbate the multitasking problem in higher education. In addition, Oreopoulos and Salvanes (2011) document a number of nonpecuniary benefits of postsecondary study that are not captured by labor market outcomes. It is not difficult to imagine that some colleges provide great value-added for "nonmarket outcomes" that do not show up on the balance sheet.

Another limitation of using employment outcomes for accountability is the long time horizon required to measure post-college earnings. The College Scorecard measures earnings ten years after initial enrollment. If colleges are evaluated based on earnings in a student's late 20s, a relatively easy way to "game the system" is to emphasize fields of study where early career earnings are high and graduate education is uncommon—the Scorecard excludes individuals known to be enrolled in school at the point of measurement so graduate students ten years out don't help the school's ratings—or by counseling students into higher-earning options. On the other hand, a substantially longer time horizon means that the information on earnings is what happened to those who enrolled more than a decade earlier, and colleges are unlikely to alter their behavior based on predictions of outcomes far in the future.

A final possibility is to hold postsecondary institutions accountable for learning outcomes directly, using assessments such as the National Survey of Student Engagement or the Collegiate Learning Assessment (for example, Arum and Roksa 2010). This approach presents many opportunities for institutional strategic behavior observed at the K–12 level, both in terms of emphasizing the types of skills that are more likely to be represented on the assessment as well as in terms of selecting which students enroll and actually take the assessment. Moreover, while tools like the Collegiate Learning Assessment are surely valuable indicators of one aspect of student learning growth over the course, they do not reflect the wide range of objectives of postsecondary institutions or the considerable heterogeneity in the purposes of these institutions. It is possible to construct field-specific exit exams that would at least present the opportunity to capture one aspect of skill; MacLeod, Riehl, Saavedra, and Urquiola (2015) present evidence from Colombia that the rollout of a field-specific college exit exam reduced some of the labor market returns to college reputation.[3] But carrying out this form of exit exam is extremely expensive, and the introduction of such an exam brings with it the risk of new manipulative behaviors on the part of educational institutions, along with the challenges associated with measuring institutional value added, which are surely more difficult in the postsecondary setting.

Some Design Principles for Accountability in Higher Education

Some of the adjustments and tradeoffs from greater accountability in higher education may be welcome. After all, if students who are undecided about majors get a nudge toward a choice that pays better, or if schools put more emphasis on a high graduation rate and a lower debt burden, such steps may overall be beneficial. Indeed since all college students are paying customers—both directly and indirectly through the opportunity cost of foregone immediate earnings—we might expect accountability to have a larger impact in higher education than in K–12. Here, we draw on insights from economic theory and from lessons learned in K–12 education to lay out some design principles for accountability in US higher education.

A first design principle is that a college rating system should be kept as simple as possible to reduce the scope for strategic responses. While some risk adjustment is probably necessary, it should be as transparent as possible to facilitate consumer choice. Rather than constructing college "value added" through regression adjustment, a simpler alternative is to construct groups of postsecondary institutions that represent likely choice sets for certain types of students. Equivalence classes could be created based on geographical proximity and measures of selectivity, or alternatively they could be constructed empirically using overlap in actual students' choice sets (for example, Avery, Glickman, Hoxby, and Metrick 2012).

A second design principle is to target the postsecondary institutions that are least likely to respond to market forces in the absence of accountability. In the K–12

[3] Hoekstra (2009) and others demonstrate that there exists a labor market return to higher education institutional reputation in the United States.

setting, most of the benefits of accountability come from pressure on educators to avoid a failing grade, perhaps because families with higher socioeconomic status monitor schools more closely and are more likely to "vote with their feet." Likewise, elite colleges already compete fiercely for students, and a government rating is unlikely to change their incentives much. Thus one idea is to focus on certifying a minimum standard of quality, rather than assigning grades or ratings to institutions all along the spectrum. Similar to health inspections or the consumer drug approval process, the job of a higher education accountability system could be to certify that schools are good enough to receive public support.

Public certification of postsecondary institutions already exists in the form of accreditation. The US Department of Education keeps a list of regional and national accreditors, and in principle institutions must be approved by an accreditor's regular inspections to distribute federal financial aid. Yet in practice, accreditors—who are paid by the institutions themselves—appear to be ineffectual at best, much like the role of credit rating agencies during the recent financial crisis. As a case in point, the Accrediting Council for Independent Colleges and Schools (ACICS) has come under considerable scrutiny for continuing to accredit branches of Corinthian Colleges right up until the company's collapse in April 2015 amid allegations of fraud and financial misconduct.

While we are unaware of well-identified studies of the consequences of independent accreditation in the higher education sector, Hussain (2015) demonstrates that in the K–12 sector in the United Kingdom, inspectorate systems led to measurable and lasting improvements in student outcomes. One possible approach is to design an inspectorate system that is "turned on" when an institution falls below quantitative benchmarks. While school inspections are resource-intensive, targeting toward the lowest performers would help to limit the cost of such a program. Duflo, Greenstone, Pande, and Ryan (2013) present experimental evidence from environmental inspections of industrial plants in India. They find that regulation works much better when auditors are randomly assigned and centrally compensated, rather than chosen and paid by firms themselves. Similar reforms to accreditation might have sizeable benefits in terms of improved higher education outcomes.

Another way that higher education accountability can target the lowest performing institutions is by setting a relatively low bar for performance yet enforcing it vigorously. The federal Gainful Employment regulations that went into effect in 2015 are one—albeit imperfect—example. The purpose of the Gainful Employment regulations is to link the costs and benefits of postsecondary programs explicitly (US Department of Education 2015), and while Gainful Employment regulates on debt burden alone, it is still a step in this direction. The rules specify that, on average, graduates of nearly all for-profit programs (along with certificate programs at not-for-profit and public institutions) must have an annual loan payment that does not exceed 20 percent of discretionary income or 8 percent of total earnings. The penalty for repeatedly falling below this debt-to-earnings threshold is the withdrawal of eligibility for that institution to disburse federal Title IV financial aid.

This bar seems relatively low, yet the Department of Education estimated that 840,000 students were enrolled in programs that would not have met the standard in 2013. While this represents fewer than 5 percent of all US postsecondary enrollment, more than 99 percent of the students were concentrated within a small number of for-profit programs. The combination of high prices and low labor market returns is unique to the for-profit higher education sector, making it a prime target for increased accountability. Looney and Yannelis (2015) show that for-profit institutions are responsible for a disproportionate share of the increase in student debt and loan defaults since 2000, and Deming et al. (2016) use a resume audit experiment to show that employers are less likely to express interest in a resume with a degree from a for-profit institution of higher education compared to identical resumes with degrees from public institutions.

The design of Gainful Employment is simple and straightforward, and the regulation successfully concentrates on the worst offenders. However, a legitimate criticism is that it unfairly targets the for-profit sector and leaves poorly performing public institutions untouched.[4]

Ideally, regulations like Gainful Employment would focus on institutions that rely heavily on public subsidies—regardless of their for-profit or public status. At present, for-profit institutions derive about 75 percent of their revenue from federal Title IV Pell Grants and Stafford Loans, which are disbursed to eligible students based on financial need (Deming, Goldin, and Katz 2012). A federal regulation known as the 90/10 rule prohibits these colleges from deriving more than 90 percent of revenue from Title IV aid. Yet the largest for-profit colleges bump right up against this 90 percent cap.[5] This dependence on taxpayer largesse, more than for-profit status, justifies tighter regulation. Many smaller for-profit institutions attract paying customers without needing federal financial aid subsidies, and these schools are rightly free from the Gainful Employment regulations (for example, Cellini and Goldin 2014).

In most states, community colleges and less-selective four-year publics are also heavily subsidized by taxpayers. Tuition is kept much lower than the resource cost of college, and is in some cases close to zero after accounting for federal financial aid.

[4] The Gainful Employment program focuses only on for-profits and certificate programs in nonprofit and public institutions. This targeting was partly a regulatory necessity—the phrase "gainful employment" originates from language in the Higher Education Act of 1965 that specifies which institutions are allowed to distribute Title IV aid—but was also deliberately aimed at the for-profit sector. For-profits have criticized the Gainful Employment regulations for unfairly targeting the sector. In 2012, a for-profit college trade group sued, and the initial regulation from the US Department of Education, which included a rule about loan repayment rates in addition to the debt-to-earnings ratio, was struck down in federal district court. The follow-up effort, which eliminated repayment rates as an accountability metric, was upheld in May 2015. This court decision set the stage for Gainful Employment rules to become law in July 2015, although the ruling remains legally tenuous.

[5] In addition, the large for-profits engage in strategic behavior that seems aimed at maximizing loan support, such as targeted recruitment of GI bill-subsidized military students, Military students receive higher education subsidies from the GI bill, which is federal aid to students in higher education that does not fall under Title IV. Thus every $1 of GI bill subsidies allows schools to bring in another $9 of Title IV aid while remaining under the 90 percent revenue cap.

Schools that depend more on taxpayer support should justifiably be targeted with increased government regulation, even if they are public institutions. Concretely, one could design an accountability system where regulatory control is increasing in the share of institutional revenue that comes from public sources.

Thus a third principle, following from the discussion above, is that accountability in higher education should be designed to ensure that both students and postsecondary institutions have some "skin in the game." Here there is no exact parallel with K–12 schooling, because most college students and few primary and secondary school students pay for their education. However, "skin in the game" can be a guiding principle for regulators in thinking about how much control to exert over postsecondary institutions. Colleges that can attract full-paying customers—either out-of-state students or students who do not qualify for financial aid—have implicitly survived a market test and should be allowed to operate more freely. This principle does not mean that public institutions cannot be heavily subsidized, but it does suggest that scrutiny should be greater when taxpayers are footing more of the bill.

A more direct approach is risk-sharing, where institutions would be responsible for paying a share of student loans that subsequently end up in default. As with all accountability metrics, risk-sharing programs would probably lead to lower lending, but also with some potential for strategic responses. Institutions would be more likely to enroll students whom they suspect will stand the best chance of completing college and repaying their student loans. Institutions might also offer fewer programs in professions with high social value but low downstream income potential.

Conclusion

The rationale for increased accountability in the higher education sector is clear. However, designing a well-functioning accountability system is extremely difficult. The experience from accountability in K–12 education and other industries demonstrates that "what gets measured gets done," in both socially desirable and undesirable ways. Also, figuring out what *should* get measured and what *should* get done is no easy matter. The outcomes desired by parents and students might differ from the outcomes chosen by policymakers, and any one-size-fits-all solution will not do full justice to the multidimensional nature of higher education.

One main lesson we take from the research evidence is that accountability is likely to be most important in the education markets that are the least competitive. At the K–12 level, accountability works best in low-performing schools with weak systems of support, and when students have relatively few options other than their local public school. Similarly, we suspect that accountability for selective colleges will have little impact, because both elite colleges and the students who attend them already have plenty of "skin in the game." Just to be clear, this does not mean that we believe all students at such institutions receive a high-quality and cost-effective education. Rather, we believe that additional accountability measures aren't likely to lead to improvements among schools that are already facing other kinds of pressure, although they may respond strategically to improve their position in the ratings hierarchy.

This same logic also applies to less-selective institutions of higher education that can attract paying customers with only modest help from public funds or federal financial aid. In contrast, for postsecondary institutions that are heavily dependent on taxpayer support, or that have a poor record on metrics like graduation rates, an accountability system with explicit consequences could improve student outcomes.

If performance measures work, they will provoke a mix of real improvement and strategic responses. Thus, an ebb-and-flow of defining accountability, backing away, and then redefining accountability and backing away again, is to be expected. As this process evolves, the ongoing challenge is to maximize the benefits of accountability while minimizing its unintended side effects. In higher education, this may be the time for state-level policy experimentation: if different states try different forms of accountability for their higher education institutions and programs, we will have the opportunity to learn more about which approaches to accountability in the higher education sector yield the greatest net benefits.

■ *The authors are grateful to Mark Gertler, Claudia Goldin, Gordon Hanson, Larry Katz, Jordan Matsudaira, Enrico Moretti, Timothy Taylor, and Miguel Urquiola for helpful comments and suggestions, though all errors are their own.*

References

Allen, Rebecca, and Simon Burgess. 2012. "How Should We Treat Under-performing Schools? A Regression Discontinuity Analysis of School Inspections in England." Working Paper no. 12/287, Center for Market and Public Organization, University of Bristol.

Arum, Richard, and Josipa Roksa. 2010. *Academically Adrift: Limited Learning on College Campuses.* University of Chicago Press.

Avery, Christopher N., Mark E. Glickman, Caroline M. Hoxby, and Andrew Metrick. 2013. "A Revealed Preference Ranking of American Colleges and Universities." *Quarterly Journal of Economics* 128(1): 425–67.

Avery, Christopher, and Sarah Turner. 2012. "Student Loans: Do College Students Borrow Too Much—Or Not Enough?" *Journal of Economic Perspectives* 26(1): 165–92.

Baker, George P. 1992. "Incentive Contracts and Performance Measurement." *Journal of Political Economy* 100(3): 598–614.

Barnow, Burt S., and Carolyn J. Heinrich. 2010. "One Standard Fits All? The Pros and Cons of Performance Standard Adjustments." *Public Administration Review* 70(1): 60–71.

Bastedo, Michael N., and Nicholas A. Bowman. 2010. "*U.S. News and World Report* College Rankings: Modeling Institutional Effects on Organizational Reputation." *American Journal of Education* 116(2): 163–83.

Bayer, Patrick, Fernando Ferreira, and Robert McMillan. 2007. "A Unified Framework for

Measuring Preferences for Schools and Neighborhoods." *Journal of Political Economy* 115(4): 588–638.

Bennear, Lori S., and Sheila M. Olmstead. 2008. "The Impacts of the 'Right to Know': Information Disclosure and the Violation of Drinking Water Standards." *Journal of Environmental Economics and Management* 56(2): 117–30.

Black, Sandra E. 1999. "Do Better Schools Matter? Parental Valuation of Elementary Education." *Quarterly Journal of Economics* 114(2): 577–99.

Black, Sandra E., and Stephen Machin. 2011. "Housing Valuation and School Performance." Chap. 10 in *Handbook in Economics of Education*, vol. 3, edited by Eric Hanushek, Stephen Machin, and Ludger Woessmann. Elsevier.

Booher-Jennings, Jennifer. 2005. "Below the Bubble: 'Educational Triage' and the Texas Accountability System." *American Educational Research Journal* 42(2): 231–68.

Bowman, Nicholas A., and Michael N. Bastedo. 2009. "Getting on the Front Page: Organizational Reputation, Status Signals, and the Impact of *U.S. News and World Report* on Student Decisions." *Research in Higher Education* 50(5): 415–36.

Brehm, Margaret, Scott A. Imberman, and Michael F. Lovenheim. 2015. "Achievement Effects of Individual Performance Incentives in a Teacher Merit Pay Tournament." NBER Paper 21598.

Campbell, Donald T. 1976. "Assessing the Impact of Planned Social Change." Occasional Paper 8, December. (Reprinted February 2011 in the *Journal of MultiDisciplinary Evaluation*, vol. 7, no. 15.)

Carnoy, Martin, and Susanna Loeb. 2002. "Does External Accountability Affect Student Outcomes? A Cross-State Analysis." *Educational Evaluation and Policy Analysis* 24(4): 305–31.

Chay, Kenneth Y., Patrick J. McEwan, and Miguel Urquiola. 2005. "The Central Role of Noise in Evaluating Interventions That Use Test Score to Rank Schools." *American Economic Review* 95(4): 1237–58.

Cellini, Stephanie Riegg, and Claudia Goldin. 2014. "Does Federal Student Aid Raise Tuition? New Evidence on For-Profit Colleges." *American Economic Journal: Economic Policy* 6(4): 174–206.

Chiang, Hanley. 2009. "How Accountability Pressure on Failing Schools Affects Student Achievement." *Journal of Public Economics* 93(9–10): 1045–57.

Cullen, Julie Berry, and Randall Reback. 2006. "Tinkering toward Accolades: School Gaming under a Performance Accountability System." In *Advances in Applied Microeconomics*, vol. 14, edited by Timothy J. Gronberg and Dennis W. Jansen, 1–34. Elsevier.

Cutler, David M., Robert S. Huckman, and Mary Beth Landrum. 2004. "The Role of Information in Medical Markets: An Analysis of Publicly Reported Outcomes in Cardiac Surgery." *American Economic Review* 94(2): 342–46.

Dee, Thomas S., and Brian Jacob. 2011. "The Impact of No Child Left Behind on Student Achievement." *Journal of Policy Analysis and Management* 30(3): 418–46.

Deere, Donald, and Wayne Strayer. 2001. "Putting Schools to the Test: School Accountability, Incentives and Behavior." Unpublished paper, Texas A&M University.

Deming, David J., Sarah Cohodes, Jennifer Jennings, and Christopher Jencks. Forthcoming. "School Accountability, Postsecondary Attainment and Earnings." *Review of Economics and Statistics.*

Deming, David J., Claudia Goldin, and Lawrence F. Katz. 2012. "The For-Profit Postsecondary School Sector: Nimble Critters or Agile Predators?" *Journal of Economic Perspectives* 26(1): 139–64.

Deming, David J., Noam Yuchtman, Amira Abulafi, Claudia Goldin, and Lawrence F. Katz. 2016. "The Value of Postsecondary Credentials in the Labor Market: An Experimental Study." *American Economic Review* 106(3): 778–806.

Diamond, John B. 2007. "Where the Rubber Meets the Road: Rethinking the Connection between High-Stakes Testing Policy and Classroom Instruction." *Sociology of Education* 80(4): 285–313.

Diamond, John B., and James Spillane. 2004. "High-Stakes Accountability in Urban Elementary Schools: Challenging or Reproducing Inequality?" *Teachers College Record* 106(6): 1145–76.

Dranove, David, Daniel Kessler, Mark McClellan, and Mark Satterthwaite. 2003. "Is More Information Better? The Effects of 'Report Cards' on Health Care Providers." *Journal of Political Economy* 111(3): 555–88.

Duflo, Esther, Michael Greenstone, Rohini Pande, and Nicholas Ryan. 2013. "Truth-Telling by Third-Party Auditors and the Response of the Polluting Firms: Experimental Evidence from India." *Quarterly Journal of Economics* 128(4): 1449–98.

Espeland, Wendy Nelson, and Michael Sauder. 2007. "Rankings and Reactivity: How Public Measures Recreate Social Worlds." *American Journal of Sociology* 113(1): 1–40.

Evers, William, and Herbert Walberg. 2002. *School Accountability*. Palo Alto, CA: Hoover Press.

Ferreyra, Maria Marta, and Pierre Jinghong Liang. 2012. "Information Asymmetry and Equilibrium Monitoring in Education." *Journal of Public Economics* 96(1–2): 237–54.

Figlio, David N. 2006. "Testing, Crime and Punishment." *Journal of Public Economics* 90(4–5):

837–51.

Figlio, David N., and Lawrence S. Getzler. 2006. "Accountability, Ability and Disability: Gaming the System?" In *Advances in Applied Microeconomics*, Vol 14: *Improving School Accountability* edited by Timothy J. Gronberg and Dennis W. Jansen, 35–39. Emerald.

Figlio, David N, and Lawrence W. Kenny. 2009. "Public Sector Performance Measurement and Stakeholder Support." *Journal of Public Economics* 93(9–10): 1069–77.

Figlio, David N., and Helen F. Ladd. 2015. "School Accountability and Student Achievement." In *Handbook of Research in Education Finance and Policy*, 2nd edition, edited by Helen F. Ladd and Margaret Goertz, pp. 194–210. AEFP and Routledge.

Figlio, David, and Susanna Loeb. 2011. "School Accountability." In *Handbook in Economics of Education*, vol. 3, edited by Eric Hanushek, Stephen Machin, and Ludger Woessmann, 383–421. Elsevier.

Figlio, David N., and Maurice E. Lucas. 2004. "What's in a Grade? School Report Cards and the Housing Market." *American Economic Review* 94(3): 591–604.

Figlio, David N., and Cecilia Elena Rouse. 2006. "Do Accountability and Voucher Threats Improve Low-Performing Schools?" *Journal of Public Economics* 90(1–2): 239–55.

Figlio, David N., and Joshua Winicki. 2005. "Food for Thought: The Effects of School Accountability Plans on School Nutrition." *Journal of Public Economics* 89(2–3): 381–94.

Friedman, Milton. 1962. *Capitalism and Freedom*. University of Chicago Press.

Gardner, David P., et al. 1983. *A Nation at Risk*. Report of the National Commission on Excellence in Education, US Department of Education, Washington, DC. ERIC, Institute of Education Sciences.

Greene, Jay, Marcus Winters, and Greg Forster. 2004. "Testing High-Stakes Tests: Can We Believe the Results of Accountability Tests?" *Teachers College Record* 106(6): 1124–44.

Hamilton, Laura, Mark Berends, and Brian M. Stecher. 2005. "Teachers' Responses to Standards-Based Accountability." RAND Working Paper WR-259-EDU.

Hanushek, Eric A., and Margaret F. Raymond. 2005. "Does School Accountability Lead to Improved Student Performance?" *Journal of Policy Analysis and Management* 24(2): 297–327.

Hastings, Justine S., and Jeffrey M. Weinstein. 2008. "Information, School Choice, and Academic Achievement: Evidence from Two Experiments." *Quarterly Journal of Economics* 123(4): 1373–1414.

Heckman, James, Carolyn Heinrich, and Jeffrey Smith. 1997. "Assessing the Performance of Performance Standards in Public Bureaucracies." *American Economic Review* 87(2): 389–95.

Heckman, James J., Carolyn Heinrich, and Jeffrey Smith. 2002. "The Performance of Performance Standards." *Journal of Human Resources* 37(4): 778–811.

Hershbein, Brad, and Melissa Kearney. 2014. "Major Decisions: What Graduates Earn Over their Lifetimes." The Hamilton Project, Brooking Institution. http://www.hamiltonproject.org/papers/major_decisions_what_graduates_earn_over_their_lifetimes/.

Hoekstra, Mark. 2009. "The Effect of Attending the Flagship State University on Earnings: A Discontinuity-Based Approach." *Review of Economics and Statistics* 91(4): 717–24.

Holmstrom, Bengt, and Paul Milgrom. 1991. "Multitask Principal–Agent Analyses: Incentive Contracts, Asset Ownership, and Job Design." *Journal of Law, Economics, and Organization*, 7(Special issue: [Papers from the Conference on the New Science of Organization, January 1991]): 24–52.

Hout, Michael, and Stuart W. Elliott. 2011. *Incentives and Test-based Accountability in Education.* National Academies Press.

Hoxby, Caroline M. 2003. "School Choice and School Productivity (Or, Could School Choice Be a Rising Tide that Raises All Boats?)" Chap. 8 in *The Economics of School Choice*, edited by C. Hoxby. University of Chicago Press.

Hoxby, Caroline M. 2009. "The Changing Selectivity of American Colleges." *Journal of Economic Perspectives* 23(4): 95–118.

Hussain, Iftikhar. 2015. "Subjective Performance Evaluation in the Public Sector: Evidence from School Inspections." *Journal of Human Resources* 50(1): 189–221.

Jacob, Brian A. 2005. "Accountability, Incentives and Behavior: The Impact of High-Stakes Testing in the Chicago Public Schools." *Journal of Public Economics* 89(5–6): 761–96.

Jacob, Brian A., and Steven D. Levitt. 2003. "Rotten Apples: An Investigation of the Prevalence and Predictors of Teacher Cheating." *Quarterly Journal of Economics* 118(3): 843–77.

Jin, Ginger Zhe, and Phillip Leslie. 2003. "The Effect of Information on Product Quality: Evidence from Restaurant Hygiene Grade Cards." *Quarterly Journal of Economics* 118(2): 409–51.

Kane, Thomas J., and Douglas O. Staiger. 2002. "The Promise and Pitfalls of Using Imprecise School Accountability Measures." *Journal of Economic Perspectives* 16(4): 91–114.

Kurlaender, Michal, Scott Carrell, and Jacob Jackson. 2016. "The Promises and Pitfalls of

Measuring Community College Quality." *RSF: The Russell Sage Foundation Journal of the Social Sciences* 2(1): 174–90.

Ladd, Helen F. 1999. "The Dallas School Accountability and Incentive Program: An Evaluation of Its Impacts on Student Outcomes." *Economics of Education Review* 18(1): 1–16.

Ladd, Helen F., and Douglas L. Lauen. 2010. "Status versus Growth: The Distributional Effects of School Accountability Policies." *Journal of Policy Analysis and Management* 29(3): 426–50.

Lauen, Douglas Lee, and S. Michael Gaddis. 2012. "Shining a Light or Fumbling in the Dark? The Effects of NCLB's Subgroup-Specific Accountability on Student Achievement." *Educational Evaluation and Policy Analysis* 34(2): 185–208.

Lemons, Richard, Thomas F. Luschei, and Leslie Santee Siskin. 2003. "Leadership and the Demands for Standards-based Accountability." Chap. 4 in *The New Accountability: High Schools and High-Stakes Testing*, edited by Martin Carnoy, Richard Elmore, Leslie Santee Siskin, 99–128. RoutledgeFalmer.

Looney, Adam, and Constantine Yannelis. 2015. "A Crisis in Student Loans? How Changes in the Characteristics of Borrowers and in the Institutions They Attended Contributed to Rising Loan Defaults." Brookings Paper on Economic Activity, Fall 2015 Conference.

Lu, Susan Feng. 2012. "Multitasking, Information Disclosure, and Product Quality: Evidence from Nursing Homes." *Journal of Economics and Management Strategy* 21(3): 673–705.

Luca, Michael, and Jonathan Smith. 2013. "Salience in Quality Disclosure: Evidence from the U.S. News College Rankings." *Journal of Economics and Management Strategy* 22(1): 58–77.

Macartney, Hugh. 2016. "The Dynamic Effects of Educational Accountability." *Journal of Labor Economics* 34(1): 1–28.

MacLeod, W. Bentley. 2003. "Optimal Contracting with Subjective Evaluation." *American Economic Review* 93(1): 216–40.

MacLeod, W. Bentley, Evan Riehl, Juan E. Saavedra, and Miguel Urquiola. 2015. "The Big Sort: College Reputation and Labor Market Outcomes." NBER Working Paper 21230.

MacLeod, W. Bentley, and Miguel Urquiola. 2015. "Reputation and School Competition." *American Economic Review* 105(11): 3471–88.

Meredith, Marc. 2004. "Why Do Universities Compete in the Rankings Game? An Empirical Analysis of the Effects of the *US News and World Report* College Rankings." *Research in Higher Education* 45(5): 443–61.

Mintrop, Heinrich, and Tina Trujillo. 2005. "Corrective Action in Low Performing Schools: Lessons for NCLB Implementation from First-Generation Accountability Systems." *Education Policy Analysis Archives* 13(48).

Monks, James, and Ronald G. Ehrenberg. 1999. "The Impact of *U.S. News and World Report* College Rankings on Admissions Outcomes and Pricing Policies at Selective Private Institutions." NBER Working Paper 7227.

Murnane, Richard J. 2013. "US High School Graduation Rates: Patterns and Explanations." *Journal of Economic Literature* 51(2): 370–422.

Neal, Derek. 2010. "Aiming for Efficiency Rather than Proficiency." *Journal of Economic Perspectives* 24(3): 119–31.

Neal, Derek. 2013. "The Consequences of Using One Assessment System to Pursue Two Objectives." NBER Working Paper 19214.

Neal, Derek, and Diane Whitmore Schanzenbach. 2010. "Left Behind by Design: Proficiency Counts and Test-Based Accountability." *Review of Economics and Statistics* 92(2): 263–83.

Obama, Barack. 2015. "Weekly Address: A New College Scorecard." September 12. https://www.whitehouse.gov/the-press-office/2015/09/12/weekly-address-new-college-scorecard.

Oreopoulos, Philip, and Kjell G. Salvanes. 2011. "Priceless: The Nonpecuniary Benefits of Schooling." *Journal of Economic Perspectives* 25(1): 159–84.

Özek, Umut. 2012. "One Day Too Late? Mobile Students in an Era of Accountability." Working Paper 82, CALDER.

Reback, Randall. 2008. "Teaching to the Rating: School Accountability and the Distribution of Student Achievement." *Journal of Public Economics* 92(5–6): 1394–1415.

Reback, Randall, Jonah Rockoff, and Heather L. Schwartz. 2014. "Under Pressure: Job Security, Resource Allocation, and Productivity in Schools under NCLB." *American Economic Journal: Economic Policy* 6(3): 207–41.

Rockoff, Jonah, and Lesley J. Turner. 2010. "Short-Run Impacts of Accountability on School Quality." *American Economic Journal: Economic Policy* 2(4): 119–147.

Rouse, Cecilia Elena, Jane Hannaway, Dan Goldhaber, and David Figlio. 2013. "Feeling the Florida Heat? How Low-Performing Schools Respond to Voucher and Accountability Pressure." *American Economic Journal: Economic Policy* 5(2): 251–81.

Stecher, Brian M., Sheila Barron, Tammi Chun, and Karen E. Ross. 2000. *The Effects of the Washington State Education Reform on Schools and Classrooms.* Santa Monica, CA: RAND Corporation.

Toenjes, L. A., and Garst, J. E. 2000. "Identifying High Performing Texas Schools and School

Districts and their Methods of Success." Texas Education Agency.

US Department of Education. 2015. "Fact Sheet: Obama Administration Increases Accountability for Low-Performing For-Profit Institutions." July 1. http://www.ed.gov/news/press-releases/fact-sheet-obama-administration-increases-accountability-low-performing-profit-institutions.

White House. 2013. "Fact Sheet on the President's Plan to Make College More Affordable: A Better Bargain for the Middle Class." August 22. https://www.whitehouse.gov/the-press-office/2013/08/22/fact-sheet-president-s-plan-make-college-more-affordable-better-bargain-.

Winston, Gordon C. 1999. "Subsidies, Hierarchy and Peers: The Awkward Economics of Higher Education." *Journal of Economic Perspectives* 13(1): 13–36.

Winters, Marcus A., Julie R. Trivitt, and Jay P. Greene. 2010. "The Impact of High-Stakes Testing on Student Proficiency in Low-Stakes Subjects: Evidence from Florida's Elementary Science Exam." *Economics of Education Review* 29(1): 138–46.

Journal of Economic Perspectives—Volume 30, Number 3—Summer 2016—Pages 57–84

What Can We Learn from Charter School Lotteries?

Julia Chabrier, Sarah Cohodes, and
Philip Oreopoulos

Publicly funded charter schools, which set their own curriculum, financial management, and staffing, were originally designed as testing grounds for trying out new and innovative approaches for improving academic achievement. From the first few charter schools started in Minnesota in 1993 with a few dozen students, enrollment has increased to about three million across 7,000 schools (National Center for Education Statistics 2015), which is more than 5 percent of all public elementary and secondary students in the country. In some large urban districts, like Indianapolis, Philadelphia, Detroit, and Washington, DC, more than 30 percent of students attend charter schools. In the 2014–2015 school year, the New Orleans Recovery School District became the first US district to be comprised entirely of charter schools (National Alliance for Public Charter Schools 2015a; Abdulkadiroğlu, Angrist, Hull, and Pathak 2016).

All charter schools are free to students. Anyone residing in a given geography (which, depending on state law, would be the district, region, or state where the charter school is located) is eligible to attend. Increasingly, however, applicants

■ *Julia Chabrier is a Policy Manager, Abdul Latif Jameel Poverty Action Lab (J-PAL), Massachusetts Institute of Technology, Cambridge, Massachusetts. Sarah Cohodes is Assistant Professor of Education and Public Policy, Teachers College, Columbia University, New York City, New York. Philip Oreopoulos is Professor of Economics, University of Toronto, Toronto, Canada. He is also Faculty Research Associate, National Bureau of Economic Research, Cambridge, Massachusetts, and Co-director, Canadian Institute for Advanced Research, Toronto, Canada. Their email addresses are jchabrier@povertyactionlab.org, cohodes@tc.columbia.edu, and philip.oreopoulos@utoronto.ca.*

† For supplementary materials such as appendices, datasets, and author disclosure statements, see the article page at
http://dx.doi.org/10.1257/jep.30.3.57 doi=10.1257/jep.30.3.57

exceed the spots available. When faced with too many applicants, charters must admit students by lottery. Systematic evidence on what share of charters are oversubscribed is scant, but the authors of a national evaluation of charter school impacts estimated that about 26 percent of charter middle schools were likely to be oversubscribed in the 2006–2007 school year (Gleason, Clark, Clark Tuttle, Dwoyer, and Silverberg 2010; see also Clark Tuttle, Gleason, and Clark 2012). However, in disadvantaged urban neighborhoods, some charter schools admit fewer than 20 percent of the applicants. Lotteries are sometimes held in large auditoriums in front of anxious parents and children, leading to heartbreaking scenes of disappointment like those in the 2010 documentary, *Waiting for Superman*. Lottery losers often must default back to attending some of the worst performing schools in the country.[1] To remove the incentive for parents to apply separately to multiple schools and to maximize the number of students who get into at least one school, a few school districts now centralize the lottery process, often using mechanisms that draw upon 2012 Nobel prize-winner Alvin Roth's work on market design. Results from the most recent District of Columbia's common lottery provide an indicator of oversubscribed demand: of the 17,000 students that entered the unified lottery, 71 percent of students received an offer from at least one school on their list, but only 60 percent received an offer from one of their top three choices (as reported in Brown 2014).

Charter school authorizers, as designated by state law, choose which charters to grant, provide ongoing oversight of charter schools, and make renewal decisions at the end of the charter contract term (typically every five years). Charter schools are allowed to operate with a degree of autonomy from some of the rules and regulations governing traditional public schools, and so those who want to start a charter school typically must submit a lengthy application, including a mission or statement about what will differentiate their proposed school. Decisions about whether to renew are often based on relative test score measures or financial health (including enrollment). Schools do close—sometimes suddenly—compelling students to find another charter school option or revert back to their local traditional public school. For example, about 3 percent of all charter schools closed in 2014 (National Alliance for Public Charter Schools 2015b, p. 2). In Texas and North Carolina, respectively, Baude, Casey, Hanuskek, and Rivkin (2014) and Ladd, Clotfelter, and Holbein (2015) conclude that charters that close are disproportionately less effective, while those that remain open improve in value-added over time.

The required process of random assignment for charter schools with too many applicants can bring worry and letdown for lottery participants, but it also generates an opportunity for research. Over the past decade, a number of studies have been able to gather data from lottery results and match them to administrative records to allow for rigorous evaluation of the impact of charter school attendance on student outcomes. Most of these studies look at 3 to 30 schools at a time. The results show wide dispersion. Some charter schools are estimated to increase performance on

[1] For examples of oversubscribed demand at popular charter schools in Baltimore, see Wiltenburg (2015); for examples in New York City, Chapman and Brown (2014); for examples in Massachusetts, Pisano (2015); for examples from in Houston, Rahman (2015).

state-required tests (especially math scores) by more than half a standard deviation per year of attendance, while others are estimated to have substantial negative effects. The estimates are often imprecise, with large standard errors.

In this paper, we look at the results from the research on charter schools which has taken advantage of evidence from lotteries and also take a more in-depth look at school-level differences. We do not attempt to answer the controversial question of whether more (or fewer) charter schools would benefit students, *on average*, since lottery studies are limited by the fact that they examine only schools that are oversubscribed and do not examine impacts for students who do not apply (for a discussion of different sides of this debate, see the website "Charter Schools in Perspective"). Rather, our intent is to ask *which* charter schools benefit which kinds of students. In so doing, we hope to learn what sorts of activities happening at successful charters might be worthwhile expanding into other schools.

A general conclusion emerging from the previous literature, which we will discuss more in this paper, is that the distinguishing feature of the charter schools with the largest positive effects is their adoption of an intensive "No Excuses" approach with strict and clear disciplinary policies, mandated intensive tutoring, longer instruction times, frequent teacher feedback, and a relentless effort to help all students. These factors need not be exclusive to charter schools: for example, Fryer (2014, 2016) offers evidence that reinventing traditional public schools in urban settings to have these characteristics can lead to similarly large performance improvements.

In line with the earlier literature, we also find that schools that have adopted a No Excuses approach are correlated with large positive effects on academic performance. However, we find that No Excuses schools are concentrated in urban neighborhoods with very poor-performing schools and are scarce in nonurban areas. Thus one reason for the large effects achieved by No Excuses urban schools is that fallback public schools for urban students have such poor performance. Neal (2009) makes a similar point that private school returns are largest for urban minority students. Once the performance levels of fallback schools are taken into account, and we look at the individual components of a No Excuses approach using charter school level data, we find that intensive tutoring is the only characteristic that remains significant in improving student performance. Tutoring offered at charter schools is typically more intense than tutoring offered at traditional public schools. Charter schools often use paid tutors, add tutoring on top of already long school days, and require all students to participate. This finding about the importance of tutoring is in line with other recent evidence pointing to dramatic effects from intensive tutoring on its own, suggesting a good place to start for effective and practical reform at traditional public schools.

Lottery Studies of Charter Schools

When the first charter school legislation was enacted in 1991 in Minnesota, the law specified that oversubscribed schools would be filled by lottery (Junge 2014), although some states allow charters to give preference to certain students, such as siblings, children of employees, or educationally disadvantaged students (National

Alliance for Charter Public Schools 2015c). We know of 16 studies of charter schools that have used lotteries as a way to draw conclusions about their efficacy. Some of these studies also include results using a matching on observables approach, which we consider less-convincing; for the purpose of this paper, we focus on the lottery-based findings. First, we sketch how such lottery studies are conducted and then review the results.

The Methodology of Lottery Studies

In broad terms, the methodology of these studies is to compare those who won a charter school lottery with those who did not. Of course, complexities arise. One challenge is that researchers must take into account that not all winners attend charter schools and not all losers end up at traditional public schools. In Boston, for example, Abdulkadiroğlu, Angrist, Dynarski, Kane, and Pathak (2011) find one-fifth of lottery winners never attend a charter school and some lottery losers eventually end up in one (by moving off a waitlist, entering a future admissions lottery, or gaining sibling preference when a sibling wins the lottery). Therefore, in most studies of how charter schools affect test scores, researchers measure the effects in two stages, first estimating how winning a lottery predicts increased attendance at charter schools and, second, estimating how this predicted increased attendance affects achievement.[2] Because effects of attending a charter school are identified based on differences between initial lottery winners and losers, selection in who enrolls or persists in charter schools does not bias the causal estimates. While this approach addresses internal validity, external validity concerns may arise if the potential impact of charters is weaker for those who do not apply (but would have gotten in had they done so).

Fixed effects are usually added to the estimating equation for each group of students that applied to the same set of school lotteries to ensure that winner–loser comparisons are between those who had an equal chance of being selected (to the set of schools they applied). In many cases, test score data from different grade levels are stacked together, implicitly assuming that attendance effects increase equally for each year spent in a charter school versus not. Pooling data from multiple test results while clustering standard error estimates by grouping at the student level may also help increase precision.

An Overview of the Studies

We summarize lottery-based charter school research in Table 1. The studies described in Table 1 do not include all charter schools that have held lotteries. To do research on outcomes of winners and losers in a charter school lottery;

[2] In other words, winning a charter school lottery is used as an instrumental variable for charter school attendance. Conceptually, researchers estimate the "intention-to-treat" (ITT) effect of winning a lottery for a charter school seat on the outcome of interest (for example, student test scores) by calculating the difference in average outcomes between lottery winners and losers. The "local average treatment effect" (LATE) of charter school attendance on the outcome of interest is calculated by scaling up the ITT estimate by the difference in charter school attendance between lottery winners and lottery losers (this is sometimes called the treatment on the treated (TOT) when no or few lottery losers gain entry to charter schools).

Table 1

Summary of Lottery-Based Charter School Estimates of Reading and Math Test Score Impacts

Setting (1)	*Sample* (2)	*Paper* (3)	*Two-stage least squares impacts of per-year charter attendance (all effects significant at 5% level unless otherwise noted)* (4)
Massachusetts	Boston (8 schools)	Abdulkadiroğlu, Angrist, Dynarski, Kane, and Pathak (*QJE*, 2011)	MS: 0.198 sd ELA, 0.359 sd math HS: 0.265 sd ELA, 0.364 sd math
	Boston (13 schools)	Cohodes, Setren, Walters, Angrist, and Pathak (Boston Foundation, 2013)	MS: 0.138 sd ELA, 0256 sd math HS: 0.271 sd ELA, 0.354 sd math
	Massachusetts (26 schools)	**Angrist, Pathak, and Walters (*AEJ: Applied Economics*, 2013)**	MS: 0.075 sd ELA, 0.213 sd math HS: 0.206 sd ELA, 0.273 sd math
	KIPP Lynn	Angrist, Dynarski, Kane, Pathak, and Walters (*JPAM*, 2012)	MS: 0.133 sd ELA, 0.352 sd math
	UP Academy Charter School of Boston	**Abdulkadiroğlu, Angrist, Hull, and Pathak (NBER Working Paper, 2014)**	MS: 0.118 sd ELA, 0.270 sd math
National	15 states (36 schools)	**Gleason, Clark, Clark Tuttle, Dwoyer, and Silverberg (2010)**	MS: –0.04 sd reading, –0.04 sd math (not significant). †Year 2 impacts divided by 2 to get a per-year estimates
	KIPP schools (24 schools)	**Clark Tuttle, Gleason, Knechtel, Nichols-Barrer, Booker, Chojnacki, Coen, and Goble (Mathematica Policy Research, 2015)**	ES: 0.11 sd on letter-word identification and 0.10 sd on passage comprehension test in reading, 0.14 sd on calculation, 0.02 sd (not significant) on applied problems in math. From study-administered Woodcock-Johnson exam. †Year 3 impacts divided by 3 to get a per-year estimate MS: 0.08 sd reading, 0.12 sd math. †Year 2 impacts divided by 2 to get a per-year estimate
	KIPP middle schools (12 schools)	**Clark Tuttle, Gill, Gleason, Knechtel, Nichols-Barrer, Resch (Mathematica Policy Research 2013)**	0.08 reading (not significant), 0.18 math. †Year 2 impacts divided by 2 to get a per-year estimate
	Charter schools that were members of charter management organizations in 14 states (16 schools in 6 sites; estimates aggregated by site)	**Furgeson, Gill, Haimson, Killewald, McCullough, Nichols-Barrer, Teh, Verbitsky-Savitz, Bowen, Demeritt, Hill, and Lake (Mathematica Policy Research 2012)**	Intention-to-treat estimates: MS/HS: –0.02 reading (not significant), –0.05 math (not significant).
New York City	New York City (42 schools)	Hoxby, Murarka, Kang (2009)	ES/MS: 0.09 sd ELA, 0.12 sd in math HS: 0.18 sd ELA, 0.19 sd math
	New York City (29 schools)	**Dobbie and Fryer (*AEJ: Applied Economics*, 2013)**	ES: 0.058 sd ELA, 0.113 sd math MS: 0.048 ELA (not significant), 0.126 math
	Harlem Children's Zone Promise Academy middle school	Dobbie and Fryer (*JPE* 2015)	0.031 sd (not significant) reading, 0.075 sd math. From study-administered Woodcock-Johnson exam.
	Harlem Children's Zone Promise Academy middle and elementary schools	Dobbie and Fryer (*AEJ: Applied Economics*, 2011)	ES: 0.114 sd ELA (not significant), 0.191 sd math (not significant) MS: 0.047 sd ELA (not significant), 0.229 sd math

(continued)

Table 1 (continued)

Summary of Lottery-Based Charter School Estimates of Reading and Math Test Score Impacts

Setting (1)	*Sample* (2)	*Paper* (3)	*Two-stage least squares impacts of per-year charter attendance (all effects significant at 5% level unless otherwise noted)* (4)
Chicago	Chicago International Charter School schools (3 schools)	Hoxby and Rockoff (Unpublished paper, 2004)	No significant impacts on math or reading (dependent variable is percentile score on Iowa Test of Basic Skills)
Unknown	Anonymous No Excuses charter schools run by prominent CMO in mid-sized urban school district (4 schools)	Hastings, Nielson, Zimmerman (NBER Working Paper, 2012)	0.346 sd reading, –0.092 sd math (not significant), estimates are a mix of different years
Washington, DC	SEED School	**Curto and Fryer (*JLE*, 2014)**	0.211 sd reading, 0.229 sd math

Notes: This table only includes studies that use charter school lotteries to estimate effects on test scores. Some of these studies also include or focus on observational results, which are not reported here. In some cases where there are multiple studies of the same setting, we focus on published academic studies, adding studies when it appears that a substantial number of additional schools have been added. All impacts are second stage estimates reported in standard deviations and are statistically significant unless noted otherwise. Citations in boldface type indicate that this study contributes to the analyses presented in this paper. See Appendix Table 1 for more details on the studies indicated in boldface. ES = elementary school, MS = middle school, HS = high school, sd = standard deviation, ELA = English/language arts, CMO = charter management organization.

records must be in suitable condition; enough time must elapse to observe student outcomes of interest; researchers must obtain permission from schools to work with their lottery records; and, because of federal privacy law, the matching of lottery records to student test scores often requires either individual consent from study participants or collaboration with state or school district administrators who can conduct or supervise the match. In cases of multiple studies working with the same data or location, we focus here on the most recent published academic study or report, or if not that is not available, the most recent unpublished study. In some cases in the discussion that follows, we will rescale the estimates of charter school effects to be comparable across studies.[3]

Hoxby and Rockoff (2004) collected admissions lottery data from three No Excuses–style Chicago International Charter Schools (CICS), which deliberately

[3] More specifically, in cases where a study reported only the intention-to-treat effect (the outcome effect from winning a lottery) and no first stage estimate (the effect of winning a lottery on attendance), we noted this in Table 1. If the first stage and intention-to-treat are reported but a local average treatment effect is not, we divide by the best estimate of the first stage. In cases where a study reported only cumulative estimates, we divided the final year estimate by the number of years observed to obtain a per-year estimate. When we convert estimates to per year or second stage estimates, we also divide the standard errors by the same factors we divide the coefficients. In the cases where we are converting intention-to-treat estimates to second stage estimates, this *will not* correct the standard errors as a typical two-stage least squares procedure would in a statistical software program. Thus our standard errors are likely slightly too small for a subset of the charter school impact estimates that are based on intention-to-treat estimates—those from the Knowledge is Power Program (KIPP) (Clark Tuttle et al. 2013) and charter management organization (Furgeson et al. 2012) studies. We follow these conventions in our data analysis as well. Means and standard deviations are weighted by the inverse of the standard error of the relevant point estimates, both here and throughout our study.

locate in disadvantaged urban communities to target low-income families. Hoxby and Rockoff had admissions lottery data matched to Chicago Public School administrative data on test score outcomes. They find small positive changes due to charter school attendance, not statistically significant at standard levels.

Around the same time as Hoxby and Rockoff's study, another team of economists began collecting charter school lottery data from Massachusetts and, with support from state officials, obtained access to administrative public school data for matching. Abdulkadiroğlu et al. (2011) focus on students residing in Boston prior to applying to at least one of five charter middle schools or one of three charter high schools where high demand cause the schools to be oversubscribed. They find very large average effects: charter school attendance increases state-level English/language arts and math performance test scores by 0.2 and 0.35 standard deviations per year respectively.

Given that that the achievement gap between black and white students in Massachusetts is about 0.7 to 0.8 standard deviations, these estimates suggest that three years of charter school attendance for blacks would eliminate the black-white performance gap. Angrist, Pathak, and Walters (2013) update this analysis to include urban and nonurban schools across Massachusetts, along with additional years of test score data. They continue to find positive average charter school effects on test scores, but these effects appear in urban schools only and with wide variance across schools—a finding we revisit later in this paper.

The New York City Department of Education also facilitated the matching of charter school lottery data with standardized test scores in English/language arts and math. Dobbie and Fryer (2013) collected data from 19 elementary and 10 middle schools that were oversubscribed. They also find that charter school attendance increases test scores, especially for math scores, though again with large variance across schools. In an earlier lottery-based study of New York City charter schools, Hoxby, Muraka, Kang (2009) also found large and significant results for middle schools and report even larger positive effects for charter high schools.

Studies that use survey data for national samples of charter schools tend to find positive but not statistically significant overall impacts. Both Gleason et al. (2010) and Furgeson et al. (2012) contacted charter schools asking for permission to survey lottery applicants and obtain consent prior to randomization. The Furgeson et al. group also collected retrospective data to match directly with administrative data. Among the 77 charter middle schools that agreed to participate in Gleason et al. (2010), only 36 ended up with a large enough waiting list to use in their study. On average, lottery winners performed no better and no worse in math and reading scores than lottery losers two years after students applied, though as in Massachusetts, urban charters outperformed nonurban ones. Furgeson et al. (2012) identified 16 charter schools (of 109 schools run by charter management organizations) with adequate records and also find insignificant overall test score effects from winning the lottery. Estimates from survey data, however, are generally more imprecise than those using administrative data.

Seven additional lottery-based studies estimate charter impacts for specific schools or organizations. Three of these studies examine the Knowledge Is Power

Program (KIPP) charter schools. KIPP is the largest network of charter schools in the country and is often described as the source of the No Excuses movement (as reported in Rotherham 2011). In KIPP schools, principals and teachers have high behavioral and academic expectations for all students. Further, parents, students, and teachers sign a "learning pledge" and follow a strict disciplinary code. School hours are extended typically to between 7:30AM and 5:00PM and include occasional Saturdays and summer weeks, and tutoring is also offered during these times. In the 2014–2015 school year, KIPP's network included 162 schools serving 58,495 students in prekindergarten through grade 12 (Clark Tuttle et al. 2015, xiii). All three KIPP lottery studies listed in Table 1 find significant positive charter attendance effects on achievement (Angrist, Dynarski, Kane, Pathak, and Walters 2012; Clark Tuttle et al. 2013; Clark Tuttle et al. 2015). In addition to the test score results, Clark Tuttle et al. (2013) also find that KIPP attendance increases the amount of homework per night by about 45 minutes and increases school satisfaction but does not affect effort or engagement.

The Promise Academy charter schools in the Harlem Children's Zone (HCZ) contain many similar No Excuses elements. Dobbie and Fryer (2011) estimate that attendance at the Promise Academy raises test scores by about 0.20 standard deviations per year, although effects on English/language arts were not significant. The study also finds that attendance at the Promise Academy reduces absenteeism.

Two other charter schools aligned with the No Excuses model have been evaluated. The Unlocking Potential (UP) Network focuses on in-district school turnaround for chronically underperforming schools. In 2011, UP Academy Charter School of Boston replaced a failing traditional public school in Boston; within a year, the school was required to hold a lottery to address oversubscription (as reported in Nix 2015). Abdulkadiroğlu, Angrist, Hull, and Pathak (2016) find lottery-based UP attendance effects of 0.12 standard deviations per year for English/language arts scores and 0.27 standard deviations for math. SEED schools are No Excuses boarding schools in Baltimore and Washington, DC, for students from disadvantaged backgrounds in grades 6 through 12. At the Washington, DC, school, Curto and Fryer (2014) estimate increases in math scores of 0.23 standard deviations and reading scores of 0.21 standard deviations per year of attendance.

Many of the estimated effects in Table 1 are impressive. Attendance at some charter schools leads to large test score effects of more than half a standard deviation after two years of attendance. Most educational interventions such as class size reductions, teacher or student incentives, more resources, or extended time, generate gains that are less than one-quarter of this amount (Fryer 2016). However, while the large impacts from attending No Excuses schools like KIPP, UP Academy, and the Promise Academy are encouraging, some of the other charters generate no effect or even negative effects. Overall, the per-year average effect of attending a charter school in our sample of 113 schools is 0.080 standard deviations in math and 0.046 standard deviations in English/language arts. Our real interest from these papers, however, is not whether charter schools are effective on average, but rather what makes an effective charter school. Therefore, we dig a little deeper.

School-Specific Effects

The main focus of the studies of charter schools that use lottery-based evidence is usually to compare a group of charter schools to a group of alternatives. However, we want to look at how school-level characteristics of charter schools may influence the results—in particular, whether the estimated effects of charter schools are larger in poor-performing urban neighborhoods—and at the effects on certain subsets of students: blacks, Hispanics, students who were performing poorly in the past, and students who didn't apply but would have gotten in had they applied. We also look at some of the estimated effects of charter schools on nontest outcomes. We will refer to some individual studies from Table 1 that do this, and in addition, we combine school-based data from several of these studies (indicated in boldface type) to gain insight and statistical power.[4]

Larger Effects in Poor-Performing Urban Neighborhoods

We estimate charter school impacts relative to the experience of students who lose the lottery at that charter school. A charter school that attracts students who would have otherwise attended a particularly poor-performing traditional public school would appear more effective than an identical charter school that draws students who would have otherwise attended a better performing school (due to declines or less growth at the fallback school).[5]

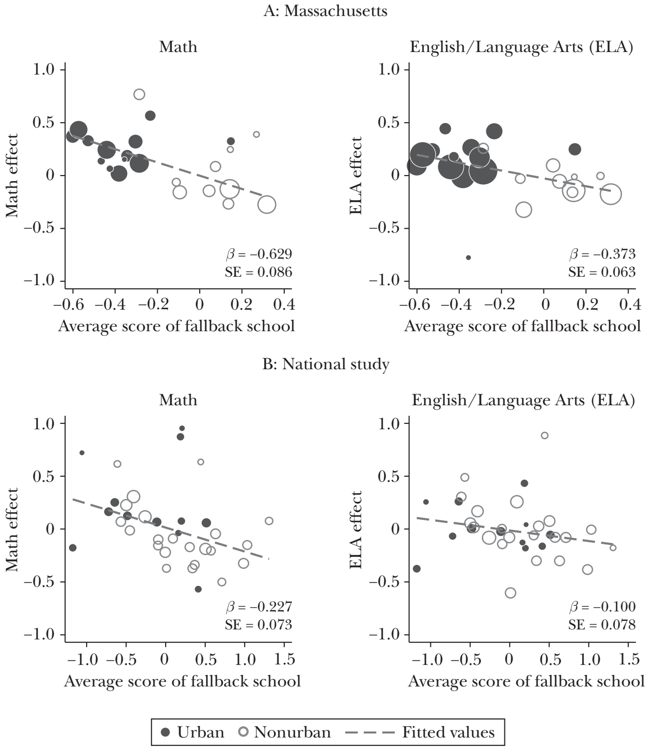
As mentioned earlier, Angrist et al. (2013) find stark differences in the positive effects that can be attributed to a charter school according to whether the school is located in an urban or nonurban setting.[6] The large positive gains from

[4] The online Appendix provides details of the data used in the rest of the analysis. Online Appendix Table 1 lists only the eight studies from Table 1 in boldface type that are included in our quantitative analyses; these studies cover 113 schools in total. Some studies in Table 1 were excluded from our school-based analysis because they were superseded by another paper: for example, the Boston schools are included in the Massachusetts study, and the 2009 study of New York City schools was replaced by a more recent 2013 study. Also, the results of Hoxby and Rockoff (2004) could not be converted to standard deviations and Hastings et al. (2012) could not be converted to per-year second stage effects. Online Appendix Figures 1A and 1B are histograms showing the wide range of estimated standardized effects on math or English/language arts tests from a year of attending these schools, which show an average mean effect that is positive but imprecisely estimated and with large standard errors: the average per-year math test score effect is 0.080 and its standard deviation is 0.23; the average English/language arts effect is 0.046 with a standard deviation of 0.21. Online Appendix Figures 2A and 2B plot the math and English/language arts effect sizes against their corresponding standard errors to show that large point estimates are often accompanied by large standard errors. Online Appendix Figure 3 shows that if we focus on charter schools where the standard errors are estimated with some precision—less than or equal to 0.1 standard deviations—the charter school effect on math and English/language arts scores show a positive correlation of 0.64. The high correlation implies schools good at improving one subject are often good at improving others, and that these estimates have good signal-to-noise ratios. Online Appendix Table 2 shows how school characteristic variables are defined for the studies included in our regression analyses.

[5] Hastings, Nielson, and Zimmerman (2012) examine whether winning a school choice lottery impacts students' academic achievement even before they enroll in their chosen schools by raising their intrinsic motivation. They find that charter and magnet school lottery winners in an anonymous urban school district had truancy rates that were 7 percent lower than lottery losers in the period after the lottery was held but before winners enrolled in their new schools.

[6] Angrist, Pathak, and Walters (2013) define urban schools as those located in areas where the district superintendent participates in the Massachusetts Urban Superintendents Network. This includes Boston, as well as smaller districts such as Cambridge, Holyoke, Lawrence, and Worcester. In Massachusetts, urban

Figure 1
School-Level Charter School Effects by Scores of Fallback Schools



Notes: This graph shows school-level lottery-based charter school effects, where the effects are per-year school-level second stage point estimates, plotted against the average scores of fallback schools attended by noncharter students that applied to the charter school. The size of the point is weighted by the inverse of the standard error (larger points are more precise estimates). The following studies are included in this figure: the national study (Gleason et al. 2010) and Massachusetts (Angrist et al. 2013). See online Appendix Table 1 for details on these studies and for notes on modifications of published point estimates which put estimates on the same scale. See online Appendix Table 2 for description of the calculation of the fallback school scores.

the Massachusetts studies are concentrated among urban charter schools, while nonurban charters are generally ineffective and some may even make students worse off than if they had lost the lottery. We show this pattern in the top two panels of Figure 1, which plots the Massachusetts estimates by average achievement levels at the fallback schools for lottery losers. The fallback school achievement level is measured as the average test score at the noncharter school that lottery losers attend the following year, weighted by the number of students that attend. Students at urban schools that lottery losers attend score well below average in test scores, while

charter schools are almost uniformly located in areas with high poverty rates and high minority enrollment. We follow the definitions of variables as defined in their original studies. See Online Appendix Table 2 for a full list of variable definitions across studies.

the students at fallback nonurban schools generally score above average. The solid circles indicate effects from attendance at urban charter schools, which are almost all uniformly positive. Larger circles indicate more precise estimates (that is, smaller standard errors). The average urban charter school math effect is 0.25 (s.e. = 0.044). The open circles that indicate nonurban effects are mostly close to zero or even negative. The average math impact at nonurban charters is –0.07 (s.e. = 0.092).[7]

The top left graph in Figure 1 shows that when regressing the charter school effect in math on averages scores at the fallback schools, we get a strong negative relationship (–0.629, s.e.= 0.086). The R^2 is more than half (0.513). An indicator for whether a school is in an urban area has no additional explanatory power.[8] The top right graph of Figure 1 shows a qualitatively similar pattern, although less extreme, for charter school English/language arts impacts by test scores at the fallback institutions. Clearly, the most impressive charter school effects are found where fallback schools have the least impressive academic performance.

The national charter school study by Gleason et al. (2010) also displays a noticeable negative relationship between charter school effects and conditions at fallback schools. In this case, we use their dummy variable indicating "Large City" to define urban versus nonurban areas. For the performance level of fallback schools, we use the standardized average proficiency rate of the traditional public schools attended by lottery applicants in the year and grade level after losing a charter lottery (which is not on the same scale as the Massachusetts variable). The bottom left graph of Figure 1 shows that the slope from regressing charter math impacts on performance levels at fallback schools is also negative in this data (–0.227, s.e = .073). Again, the slope remains essentially the same when adding the urban dummy (–0.191, s.e.= 0.088). The slope for English/language arts test score impacts regressed on fallback school performance is also negative, but less steep and not significant.

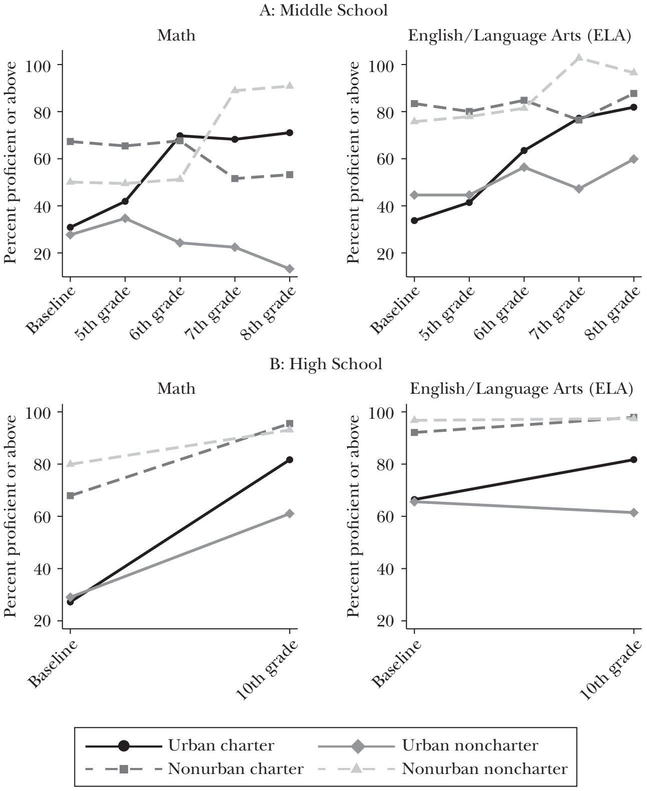
The importance of the fallback school to the size of the effect of enrolling in a charter school can also been seen in the top two graphs of Figure 2, which trace the accumulation of charter school effects over time for applicants to urban and nonurban middle schools in Massachusetts. We calculate percent proficient[9] on the state standardized exam for urban charter attendees who were offered a seat in the lottery (solid, dark line) and noncharter attendees who were not offered a seat in the lottery (solid, lighter line) at each grade level, with similar calculations for the nonurban charter applicants (dashed lines), using the methods from Abadie (2002, 2003) as described in Angrist, Cohodes, Dynarski, Pathak, and Walters (2016). In both subjects in middle school, applicants to urban and nonurban charters have

[7] Results are very similar, though less precise, when we use control complier test scores rather than the average school outcome of lottery losers to measure scores at fallback schools. Additionally, to address the concern that these findings reflect a mechanical correlation due to the presence of lottery losers' outcomes in the fallback school scores, we recalculate the fallback school scores using the prior year's scores (which the lottery losers do not contribute to). The findings are essentially identical, likely due to the relatively small proportion of lottery losers in any given school.

[8] Specifically, the slope and R^2 remain about the same when adding a dummy variable for the school being in an urban area (–0.658, s.e. = 0.375).

[9] We use percent proficient as opposed to mean scores, so we are making comparisons to a set standard rather than the state mean. However, mean scores show a very similar pattern.

Figure 2
Middle School Urban and Nonurban Charter School Effects over Time



Notes: This graph shows charter school effects for urban (solid lines) and nonurban charters (dashed lines) in Massachusetts over time. The darker line in each pair shows mean scores for charter school attendees who were offered a seat in the lottery (compliers) over time and the light line shows mean scores for noncharter attendees who were not offered a seat in the lottery (compliers). Scores for compliers were calculated using the methods from Abadie (2002, 2003). The gap between the lines is the second stage charter school effect at that grade level, using a dummy variable endogenous variable for charter school attendance. Percent proficient or above is the percentage of students who score at least 240 or higher on the scaled score of their state administered standardized test (MCAS).

very low proficiency rates at baseline. Then the proficiency rates diverge, with lottery winners that attend urban charter schools increasing their qualifications over time substantially, from about 30 to 70 percent. In math, lottery losers that attend urban noncharter schools actually have proficiency rates lower than their baseline rate by 8th grade. In nonurban schools (dashed lines), the opposite is true: noncharter schools (the light dashed lines) improve over time, and charter schools (the dark dashed line) do worse. The figure also shows that, by 8th grade, the proficiency of urban charter school attendees is in the range of children in the suburbs. The pattern for urban high schools shown in the bottom panels of Figure 2, also indicates a large proficiency gap of about 20 percentage points that opens up between charter and noncharter schools after two years for both math and English.

Larger Effects for Black, Hispanic, and Previously Poor-Performing Students

The urban charter school advantage is fairly consistent across subgroups. Table 2 reports per-year local average charter school treatment effects, but for different subgroups of students after combining data from the Massachusetts (Angrist, Pathak, and Walters 2013) and national (Gleason et al. 2010) charter school studies. The dependent variables for the four columns are math and English/language arts test scores separated for urban and nonurban charter schools.

For urban charter schools, the coefficients reveal positive and statistically significant effects across each of the subgroups we examine, with the exception of white students, for whom the charter school effect is positive and marginally significant in math and essentially zero in English/language arts. Effects are generally larger for less-advantaged students, including black and Hispanic students, those with low baseline scores, those who receive subsidized lunch, and English language learners. Special education and non-special-education students in urban charters have essentially the same test score impact estimates (for more details and updated impacts on English language learners and special education students, including effects on classification, see Setren 2015).

For nonurban charter schools, we find negative and statistically significant effects for female students, white students, and those without low baseline test scores, who do not receive subsidized lunch, who are not in special education, or who are not English language learners. There are marginally positive effects in math in nonurban schools for black students and those with low baseline scores.[10]

Similar Estimated Effects for Students Who Do Not Apply

Using lottery outcomes to estimate charter school effects provides a useful estimate of the advantage from charter schools for those who students who applied to oversubscribed charter schools. However, the lottery studies cannot clearly tell us adopting approaches practiced by the most successful oversubscribed charters would help the type of students who *don't* apply to charter schools. For example, charter schools often try to engage parents in their child's learning; if students who do not apply to charter schools have less involved parents, these types of parental engagement strategies may not work for these students.

In fact, there are a few studies suggesting that charters also benefit those who end up in them without applying. Abdulkadiroğlu et al. (2016) examine charter takeovers in New Orleans and Boston, where chronically poor-performing schools were replaced with charters, most of which follow the No Excuses pedagogy. By comparing students at schools not yet taken over with students at schools that were taken over and turned into charter schools and excluding attendance at other charters, the authors estimate charter school effects for students who passively enroll. They calculate estimates of charter school impacts at New Orleans takeover

[10] See the appendix to Chabrier, Cohodes, and Oreopoulous (2016), the NBER Working Paper version of our paper, for results for subgroups by each individual study, as well as other results by individual study.

Table 2

Per-Year Lottery Estimated Charter School Attendance Effects for Subgroups

		Urban		*Nonurban*	
		Math (1)	*English/ Language Arts* (2)	*Math* (3)	*English/ Language Arts* (4)
Male		0.228***	0.122***	–0.039	–0.046
		(0.046)	(0.043)	(0.049)	(0.042)
	N	8,310	8,180	4,020	4,050
Female		0.299***	0.117***	–0.126***	–0.097**
		(0.045)	(0.040)	(0.045)	(0.039)
	N	8,800	8,690	4,230	4,260
Black/Hispanic		0.337***	0.126***	0.107*	0.003
		(0.046)	(0.042)	(0.062)	(0.055)
	N	9,460	9,220	1,140	1,150
White		0.098*	–0.005	–0.128***	–0.097***
		(0.059)	(0.051)	(0.036)	(0.032)
	N	3,830	3,790	7,130	7,190
Low Baseline Score		0.289***	0.123**	0.003	0.022
		(0.051)	(0.050)	(0.052)	(0.051)
	N	4,370	4,380	2,030	2,090
Not Low Baseline Score		0.250***	0.100***	–0.180***	–0.130***
		(0.034)	(0.030)	(0.034)	(0.029)
	N	12,200	11,730	5,780	6,080
Subsidized Lunch		0.315***	0.156***	0.126*	0.075
		(0.039)	(0.035)	(0.066)	(0.062)
	N	11,650	11,500	1,320	1,340
Not Subsidized Lunch		0.171***	0.042	–0.130***	–0.107***
		(0.057)	(0.051)	(0.037)	(0.032)
	N	5,460	5,370	6,930	6,970
Special Education		0.246***	0.117	0.025	–0.117
		(0.073)	(0.074)	(0.095)	(0.093)
	N	3,120	3,090	1,310	1,330
Not Special Education		0.277***	0.123***	–0.108***	–0.074**
		(0.036)	(0.032)	(0.035)	(0.030)
	N	13,990	13,790	6,940	6,990
English Language Learner		0.382***	0.204**	0.166	–0.123
		(0.088)	(0.090)	(0.168)	(0.142)
	N	1,400	1,390	240	250
Not English Language Learner		0.253***	0.101***	–0.105***	–0.081***
		(0.035)	(0.032)	(0.033)	(0.029)
	N	15,710	15,480	8,000	8,070

Notes: This table shows per-year two-stage least squares estimates of charter school impacts for various subgroups, by urban and nonurban schools. Standard errors are clustered by student and school by grade and by year. The following studies are included in this figure: the national study (Gleason et al. 2010) and Massachusetts (Angrist et al. 2013). Individual study results are estimated with the microdata. Since data security restrictions preclude combining the microdata from these two studies, the combined estimates are the inverse variance weighted average. Sample sizes are rounded to the nearest 10.
***, **, and * indicate significance at the 1, 5, and 10 percent levels, respectively.

charters of 0.36 standard deviations in math and 0.15 standard deviations in English/language arts per year of takeover charter school attendance. These estimates are similar to or larger than lottery estimates for the sample of Massachusetts urban charters schools in Angrist et al. (2013). At UP Academy Boston, Abdulkadiroğlu et al. (2016) find that students who passively enroll in UP due to being grandfathered into the school have even larger English/language arts test scores impacts than students who attend due to winning an admissions lottery. Students who have been grandfathered have baseline English/language arts achievement 0.24 standard deviations below that of their lottery counterparts; attendance at UP effectively closes this gap.

Indeed, evidence from the lottery studies suggests that charter schools may actually be more effective at increasing the achievement of students who are less likely to apply. In Massachusetts prior to 2011, charter applicants were slightly less likely to participate in special education programs or to qualify for a subsidized lunch and had slightly higher test scores at baseline, compared to their traditional public school counterparts (Angrist et al. 2013). However, these subgroups tend to have a larger increase in test scores relative to the counterfactual. In their study of KIPP Lynn, Angrist et al. (2012) find that students with special needs or those who have limited English proficiency experience larger positive effects in reading (0.42 and 0.27 standard deviations for students with special needs and with limited English proficiency, respectively, compared to an average of 0.12 standard deviations) and math (0.47 and 0.42 standard deviations, respectively, compared to an average of 0.35 standard deviations) for each year of attendance. They also find that the effects of attendance at KIPP Lynn are larger for students with lower baseline scores. In Boston, Walters (2014) finds that high-achieving students from higher-income families are more likely to apply to charter schools, but charter schools generate larger positive effects for disadvantaged, low-achieving, and nonwhite applicants. These results are promising because they suggest these charter schools may be good at helping the most disadvantaged among the group of disadvantaged students.

Evidence is mixed as to whether charter schools for which lottery estimates are not available—either because the schools are not oversubscribed or because lottery records are not available—are more or less effective than the charter schools included in lottery-based studies. Angrist, Pathak, and Walters (2011) find that, for Massachusetts, urban charter middle and high schools, observational estimates, calculated using a combination of matching and regression, and lottery-based estimates are very similar. However, for nonurban charter middle schools, the observational and lottery-based estimates are not as close, with the observational estimates seeming to overstate the effect of charter schools. Using observational estimates, they find that for urban charter schools, positive effects are larger in the lottery sample, relative to the set of schools that are undersubscribed or have poorly documented lotteries. Following Abdulkadiroğlu et al. (2011), Dobbie and Fryer (2013) also find that the observational estimates for the lottery sample are somewhat higher than for the full sample of New York City charter schools, but the difference is quite small. In their study of KIPP middle schools, Clark Tuttle et al. (2013) find that matching-based estimates for the 10 schools in their lottery sample are similar to the matching-based estimates for all 41 study schools.

Effects on Non-Test-Score Outcomes

Most of the available research focuses on how charter school attendance affects scores on state-mandated tests, but some studies look at subsequent educational attainment and other outcomes likely linked to adult well-being (for example, Oreopoulos and Salvanes 2011). Angrist et al. (2016) find that charter attendance increases pass rates on the state high school graduation exam (which also qualifies students for state-sponsored college scholarships), as well as increasing SAT scores, advanced placement exam test taking, and advanced placement scores. While charter school attendance does not result in a statistically significant increase in overall college enrollment, it shifts enrollment from two-year to four-year colleges: charter school attendance decreases immediate enrollment in a two-year college by 11 percentage points and increases immediate enrollment in a four-year college by 17 percentage points.

Dobbie and Fryer (2015) collect longer-term survey and administrative data for the earliest cohorts of the Promise Academy middle school. Six years after the admissions lottery, the authors estimate a 0.075 standard deviation increase in math achievement among youth offered admission to Promise Academy, higher college enrollment immediately following high school graduation, higher rates of immediate enrollment in a four-year college, a 10.1 percentage point drop in female pregnancy, and a 4.4 percentage point drop in male incarceration. Together, these findings suggest that charter schools with large impacts on test scores can also change educational attainment and wellness outcomes. Charter schools without positive test score impacts may well influence other outcomes—however, there is no lottery-based evidence for longer-term outcomes for these types of charters, though Sass, Zimmer, Gill, and Booker (2016) find positive charter effects on earnings for charter schools in Florida that have few test score gains through a matching and instrumental variables strategy.

Why Are Some Charter Schools Effective But Not Others?

No Excuses Studies

Lottery studies that use admissions data from identifiable schools, like KIPP Lynn, UP Academy, SEED, and the Promise Academy charter schools, allow for a more in-depth analysis of the mechanisms behind the greater effectiveness of some types of charter schools. All four of these charters boost student performance substantially (especially in math) compared to the low-performing urban schools that lottery losers attend. Because each of these charter schools targets disadvantaged areas, they also have a competitive advantage against surrounding traditional public schools. Because these charters are all trying to turn around the prospects of youth from disadvantaged neighborhoods, it is perhaps not surprising that they have adopted similar No Excuses strategies, which have been cited for decades by qualitative researchers as important for improving student performance (Dobbie and Fryer 2013). As noted earlier, these strategies include uniforms, high expectations from principals and teachers, a tightly enforced discipline code, along with intensive tutoring, longer instruction time, regular feedback, college preparation services, and an energetic commitment to ensuring the academic success of all students. Another feature of these schools are empowered, flexible, and inspiring principals, whose presence may be necessary to implement No Excuses schools successfully (Carter 2000).

There is some question about the extent to which the No Excuses framework captures what is different about these schools. While these schools share many similarities, they also exhibit distinct differences in curricula and culture—for example, KIPP schools follow a particularly unique setup, with middle schools starting in Grade 5 instead of 6, students receiving "paychecks" for exhibiting good behavior that can be used for participation in school activities, and classrooms requiring students to SLANT (that is, Sit up straight, Listen, Ask questions, Nod, and Track the person speaking with your eyes). At HCZ's Promise Academy, students receive a free daily breakfast and regular instruction on character and social/emotional issues in gender-based groups, and all classrooms are equipped with smart boards. Suspension rates also differ. UP Academy and SEED report relatively high suspension rates (33.5 percent in 2013 for UP compared to a 2.8 percent state average, and 52 percent for SEED compared to a 23 percent city average), while KIPP Lynn and HCZ's Promise Academy report low suspension rates that are close to state averages (4.7 and 2.5 percent, respectively).[11]

Moreover, some evidence suggests that these four charter schools may spend more per student than the traditional public schools, because they receive additional funding from charitable foundations. KIPP, for example, reports that 15 percent of its annual operation expenses are covered by philanthropic contributions.[12] The extent to which these revenues are pursued due to less per-student funding from public sources remains a source of debate. KIPP schools, at least in general, appear to spend significantly more per student compared to traditional schools (Miron, Urschel, and Saxton 2011; Baker, Libby, and Wiley 2012), though this pattern is not observed in Boston charter schools (Angrist et al. 2016).

[11] For UP Academy: "2015 Massachusetts School Report Card Overview: UP Academy Charter School of Boston," Massachusetts Department of Elementary and Secondary Education, accessed January 21, 2016, http://profiles.doe.mass.edu/reportcard/SchoolReportCardOverview.aspx?linkid=105&orgcode=04800405&fycode=2015&orgtypecode=6&. For SEED: "SEED PCS of Washington, DC: 2014-2015 Equity Report," District of Columbia, accessed January 21, 2016, http://learndc.org/schoolprofiles/view?s=0174#equityreport. For KIPP Lynn: "2015 Massachusetts School Report Card Overview: KIPP Academy Lynn Charter School," accessed January 21, 2016, http://profiles.doe.mass.edu/reportcard/SchoolReportCardOverview.aspx?linkid=105&orgcode=04290010&fycode=2015&orgtypecode=6&. For Promise Academy: "Charter School Suspension Rates: Way Above District Averages," United Federation of Teachers, accessed January 21, 2016, http://www.uft.org/files/charter-school-suspension-rates-way-above-most-district-averages. Note that, according to the UFT report, suspension rates for KIPP schools in New York City vary widely, from 0 percent (KIPP NYC Washington Heights Academy Charter School) to 23 percent (KIPP AMP).

[12] For details, see KIPP, "Frequently Asked Questions," http://www.kipp.org/faq; Goldman Sachs, "Supporting the Harlem Children's Zone," http://www.goldmansachs.com/citizenship/goldman-sachs-gives/building-and-stabilizing-communities/hcz/; The Giving Common, "UP Education Network (Unlocking Potential Inc),", https://www.givingcommon.org/profile/1108725/up-education-network-unlocking-potential-inc/; and The SEED Foundation, "FAQs," http://www.seedfoundation.com/index.php/about-seed/faqs. All four websites accessed January 21, 2016.

Extensive research would be needed to document and appreciate the detailed differences across these schools (for an example, see Merseth, Cooper, Roberts, Tieken, Valant, and Wynne 2009). However, the similarity in effectiveness of these charter schools suggests that it is their common set of No Excuses characteristics that matter most in boosting performance. One exception might be the higher reading score effects for SEED Academy. Curto and Fryer (2014) suggest that this may be due to the fact that SEED is a boarding school.

What Relationships Exist between Charter School Characteristics and Effectiveness?

We combine data from three studies (Massachusetts, New York City, and the national study) for which school-specific charter effects and school characteristics are available in order to explore the relationship between school characteristics and effectiveness. We use both the school-specific effects and school characteristics variable definitions from Dobbie and Fryer's (2013) New York City study. Their school characteristics include five "nontraditional" inputs that are measured on a binary basis: teacher feedback, data-driven instruction, instructional time, high-dosage tutoring, and high expectations, as well as a standardized index of the five characteristics. They also include four traditional inputs: class size, per pupil expenditures, highly qualified teachers (as measured by masters degrees), and teacher certification and an index that combines these as well. We create equivalent variables for schools in the Massachusetts study (Angrist et al. 2013) and the national study (Gleason et al. 2010). For these two studies, our method for creating dummy variables equivalent to those in the New York City study is to estimate the median of a school characteristic—for example, per pupil expenditure—and assign values of one for schools that were above the median and zero for schools that were below. We are able to create fairly similar measures in the Massachusetts study, but had fewer similar input variables in the national study.[13] When we combine the three studies (Massachusetts, New York City, and the national study), our sample size is large enough to use lottery-based rather than observational estimates as our outcome of interest, whereas Dobbie and Fryer (2013) had to use observational estimates and Angrist, Pathak, and Walters (2013) use both observational and lottery estimates but have less precision than we do.

In Table 3, we present results from regressing the estimated charter school effects from the studies themselves on their corresponding school characteristics as defined above, which include both traditional and nontraditional inputs. All regressions include study fixed effects and a control for school level (elementary, middle, high) and are weighted by the inverse of the outcome's standard error. We also cluster standard errors by school to account for the fact that a handful of the charter schools in this sample have campuses serving multiple school levels. Columns 1 and 5 include results from single variable regressions, while all other

[13] In online Appendix Table 2, we describe in detail the variables and our adaptations across the underlying studies. See Chabrier, Cohodes, and Oreopoulos (2016) for the individual study results, which tend to be similar though less precisely estimated. Of the three studies whose data we combine, the most dissimilar study is the national study (Gleason et al. 2010), where the available survey variables do not map well to the constructs from the New York City study.

Table 3
Correlation between Lottery-Based Charter School Math Effects and Key Variables from Dobbie and Fryer (2013)

	Math				*English/Language Arts*			
	Single variable regression	*Multivariable regressions*			*Single variable regression*	*Multivariable regressions*		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Teacher Feedback	0.140**	0.104**			0.050	0.023		
	(0.062)	(0.047)			(0.047)	(0.048)		
N	86				86			
Differentiated Instruction (Data Driven)	0.093	0.055			0.106**	0.081*		
	(0.072)	(0.055)			(0.049)	(0.046)		
N	82				82			
Instructional Time	0.146***	0.071			0.078**	0.027		
	(0.051)	(0.049)			(0.038)	(0.041)		
N	86				86			
High Quality Tutoring	0.260***	0.153**			0.136***	0.073		
	(0.064)	(0.069)			(0.050)	(0.056)		
N	86				86			
High Expectations	0.145**	0.080*			0.100**	0.072*		
	(0.057)	(0.047)			(0.042)	(0.042)		
N	86				86			
Index of Practice Inputs	0.109***		0.142***	0.110***	0.064***		0.067***	0.064***
	(0.026)		(0.027)	(0.027)	(0.020)		(0.023)	(0.020)
N	87				87			
Class Size	0.015		0.063		–0.079*		–0.053	
	(0.066)		(0.045)		(0.047)		(0.037)	
N	85				85			
Per-Pupil Expenditures	0.089		–0.015		0.086**		0.030	
	(0.055)		(0.054)		(0.041)		(0.045)	
N	81				81			
Teachers with Masters	0.039		0.126***		0.049		0.088***	
	(0.062)		(0.040)		(0.043)		(0.034)	
N	84				84			
Teachers with Certification	–0.020		0.034		–0.034		–0.012	
	(0.061)		(0.044)		(0.043)		(0.037)	
N	85				85			
Index of Resource Inputs	0.021	0.028		0.023	0.000	0.007		0.002
	(0.041)	(0.026)		(0.028)	(0.025)	(0.019)		(0.019)
N	87	81	78	87	87	81	78	87

Notes: This table shows estimates from regressions of school characteristics on school-level charter school effect estimates using data from the National Study (Gleason et al. 2010), Massachusetts (Angrist et al. 2013), and New York City (Dobbie and Fryer 2013). Columns (1) and (5) show results from single variable regressions; each coefficient comes from its own regression. Columns (2)–(4) and (6)–(8) show results from multivariate regressions, with the school characteristics included as indicated. Regressions are weighted by the inverse of the school-level standard error. Regressions include dummies for school levels (elementary, middle) as well as study fixed effects, and standard errors are clustered by the school level to account for schools with campuses at multiple grade levels. See online Appendix Table 1 for details on these studies and for notes on modifications of published point estimates which put estimates on the same scale. See online Appendix Table 2 for variable definitions across studies.
***, **, and * indicates significance at the 1, 5, and 10 percent levels, respectively.

columns include multiple school characteristics. We also present results using an index of school practices, equal to a standardized sum of each school practice characteristic employed, as well as an index for school resource inputs summarized by a second standardized index.

When each characteristic is considered separately, in both math (in column 1) and English/language arts (in column 5), all of the school practice inputs but one are positive and statistically significant (excluding data-driven instruction for math and teacher feedback for English/language arts). The coefficient on the index summarizing the practice inputs, which correspond to No Excuses–style practices, is positive and precise. In math, none of the school resource variables have predictive power for charter school effects. In English/language arts, there appears to be a positive association between per pupil expenditures and school level impacts, and the coefficient on class size is significant but in the "wrong" direction. For both subjects, the summary index of resource inputs in columns 1 and 5 has no explanatory power. The other columns include multiple characteristics and generally show that school practices remain important even when controlling for resource inputs. These findings are consistent with the results from Angrist, Pathak, and Walters (2013) and Dobbie and Fryer (2013).

Taking Location into Account

We pointed out earlier that the charters with the highest value-added locate in areas where lottery losers end up in some of the worst performing schools; conversely, charter schools with the lowest value-added are in more suburban areas, where neighboring traditional public schools do relatively well. Also, charter schools that are more likely to locate in highly segregated and disadvantaged areas tend to be No Excuses schools, while nonurban charter schools, in contrast, tend to emphasize other priorities, such as performing arts, interdisciplinary group projects, field work, or customized instruction. In Massachusetts, for example, no charter schools in nonurban areas identify with a No Excuses philosophy, while two-thirds of charter schools in urban areas identify as No Excuses (Angrist et al. 2013).

Thus, we condition on test performance at fallback schools to explore whether the remaining variance in estimated charter school effectiveness still relates to No Excuses practices.[14] We drop data for New York City (Dobbie and Fryer 2013), for which we have no information about fallback school performance, leaving us with a sample of 57 schools from the Massachusetts and national studies. In column 1 of Table 4 we regress charter school effect estimates on a dummy variable for whether the charter is located in an urban area, while also including study fixed effects and school level dummies, again weighted by the inverse of the school effect standard error. Urban charters increase annual math scores by 0.28 standard deviations more than nonurban charters per year of attendance, on average. In bivariate relationships shown in columns 2–4, we see that test scores in the fallback school as well as school practice inputs also have explanatory power for charter school impacts.

Beginning in column 5, we combine the additional variables with the urban indicator. When we include average test performance at fallback schools as a

[14] Several others have also pointed out the importance of the fallback, or counterfactual, option in estimating program effects. See, for example, Heckman, Hohmann, Smith, and Khoo (2000) for evidence from job training, Kirkebøen, Leuven, and Mogstad (2014) for evidence from post-secondary decisions, and Kline and Walters (2015) for evidence on Head Start.

Table 4

Correlation between Lottery-Based Charter School Effects and Urban, Scores in Fallback Schools, and School Inputs

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Panel A: Math							
Urban	0.280***				0.170*	0.113	0.111
	(0.076)				(0.088)	(0.116)	(0.121)
Scores in the Fallback Schools		–0.327***			–0.238***	–0.197**	–0.197**
		(0.076)			(0.090)	(0.080)	(0.080)
Index of Practice Inputs			0.131***			0.064	0.065
			(0.032)			(0.045)	(0.047)
Index of Resource Inputs				0.015			0.008
				(0.047)			(0.030)
N	58	57	58	58	57	57	57
R^2	0.272	0.299	0.283	0.076	0.357	0.391	0.392
Panel B: English/Language Arts							
Urban	0.145***				0.090	0.048	0.052
	(0.054)				(0.060)	(0.070)	(0.072)
Scores in the Fallback Schools		–0.169**			–0.120	–0.083	–0.084
		(0.068)			(0.076)	(0.080)	(0.080)
Index of Practice Inputs			0.077***			0.048	0.047
			(0.024)			(0.033)	(0.034)
Index of Resource Inputs				–0.007			–0.010
				(0.028)			(0.021)
N	58	57	58	58	57	57	57
R^2	0.147	0.154	0.187	0.052	0.183	0.217	0.220

Notes: This table shows estimates from regressions of school characteristics on school-level charter school effect estimates. Columns (1) and (5) show results from single variable regressions; each coefficient comes from its own regression. Columns (2)–(4) and (6)–(8) show results from multivariate regressions, with the school characteristics included as indicated. Regressions are weighted by the inverse of the school-level standard error. Regressions include dummies for school levels (elementary, middle) as well as study fixed effects, and standard errors are clustered by the school level to account for schools with campuses at multiple grade levels. The following studies are included in this figure: The national study (Gleason et al. 2010) and Massachusetts (Angrist et al. 2013). See online Appendix Table 1 for details on these studies and for notes on modifications of published point estimates which put estimates on the same scale. See online Appendix Table 2 for variable definitions across studies.

*** Significant at the 1 percent level.

** Significant at the 5 percent level.

* Significant at the 10 percent level.

conditioning variable, along with an urban indicator and index variables for practice and resource inputs, the variables for fallback school performance remain strongly significant in math. This finding is consistent with Figure 1, which shows a strong relationship between charter effects and test scores at fallback schools, even within urban areas. Recall that No Excuses characteristics, as proxied by the index of practice inputs, are strongly related to charter school effects. But when including controls for urban areas and fallback school performance, the coefficient on the index of practice inputs falls by about half (to 0.065) and becomes insignificant.[15] A similar pattern holds for English/language arts scores: the coefficient on the school practice index falls by about half (from 0.077 to 0.047) and loses statistical significance. Of course, in both cases the loss of statistical significance could be due, in part, to an increase in the standard error. The estimated impact of more resource inputs remains negligible, with or without additional controls.

We rerun these results in Table 5, this time breaking up the school practice index variable into specific charter school characteristics and adding high suspension rates to the list of variables, again using the school characteristic definitions from Dobbie and Fryer (2013). When including urban and fallback school performance controls, the individual school characteristics that remain significantly correlated with charter school math effects are teacher feedback, intensive tutoring, and above average suspension rates (significant at the 10 percent level). These variables may serve as proxies for other underlying characteristics. Notably, the importance of the high expectations variable disappears once both urban status and fallback performance is taken into account. When all of the school characteristics variables are included together in column 7, the point estimates for the tutoring and high suspension rate variables remain about the same, while the others drop or remain negligible. Charter schools that offer intensive tutoring have math test scores 0.15 standard deviations higher, on average, for each year of charter attendance. This value is large and significant at the 10 percent level. After three years of attendance, students at these schools would have test scores almost half a standard deviation higher than lottery losers at fallback schools. Charter schools with high suspension rates have math test scores that are 0.12 standard deviations higher, on average, though this measure is not statistically significant. For English/language arts test outcomes, only differentiated instruction is significant when included with urban and fallback school performance controls, and none of the school characteristics variables are significant when they are included together in the same regression.

Overall, once one accounts for surrounding neighborhood and school characteristics, many of the specific charter school practices are no longer associated with student improvement. The main exception is intensive tutoring. Its estimated impact remains large and relatively stable, especially in math, even when conditioning on other charter school characteristics. However, after conditioning on fallback school quality, it is nonurban schools that provide most of the variation in charter school effects used to identify the importance of tutoring. When the model in column 7 of Table 5 is estimated for the 21 urban schools only (conditioning on fallback quality), the coefficients for all school characteristics are statistically insignificant with large standard errors. The coefficients on school characteristics when using only the 28 nonurban schools are also insignificant except the one for intensive tutoring (0.254, with a standard error of 0.112).

[15] In other specifications we tried—with an additional squared and cubic fallback school quality term, and without the urban dummy—the coefficient for the index of practice inputs also falls by about half. For the model without the urban dummy, the coefficient is significant.

Table 5

Correlation between Lottery-Based Charter School Effects and Urban, Scores in Fallback Schools, and Detailed School Inputs

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Panel A: Math							
Teacher Feedback	0.131**						0.066
	(0.059)						(0.064)
Differentiated Instruction (Data Driven)		0.066					0.046
		(0.070)					(0.066)
Instructional Time			0.072				–0.011
			(0.071)				(0.078)
High-Quality Tutoring				0.185***			0.153*
				(0.068)			(0.091)
High Expectations					–0.021		–0.013
					(0.079)		(0.076)
High Suspensions						0.144*	0.120
						(0.083)	(0.076)
Urban	0.184**	0.114	0.104	0.097	0.181*	0.114	0.091
	(0.088)	(0.084)	(0.089)	(0.080)	(0.105)	(0.082)	(0.112)
Scores in the Fallback Schools	–0.220***	–0.272***	–0.240***	–0.223***	–0.242***	–0.250***	–0.204***
	(0.083)	(0.086)	(0.092)	(0.080)	(0.088)	(0.086)	(0.074)
N	56	55	56	56	57	50	49
R^2	0.411	0.403	0.401	0.460	0.358	0.469	0.546
Panel B: English/Language Arts							
Teacher Feedback	0.017						–0.063
	(0.057)						(0.071)
Differentiated Instruction (Data Driven)		0.124**					0.071
		(0.054)					(0.064)
Instructional Time			0.040				–0.009
			(0.046)				(0.064)
High-Quality Tutoring				0.101			0.084
				(0.067)			(0.105)
High Expectations					0.073		0.109
					(0.071)		(0.076)
High Suspensions						0.095	0.111
						(0.062)	(0.077)
Urban	0.092	0.025	0.052	0.047	0.055	0.047	–0.046
	(0.063)	(0.053)	(0.053)	(0.056)	(0.069)	(0.056)	(0.062)
Scores in the Fallback Schools	–0.117	–0.148**	–0.124	–0.112	–0.099	–0.185***	–0.154*
	(0.079)	(0.068)	(0.076)	(0.078)	(0.083)	(0.070)	(0.087)
N	56	55	56	56	57	50	49
R^2	0.182	0.250	0.198	0.226	0.199	0.284	0.371

Notes: This table shows estimates from regressions of school characteristics on school-level charter school effect estimates. Regressions are weighted by the inverse of the school-level standard error. Regressions include dummies for school levels (elementary, middle) as well as study fixed effects, and standard errors are clustered by the school level to account for schools with campuses at multiple grade levels. The following studies are included in this figure: the national study (Gleason et al. 2010) and Massachusetts (Angrist et al. 2013). See online Appendix Table 1 for details on these studies and for notes on modifications of published point estimates that put estimates on the same scale. See online Appendix Table 2 for variable definitions across studies.
***, **, and * indicate significance at the 1, 5, and 10 percent levels, respectively

This evidence in support of tutoring is of course only suggestive, based on analysis of correlations rather than on the randomized provision of tutoring services. However, the potential importance of intensive tutoring is in line with recent quasi-experimental and experimental studies that find large increases in student performance from tutoring, delivered either as part of a package of school reforms or on its own. Kraft (2015) uses two quasi-experimental methods to estimate the impact of implementing individualized tutoring classes four days a week at MATCH Charter Public High School in Boston and finds large and statistically significant impacts on English/language arts achievement. In his review of randomized experiments in education, Fryer (2016) distinguishes between low- and high-dosage tutoring, defining the latter as being tutored in groups of six or fewer for more than three days per week, or being tutored at a rate that would equate to 50 hours or more over a 36-week period. Consistent with our findings, Fryer finds that high-dosage tutoring programs have, on average, statistically significant positive treatment effects on math and reading achievement. In contrast, the meta-coefficient on low-dosage tutoring is not statistically significant for either subject. Some examples of recent randomized experiments showing gains from intensive tutoring include Lee, Morrow-Howell, Jonson-Reid, and McCrary (2010), who study the Experience Corps® (EC) program for placing older volunteers in elementary schools to tutor reading; Fryer (2014), who studied the use of intensive tutors in fourth, sixth, and ninth grades in Houston public schools; Markovitz, Hernandez, Hedberg, and Silberglitt (2014), who evaluated the Minnesota Reading Corps, a literacy tutoring program for kindergarten through third grade students; Cook et al. (2015), who studied an intensive tutoring serving male ninth and tenth graders in 12 public high schools in Chicago; and May et al. (2014), who evaluated an early-intervention literacy tutoring program called Reading Recovery.

Conclusions

Charter schools were originally intended to serve as research laboratories for learning about best practices in education. They have since become more viewed as competitive alternatives to traditional public schools. But with many charters now receiving more applications than spots available, the requirement that oversubscribed charter schools admit students through lottery has unintentionally created the research setting that the charter school movement's originators were seeking.

Our purpose in this paper is not to enter the debate on whether charter schools should exist or expand: we have not discussed issues like how increased competition from charters affects traditional public schools over time, or the possible effects of charter schools on the racial/ethnic or socioeconomic mixture of students (for an example of discussion of this point in North Carolina, see Ladd et al. 2015). Instead, our purpose is to gather existing evidence from charter lotteries to learn more about the education production function.

We confirm a finding from previous studies that a sharp divide exists between the effectiveness of charter schools in urban and nonurban settings. However, there are two important differences between the urban and nonurban charters that have been studied. One is that almost all the charter school alternatives that have been the subject of lottery studies in disadvantaged urban areas use a No Excuses approach, while there are few No Excuses schools in nonurban settings. The other main difference is that students who attend charter schools in disadvantaged urban areas are usually being compared to students who end up in very poor performing schools, while students in charter schools in nonurban areas are being compared to students who attend better performing schools. This pattern arises because the charter schools aiming to attract students from the worst performing traditional public schools often find them residing in highly segregated and disadvantaged urban neighborhoods.

Many charter schools in disadvantaged urban schools have proven to be impressively effective, often raising average test score performance by more than half a standard deviation after just two years of attendance. For less-advantaged students, including black and Hispanic students, those with low baseline scores, and English language learners, impacts are similar or higher than impacts for the more-advantaged. Other studies find corresponding improvements for longer-term outcomes, such as reductions in incarceration rates and teen pregnancies and increases in enrollment in four-year colleges (Dobbie and Fryer 2015; Angrist et al. 2016). It is unclear, however, if other types of charter schools would deliver similarly impressive results in areas with very poor-performing traditional schools, since there are currently not enough other types of charter schools in these areas to tell. It is also unclear if No Excuses schools would deliver similar results in nonurban areas; again, there are currently not enough of them to tell. For now, the kinds of charters that have been created in nonurban areas—such as those emphasizing performing arts, exploratory learning, or instruction tailored to different learning styles—may offer other benefits but do not appear to be improving standardized test scores.

After accounting for the charter school effect variation explained by urban status and performance levels at fallback schools, we examined which charter school characteristics most strongly correlate with the little remaining variation. In line with previous studies, we find no evidence that differences in class size, per pupil expenditures, or teacher certification explain charter school effectiveness. The No Excuses explanatory factor that remains significant after controlling for fallback school performance (even for nonurban schools only) is whether a charter has an intensive tutoring program (though the effect of high suspension rates is close to significant). Of course, the tutoring variable could be a proxy for other school differences, and the relationships between effectiveness and several other associations are estimated imprecisely. But a push for intensive tutoring—more frequent and convenient than currently provided at traditional public schools, and in some cases mandatory—may serve as an important complement to instruction in many different kinds of classrooms.

■ *For sharing the data from and answering question about their studies, we thank Ira Nichols-Barrer, Brian Gill, Phil Gleason, Josh Furgeson, and Christina Clark Tuttle of Mathematica Policy Research; Danielle Eisenberg of the KIPP Foundation; Josh Angrist, Parag Pathak, and Chris Walters; and Will Dobbie and Roland Fryer. We also thank Carrie Conaway of the Massachusetts Department of Elementary and Secondary Education and the National Center for Education Statistics for additional data access, and Josh Angrist, Will Dobbie, and Mike Gilane, as well as seminar participants at Teachers College Columbia University and IFN-Stockholm for helpful comments.*

References

Abadie, Alberto. 2002. "Bootstrap Tests for Distributional Treatment Effects in Instrumental Variables Models." *Journal of the American Statistical Association* 97(457): 284–92.

Abadie, Alberto. 2003. "Semiparametric Instrumental Variable Estimation of Treatment Response Models." *Journal of Econometrics* 113(2): 231–63.

Abdulkadiroğlu, Atila, Joshua D. Angrist, Susan M. Dynarski, Thomas J. Kane, and Parag A. Pathak. 2011. "Accountability and Flexibility in Public Schools: Evidence from Boston's Charters and Pilots." *Quarterly Journal of Economics* 126(2): 669–748.

Abdulkadiroğlu, Atila, Joshua D. Angrist, Peter D. Hull, and Parag A. Pathak. 2016. "Charters Without Lotteries: Testing Takeovers in New Orleans and Boston." *American Economic Review* 106(7): 1878–1920.

Angrist, Joshua D., Sarah R. Cohodes, Susan M. Dynarski, Parag A. Pathak, and Christopher R. Walters. 2016. "Stand and Deliver: Effects of Boston's Charter High Schools on College Preparation, Entry, and Choice." *Journal of Labor Economics* 34(2).

Angrist, Joshua D., Susan M. Dynarski, Thomas J. Kane, Parag A. Pathak, and Christopher R. Walters. 2012. "Who Benefits from KIPP?" *Journal of Policy Analysis and Management* 31(4): 837–60.

Angrist, Joshua D., Parag A. Pathak, and Christopher R. Walters. 2011. "Explaining Charter School Effectiveness." NBER Working Paper 17332.

Angrist, Joshua D., Parag A. Pathak, and Christopher R. Walters. 2013. "Explaining Charter School Effectiveness." *American Economic Journal: Applied Economics* 5(4): 1–27.

Baker, Bruce D., Ken Libby, and Kathryn Wiley. 2012. "Spending by the Major Charter Management Organizations: Comparing Charter School and Local Public District Financial Resources." Boulder, CO: National Education Policy Center. http://nepc.colorado.edu/publication/spending-major-charter.

Baude, Patrick L., Marcus Casey, Eric A. Hanushek, and Steven G. Rivkin. 2014. "The Evolution of Charter School Quality." http://harris.uchicago.edu/sites/default/files/Rivkin.paper_.pdf.

Brown, Emma. 2014. "Joy and Anguish for Parents as D.C. Releases School Lottery Results." *Washington Post*, March 31. https://www.washingtonpost.com/local/education/joy-and-anguish-for-parents-as-dc-releases-school-lottery-results/2014/03/31/a04aea1e-b8ff-11e3-9a05-c739f29ccb08_story.html.

Carter, Samuel Casey. 2000. *No Excuses: Lessons from 21 High-Performing, High-Poverty Schools.* Washington, DC: Heritage Foundation. Available at http://eric.ed.gov/?id=ED440170.

Chabrier, Julia, Sarah Cohodes, and Philip Oreopoulos. 2016. "What Can We Learn from Charter School Lotteries?" NBER Working Paper 22390.

Chapman, Ben, and Stephen Rex Brown. 2014. "Success Academy Charter Schools Admissions Rate is Only 20%, Lower Than NYU." *New York Daily News*, April 4.

Charter Schools in Perspective. No date. A joint project of the Spencer Foundation and Public Agenda. Website: http://www.in-perspective.org/.

Chetty, Raj, John N. Friedman, Nathaniel Hilger, Emmanuel Saez, Diane Whitmore Schanzenbach, and Danny Yagan. 2011. "How Does Your Kindergarten Classroom Affect Your Earnings? Evidence from Project Star." *Quarterly Journal of Economics* 126(4): 1593–1660.

Citizens League of Minnesota. 1998. "Chartered Schools = Choices for Educators + Quality for All Students." http://citizensleague.org/wp-content/uploads/2013/05/424.Report.Chartered-Schools-Choices-for-Education-Quality-for-All-Students.pdf.

Clark Tuttle, Christina, Brian Gill, Philip Gleason, Virginia Knechtel, Ira Nichols-Barrer, and Alexandra Resch. 2013. *KIPP Middle Schools: Impacts on Achievement and Other Outcomes*. Washington, DC: Mathematica Policy Research. http://www.mathematica-mpr.com/our-publications-and-findings/publications/kipp-middle-schools-impacts-on-achievement-and-other-outcomes-full-report.

Clark Tuttle, Christina, Philip Gleason, and Melissa Clark. 2012. "Using Lotteries to Evaluate Schools of Choice: Evidence from a National Study of Charter Schools." *Economics of Education Review* 31(2): 237–53.

Clark Tuttle, Christina, Philip Gleason, Virginia Knechtel, Ira Nichols-Barrer, Kevin Booker, Gregory Chojnacki, Thomas Coen, and Lisbeth Goble. 2015. "Understanding the Effect of KIPP as It Scales: Vol. I: Impacts on Achievement and Other Outcomes." Washington, DC: Mathematica Policy Research. http://www.mathematica-mpr.com/our-publications-and-findings/publications/executive-summary-understanding-the-effect-of-kipp-as-it-scales-volume-i-impacts-on-achievement.

Cohodes, Sarah R., Elizabeth M. Setren, Christopher R. Walters, Joshua D. Angrist, and Parag A. Pathak. 2013. "Charter School Demand and Effectiveness: A Boston Update." The Boston Foundation and New Schools Venture Fund. Available at: http://seii.mit.edu/research/study/charter-school-demand-and-effectiveness-a-boston-update/.

Cook, Philip J., Kenneth Dodge, George Farkas, Roland G. Fryer, Jr., Jonathan Guryan, Jens Ludwig, Susan Mayer, Harold Pollack, and Laurence Steinberg. 2015. "Not Too Late: Improving Academic Outcomes for Disadvantaged Youth." Institute for Policy Research Northwestern University Working Paper WP-15-01.

Curto, Vilsa E., and Roland G. Fryer, Jr. 2014. "The Potential of Urban Boarding Schools for the Poor: Evidence from SEED." *Journal of Labor Economics* 32(1): 65–93.

Dobbie, Will, and Roland G. Fryer, Jr. 2011. "Are High-Quality Schools Enough to Increase Achievement Among the Poor? Evidence from the Harlem Children's Zone." *American Economic Journal: Applied Economics* 3(3): 158–87.

Dobbie, Will, and Roland G. Fryer, Jr. 2013. "Getting beneath the Veil of Effective Schools: Evidence from New York City." *American Economic Journal: Applied Economics* 5(4): 28–60.

Dobbie, Will, and Roland G. Fryer, Jr. 2015. "The Medium-Term Impacts of High-Achieving Charter Schools." *Journal of Political Economy* 123(5): 985–1037.

Fiske, Edward B., and Helen F. Ladd. 2000. *When Schools Compete: A Cautionary Tale*. Washington, D.C.: Brooking Institute Press.

Fryer, Roland G., Jr. 2014. "Injecting Charter School Best Practices into Traditional Public Schools: Evidence from Field Experiments." *Quarterly Journal of Economics* 129(3): 1355–1407.

Fryer, Roland G., Jr. 2016. "The Production of Human Capital in Developed Countries: Evidence from 196 Randomized Field Experiments." *Handbook of Field Experiments*. Elsevier. https://www.povertyactionlab.org/handbook-field-experiments.

Furgeson, Joshua, Brian Gill, Joshua Haimson, Alexandra Killewald, Moira McCullough, Ira Nichols-Barrer, Bing-Ru Teh, Natalya Verbitsky-Savitz, Melissa Bowen, Allison Demerrit, Paul Hill, and Robin Lake. 2012. *Charter School Management Organizations: Diverse Strategies and Diverse Student Impacts*. Princeton, NJ: Mathematica Policy Research. http://www.mathematica-mpr.com/our-publications-and-findings/publications/charterschool-management-organizations-diverse-strategies-and-diverse-student-impacts.

Gleason, Philip, Melissa Clark, Christina Clark Tuttle, Emily Dwoyer, and Marsha Silverberg. 2010. *The Evaluation of Charter School Impacts: Final Report*. NCEE 2010-4029. Washington, DC: U.S. Department of Education, National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences. http://ies.ed.gov/ncee/pubs/20104029/.

Hastings, Justine S., Christopher A. Nielson, and Seth D. Zimmerman. 2012. "The Effect of School Choice on Intrinsic Motivation and Academic Outcomes." NBER Working Paper 18324.

Heckman, James, Neil Hohmann, Jeffrey Smith, and Michael Khoo. 2010. "Substitution and Dropout Bias in Social Experiments: A Study of an Influential Social Experiment." *Quarterly Journal of Economics* 115(2): 651–94.

Hoxby, Caroline M., Sonali Murarka, and Jenny Kang. 2009. "How New York City's Charter Schools Affect Achievement." Cambridge, MA: The New York City Charter Schools Evaluation Project. http://users.nber.org/~schools/charterschoolseval/.

Hoxby, Caroline M., and Jonah E. Rockoff. 2004. "The Impact of Charter Schools on Student Achievement." https://www0.gsb.columbia.edu/faculty/jrockoff/hoxbyrockoffcharters.pdf .

Junge, Ember Reichgott. 2014. "Charter School Path Paved with Choice, Compromise, Common Sense." *Phi Delta Kappan* 95(5): 13–17.

Kirkebøen, Lars, and Edwin Leuven, and

Magne Mogstad. 2014. "Field of Study, Earnings, and Self-Selection." NBER Working Paper 20816.

Kline, Patrick, and Christopher Walters. 2015. "Evaluating Public Programs with Close Substitutes: The Case of Head Start." NBER Working Paper 21658.

Kraft, Matthew A. 2015. "How to Make Additional Time Matter: Integrating Individualized Tutorials into an Extended Day." *Education Finance and Policy* 10(1): 81–116.

Ladd, Helen F., Charles T. Clotfelter, and John B. Holbein. 2015. "The Growing Segmentation of the Charter School Sector in North Carolina." NBER Working Paper 21078.

Lee, Yung Soo, Nancy Morrow-Howell, Melissa Jonson-Reid, and Stacey McCrary. 2012. "The Effect of the Experience Corps Program on Student Reading Outcomes." *Education and Urban Society* 44(1): 97–118.

Markovitz, Carrie E., Marc W. Hernandez, Eric C. Hedberg, and Benjamin Silberglitt. 2014. *Impact Evaluation of the Minnesota Reading Corps K-3 Program*. Chicago, IL. http://www.nationalservice.gov/documents/research-and-reports/2014/impact-evaluation-minnesota-reading-corps-k-3-program.

May, Henry, Heather Goldsworthy, Michael Armijo, Abigail Gray, Philip Sirinides, Toscha J. Blalock, Helen Anderson-Clark, Andrew J. Schiera, Horatio Blackman, Jessica Gillespie, and Cecile Sam. 2014. "Evaluation of the i3 Scale-up of Reading Recovery: Year Two Report, 2012–2013." Philadelphia, PA: Consortium for Policy Research in Education. http://dx.doi.org/10.12698/cpre.2014.RR79.

Merseth, Katherine K., Kristy Cooper, John Roberts, Mara Casey Tieken, Jon Valant, and Chris Wynne. 2009. *Inside Urban Charter Schools: Promising Practices and Strategies in Five High-Performing Schools*. Cambridge, MA: Harvard Education Press.

Miron, Gary, Jessica L. Urschel, and Nicholas Saxton. 2011. "What Makes KIPP Work? A Study of Student Characteristics, Attrition, and School Finance." New York, NY: National Center for the Study of Privatization in Education.

National Alliance for Public Charter Schools. 2015a. "A Growing Movement: America's Largest Charter School Communities." http://www.publiccharters.org/publications/enrollment-share-10/ .

National Alliance for Public Charter Schools. 2015b. "Estimated Number of Public Charter Schools & Students, 2014–15." http://www.publiccharters.org/publications/open-close-2015/.

National Alliance for Public Charter Schools. 2015c. *State Laws on Weighted Lotteries and Enrollment Practices*. http://www.publiccharters.org/publications/weighted-lotteries-paper/.

Neil, Derek. 2009. "The Role of Private Schools in Education Markets." Chap. 26 in *Handbook of Research on School Choice*, edited by Bark Berends, Matthew G. Springer, Dale Ballou, and Herbert J. Walberg. Lawrence Erlbaum Associates/Taylor & Francis Group.

National Center for Education Statistics. 2015. "The Condition of Education 2015: Charter School Enrollment." U.S. Department of Education. http://nces.ed.gov/programs/coe/indicator_cgb.asp (accessed January 11, 2016).

Nix, Naomi. 2015. "A Boston Breakthrough: UP Academy Goes from Failing to First." *The 74*, August 10. https://www.the74million.org/article/a-boston-breakthrough-up-academy-goes-from-failing-to-first.

Oreopoulos, Philip, and Kjell G. Salvanes. 2011. "Priceless: The Nonpecuniary Benefits of Schooling." *Journal of Economic Perspectives* 25(1): 159–84.

Pisano, Chris. 2015 "Hundreds Turn Out for Holyoke Charter School Enrollment Lottery." *WGGB*, March 4. http://www.masscharterschools.org/media/news/hundreds-turn-out-holyoke-charter-school-enrollment-lottery.

Rahman, Fauzeya. 2015. "Word of Mouth Major Factor for Parents Charting Charter School Course." *Houston Chronicle*, November 7.

Rotherham, Andrew J. 2011. "KIPP Schools: A Reform Triumph, or Disappointment." *Time*, April 27.

Sass, Tim R., Ron W. Zimmer, Brian P. Gill, and T. Kevin Booker. 2016. "Charter High Schools' Effects on Long-Term Attainment and Earnings." *Journal of Policy Analysis and Management* 35(3): 683–706.

Setren, Elizabeth. 2015. "Special Education and English Language Learners in Boston Charter Schools: Impact and Classification." School Effectiveness and Inequality Institute (SEII) Discussion Paper 2015.05. http://seii.mit.edu/wp-content/uploads/2015/12/SEII-Discussion-Paper-2015.05-Setren1.pdf

Walters, Christopher R. 2014. "The Demand for Effective Charter Schools." NBER Working Paper 20640.

Wiltenburg, Mary. 2015. "Uncertain Future for Thousands in Charter School Lottery." *WYPR*, February 13. http://news.wypr.org/post/uncertain-future-thousands-charter-school-lottery#stream/0.

Journal of Economic Perspectives—Volume 30, Number 3—Summer 2016—Pages 85–108

The Measurement of Student Ability in Modern Assessment Systems

Brian Jacob and Jesse Rothstein

Economists often use test scores to measure a student's performance or an adult's human capital. In the research literature on the economics of education, student test scores are often used to estimate teacher effectiveness, or "value-added" (for example, Chetty et al. 2014a); to measure and attempt to explain the black–white achievement gap (for example, Fryer and Levitt 2004, 2006, 2013; Rothstein and Wozny 2013); or to measure the impacts of state- or district-level educational policy choices such as finance or accountability rules (for example, Dee and Jacob 2011; Lafortune, Rothstein, and Schanzenbach 2016). In the broader labor economics literature, test scores are often used as well as proxies for human capital, for example in examining the black–white wage gap conditional on cognitive ability as in Neal and Johnson (1996).

In our experience, many researchers think of an individual's score as a noisy but unbiased measure of true ability like, for example, the simple fraction of test items a student answers correctly. Unfortunately, the student achievement measures provided in modern assessment systems are rarely—if ever—so straightforward. Assessments commonly have multiple forms and are often adaptive, meaning that the questions students receive are based on their performance on previous

■ *Brian Jacob is Walter H. Annenberg Professor of Education Policy, Professor of Economics, and Professor of Education at the University of Michigan, Ann Arbor, Michigan. Jesse Rothstein is Professor of Public Policy and Economics and Director of the Institute for Research on Labor and Employment (IRLE) at the University of California, Berkeley, California. Both authors are Research Associates, National Bureau of Economic Research, Cambridge, Massachusetts. Their email addresses are bajacob@umich.edu and rothstein@berkeley.edu.*

† For supplementary materials such as appendices, datasets, and author disclosure statements, see the article page at
http://dx.doi.org/10.1257/jep.30.3.85 doi=10.1257/jep.30.3.85

questions. As a result, students are frequently presented with different questions that may not be of comparable difficulty. Moreover, modern test-making practice disparages simple summaries like the fraction correct in favor of estimates from complex statistical models that attempt to extract more information from the pattern of correct and incorrect responses. The scores that these models produce are generally not unbiased measures of student ability, and may not be suitable for many secondary analyses that economists would like to perform.

Consider the well-known National Assessment of Educational Progress (NAEP, also known as "the Nation's Report Card"). A little-known fact is that the scores computed for students who take the NAEP are Bayesian updates of prior information about the students' ability and depend not only on the examinees' responses to test items but also on their background characteristics, including race and gender. As a consequence, if a black and white student respond identically to questions on the NAEP assessment, the reported ability for the black student will be lower than for the white student—reflecting the lower average performance of black students on this assessment. Individual NAEP scores are not reported to students, parents, or schools, and this adjustment does not affect reported aggregate statistics such as the unconditional black–white test score gap; but, as we explain below, it can introduce important biases into many secondary analyses. Other testing systems do not incorporate students' background characteristics into their scores, but report posterior mean scores for students that are biased estimates of the students' ability, and therefore unsuitable for many of the secondary analyses that economists perform, which typically use the test scores as dependent variables (for example, to estimate the effects of programs or even just the black–white test score gap).

Even in the relatively rare case that the underlying student ability measure comes from a simple statistic such as the fraction correct, assessments often present transformed "scale" scores for each individual. Research using these test scores virtually always assumes that the ability measure has an interval property—that is, a one-unit change has the same meaning at every point on the scale (for example, an increase from 400 to 450 on the SAT represents the same improvement in student knowledge as an increase from 700 to 750). However, as explained below, this assumption is entirely unwarranted. This fact, widely recognized in the testing community but often ignored, undermines many of the purposes to which test scores have been put.

And, finally, the fact that test scores are inherently "noisy" measures of student ability has important implications for analyses that use the scores as explanatory variables, such as in wage regressions. As we discuss in more detail below, a recent paper demonstrates that the failure to properly account for measurement error in individual ability, when this is used as a control in a standard wage regression, would lead an analyst to overstate the black–white wage gap conditional on ability by nearly 50 percent (Junker, Schofield, and Taylor 2012).

Our goal in this paper is to familiarize applied economists with the construction and properties of common cognitive score measures and with their potential implications for economics research using these measures. Information about how scores are constructed is often buried deep in technical manuals, if presented at all. While the literature in psychometrics (the field concerned with the theory and methodology of psychological measurement) has explored many if not all of the issues that we discuss, economists and other applied researchers are generally unaware of them and frequently misuse test score measures, with potentially serious consequences for their analyses. These issues will become even more important in the coming years as new assessments, developed in conjunction with the new Common Core State Standards, are gradually rolled out in schools around the country.

We begin by discussing the domain covered by a test, and then the problem of assigning a quantitative scale to latent student ability.[1] We next turn to the statistical models used to convert examinees' responses to a series of test items into scores on the chosen scale. We then discuss the secondary analysis of test scores, when test scores are used as either dependent or explanatory variables, focusing in particular on how the test's measurement model can influence results. We attempt to provide both applied researchers and research consumers with practical guidance for evaluating the many research studies that use test-based cognitive ability measures.

What Does the Test Measure?

The first decision that must be made in designing a test concerns what is to be measured. Historically, psychometricians have distinguished between tests of aptitude and achievement. IQ tests (like the Wechsler Intelligence Scale for Children or Raven's Progressive Matrices) are designed to measure mental aptitude, conceptualized as a fixed trait that is unaffected by educational interventions. Respondents recite long strings of digits from memory or recognize patterns in abstract figures. By contrast, achievement tests aim to capture an individual's stock of accumulated knowledge and not his or her innate ability.

The distinction between aptitude and achievement is not always clear, however. IQ scores are affected by educational interventions such as preschool attendance (Heckman, Moon, Pinto, and Savelyev 2010) or by the amount of accumulated schooling (Cascio and Lewis 2006), though one might expect innate aptitude to be invariant to both. The Peabody Picture Vocabulary Test (PPVT), which is administered to children in the National Longitudinal Survey of Youth (NLSY), measures a child's "receptive vocabulary"—that is, number of words that the child recognizes and understands. The PPVT is sometimes described (and was designed) as an aptitude test. Yet a child's receptive vocabulary is surely affected as much by the quality of that child's educational experiences as by innate aptitude, particularly given evidence of substantial variation across socioeconomic groups in the number of

[1] Throughout this paper, we use "ability," "proficiency," "achievement," and "aptitude" interchangeably to refer to a latent trait that governs test performance. In many contexts these terms have distinct meanings—for example, some argue that IQ tests measure innate aptitude but not learned achievement—but such distinctions are not important for the purposes of this paper.

words to which young children are regularly exposed (Hart and Risley 1995). In our view, all cognitive scores should be seen as measures of what a test-taker can accomplish on the day of the test, which is influenced by a combination of the subject's innate ability, the educational and noneducational inputs received in the past, and other factors extraneous to the testing process like testing conditions, health, and mood.

Two related distinctions, central to psychometrics but largely ignored by economists, concern the domain covered and the malleability of the trait being measured. It is common to have separate tests for each core academic subject, including math and language arts. Within these broad subjects, many assessments have separate questions aimed at different subdomains, like grammar versus reading comprehension, or computation versus geometric reasoning. Scores are sometimes reported for each subdomain. Given a choice of domain, tests also differ in what is known as "instructional sensitivity." For example, a history test that focuses on facts that might have been covered in class is likely to be very sensitive to the quality and nature of the instruction that the student has received. By contrast, a test of historical reasoning, divorced from specific dates, names, and places, may better measure the student's accumulated skills across several academic subjects but be less sensitive to the specific curriculum or teaching methods of the most recent class. A related idea is that some tests may be more affected by the student's familiarity with the test form and scoring method—for example, students taking multiple choice tests must decide whether and how to guess at an item when the right answer is unknown. In many cases, it may be easier to improve scores by teaching test-taking strategies than by teaching the underlying material. Barlevy and Neal (2012) argue that avoiding this outcome should be a central consideration in the design of testing systems to be used for teacher accountability.

Scaling

Test scores are reported on different and arbitrary scales. The National Assessment of Educational Progress (NAEP), which bills itself as "The Nation's Report Card," reports scale scores, ranging from roughly 100 to 400 with standard deviations around 30, as well as discrete proficiency categories (basic, proficient, and advanced). The verbal and math sections of the SAT college entrance exam are scaled to have approximately normal distributions with means around 500, standard deviations around 100, minimum scores of 200, and maximum scores of 800. The SAT's competitor, the ACT, uses integers between 1 and 36 for each of four subjects, with means around 21 and standard deviations around 6. These scales are arbitrary in their location (mean), range, and distribution. That is, there is no reason the College Board could not assign the lowest performing student on the SAT a score of 100, or have the highest score be 1000, or set the standard deviation to be 50 or 150 instead of 100, or even adopt a scale that makes scaled scores approximately uniformly distributed.

Interval or Ordinal

Researchers using test scores generally treat them as an interval scale, meaning that a one unit change in a student's score at any point on the distribution reflects the same change in the underlying knowledge or skill. This assumption is implicit in any analysis based on score averages. However, there is generally no basis for interpreting test scales as having an interval property (Stevens 1946; Thorndike 1966; Bond and Lang 2013). Like utility and unlike income or temperature, measured achievement is best thought of as ordinal, not cardinal. This fact has important implications for virtually all empirical analyses of test scores.

Bond and Lang (2013) illustrate the importance of arbitrary scaling decisions in the calculation of a widely cited statistic in education research and policy: the black–white test score gap. Consider a test of three items, each testing a different skill, with the skills ranked cumulatively: A student must master skill 1 before mastering skill 2, and skill 2 before skill 3. Students can answer zero, one, two, or all three test items correctly. Suppose we have a sample of two black students who correctly answer 0 and 2 items, respectively, and two white students who answer 1 and 2 items correctly. The count of correct items is known as the "raw score." In this example, the average raw score for black students is thus 1, while that for white students is 1.5. Hence, the gap in mean raw scores is 0.5 points, or 0.6 standard deviations.

Now suppose that the three skills are the ability to recite the alphabet, to recognize letters, and to read fluently. In this case, one might consider the incremental knowledge represented by advancing from skill one (reciting the alphabet) to skill two (recognizing letters) to be smaller than the steps from zero (no measured pre-literacy) to one (reciting the alphabet) or from two (recognizing letters) to three (reading fluently). In this example, the difference between the two groups is driven by the black student who scored 0 and the white student who scored 1. If we assume the difference between these students' achievement is much larger than that between the two white students (who also differ in one skill), the black–white gap in average achievement approaches 1 full point. By contrast, if we assume the difference in knowledge between zero and one skill is arbitrarily small relative to that between one and two skills, the black–white test score gap approaches zero. More elaborate examples, where the distribution of one group does not stochastically dominate the other, could even produce reversals of the sign of the gap as the weight put on different skills varies.

This problem worsens if one considers changes over time. Assume that over the school year each student progresses one skill level, so the black students correctly answer 1 and 3 items correctly, and the white students answer 2 and 3 items correctly. The raw gap remains unchanged at 0.5 points. But if we assign more weight to the first skill (reciting the alphabet) than the second (recognizing letters), we would conclude that the black–white gap had shrunk; if we reverse these weights, we would conclude it had grown. Empirically, estimates of the black–white gap in achievement growth across grades turn out to be extremely sensitive to transformations of the test score, in a way that varies across test and grade level. Depending on the transformation and assessment used, Bond and Lang (2013) find that the change in the black–white test score gap between kindergarten and third grade can be as small as zero or as large as 0.6 standard deviations.

As another example, consider value-added estimates of how teachers affect the achievement of students. Setting aside questions about the causal interpretation of these estimates (Rothstein 2010, 2016; Chetty, Friedman, and Rockoff 2014a, b), any comparison of value-added across teachers whose students start with different baseline scores rests, implicitly, on an assumed interval scale. Without this, one cannot compare the impact of a teacher who works with very low-scoring students and raises their scores by 10 points to the impact of a peer who raises the scores of higher-scoring students 15 points.[2]

The ordinality of test scores thus poses a serious challenge for those working with test score data, and has inspired several types of responses. A first approach, favored by many psychometricians and education researchers, is to develop a parametric model that defines an interval scale for the achievement parameter, and then treat the resulting scores as interval. (Some scholars interpret "item response theory" models, discussed below, in this way.) However, just as with similar uses of parametric utility functions, it is not clear how one might evaluate the claim that a proposed scale of knowledge generated has an interval property.

A second approach, advocated by Bond and Lang (2013), is to accept the ordinality of test scores, limiting conclusions to those that are robust to arbitrary monotonic transformations of the scores. This approach drastically limits the statements that can be made. When scores are treated as ordinal, group achievement is only partially ordered; one group's achievement can only be said to exceed another's if the former's scale score distribution stochastically dominates the latter's. A related approach focuses on students' percentile scores. Reardon (2008) calculates the probability that a randomly chosen black student will have a test score higher than a randomly chosen white student. Ho (2009) and Ho and Haertel (2006) describe how this information can be converted to a standardized metric-free gap measure. These measures permit complete orderings and are invariant to test-makers' scaling decisions, but are nevertheless noninterval; they amount to rescaling the original test, but do not avoid concerns about assigning importance weights to achievement gains at different points in the distribution.

Some value-added models—known as the "Colorado Growth Model" or the "student growth percentile model"—also rely on percentile scores to sidestep some scaling issues. In these models, each student is assigned a "growth percentile" corresponding to the student's percentile in the distribution of test scores among the sample of students who had the same test score in the prior year. A teacher's value-added is computed as the median growth percentile of students in that teacher's class. Again, this measure is insensitive to the particular test score scale chosen,

[2] Indeed, the equation of teachers' causal effects on their students with their effectiveness relies on a much stronger assumption about the test score production process: one needs to assume that a teacher would have the same impact, in scale score points, regardless of the students' initial achievement. Even with an interval scale, this assumption of homogeneity of treatment effects may not hold.

but nevertheless provides a complete ordering. Barlevy and Neal (2012) propose building teacher accountability and compensation systems around measures closely related to student growth percentiles. Interpretation of this ordering as reflecting teacher effectiveness depends on interval-like assumptions, however, as there is no assurance that a given increment to a teacher's median growth percentile is equally easy to achieve at all points in the teacher or student distribution.

While a focus on the ordinal nature of test scores is clearly more defensible from a psychometric perspective, it does limit the questions that can be answered in research and policy evaluation. An approach that has received recent attention is to translate scores into units of another measure that we are willing to assume is interval, such as adult earnings or educational attainment (Cunha and Heckman 2006; Cunha, Heckman, and Schennach 2010; Bond and Lang 2015; Nielsen 2015b). For example, an attainment-scaled test score would be the average eventual educational attainment—looking ahead in time—of all students with a particular test score. Bond and Lang (2015) use this approach to measure the black–white gap at various grades. The attainment-scaled reading gap is roughly constant from kindergarten through grade seven at around 0.7 years of predicted educational attainment, while the math gap is close to a full year.

This forward-linking approach yields an interpretable scale that is plausibly interval, but it also raises questions. First, how should one choose the specific outcome to which the test scores are linked? There is no assurance that the scale defined by educational attainment will correspond to that defined by another outcome (such as earnings), nor that either corresponds to a hypothetical scale representing units of knowledge at the time of testing. For example, it might require more inputs to move a student from 9 to 10 years of education than from 11 to 12 years or 15 to 16 years. Second, scores on a forward-linked scale depend on both the inputs that the tested students received prior to the test and the inputs that earlier students received after the test. For example, the existence of an effective intervention program for low-scoring adolescents will raise the average educational attainment of children who scored poorly on kindergarten tests, and thus compress the left tail of forward-linked kindergarten scores relative to what would be seen from the same kindergarten test responses in a setting without such an adolescent intervention. This is contrary to the standard education production function approach in which a student's ability at time t is a function of all inputs the student has received up to, but not following, time t. Thus, while forward-linked scores may seem intuitive, they can sometimes produce odd results. For example, the black–white gap in attainment-scaled achievement at every grade is likely to be larger than the actual black–white gap in educational attainment, as black students tend to wind up with higher attainment than do white students with the same test scores. Overall, we regard this forward-linking approach as promising but underdeveloped, and not yet ready for broad application.

Finally, it might be possible to assume that raw test scores are partially but not fully interval. For example, we might be willing to assume that the difference between SAT scores of 1500 and 1000 is larger than that between 1000 and 990, even if we aren't willing to assume that it is 50 times as large. The challenge then is to parameterize and define this notion in some defensible way. Nielsen (2015a) provides a first step in this direction. His empirical results, like those of Bond and Lang (2013), suggest that cross-sectional achievement gap estimates (for example, for black/white and high-/low-income gaps) are robust to scale misspecification, but that changes in achievement gaps over time are considerably more sensitive to the choice of scale.

Standardized Scores

When analyzing tests that use well-known scales, such as the SAT, researchers often use unadjusted scale scores. When the scale is not familiar, economists frequently convert (or "standardize") scores to a known scale. There are three common methods: z-scores are the difference between the examinee's scale score and the mean scale score, divided by the scale score standard deviation; percentile scores are the examinee's rank in the distribution; and normal curve equivalents (NCEs) are obtained by applying the standard normal inverse distribution function to the percentile score. These ad hoc transformations aim to be comparable across tests and samples, but they yield scales that are no more or less correct than the raw or scale scores.

Even when researchers are willing to set aside concerns about non-interval scales, there are several practical challenges to using these transformations. The challenges derive from the fact that each transformation is defined relative to some norming population, which in practice can be small and nonrepresentative. Comparability across assessments depends on the use of norming populations with identical interval-scaled ability distributions, which is difficult to assess unless the two populations are given the same test. Consider, for example, a comparison between two states that administered different exams. Z-scores constructed from samples from the two states are comparable only if the mean and standard deviation of latent achievement, if measured on the same scale, would be identical in the two states; comparison of percentile or normal curve equivalent scores requires even stronger assumptions about latent achievement distributions. The same problem arises when comparing across ages or cohorts.

Cascio and Staiger (2012) reconsider a common empirical result that interventions aimed at younger children tend to have larger effects on standardized test scores (z-scores) than do those aimed at older children. They ask whether this could be attributable to the standardization process, rather than an indication that achievement becomes less malleable as children age. Scores are typically standardized separately by age. Differences in the effects of interventions carried out upon students of different ages might therefore reflect either differences in the interventions' true effects or differences in the distribution of scores across students. If the standard deviation of achievement increases with age, a plausible hypothesis as older students have been exposed to more out-of-school influences whose effects may accumulate, this could explain the observed pattern of declining coefficients with age. Cascio and Staiger adopt a parametric, additive model of student test scores as depending on permanent child ability, long-term knowledge that decays at a constant, geometric rate, and a transitory component that combines what they refer to as "short-term knowledge" with pure measurement error on the test. Based on this model, they conclude that while the variance of latent achievement does increase with age, this cannot fully explain the age pattern of estimated treatment effects.

Even when interval-scaled ability is similarly distributed across groups, measured ability may not be. The age-specific standard deviation combines true variability of ability among children with measurement error in the test. The measurement error component may vary with age even if true ability does not. Dividing by age-specific standard deviations of measured scores will tend to make between-group differences (like, the black–white gap) in z-scores larger at ages where test measurement error is smaller.

Practical Guidance on Scaling

Both those who consume and those who carry out research routinely use test scores in a way that assumes they have interval properties, although this assumption has no compelling justification. Should one only use the ordinal information contained in test scores, and forgo making any statements about the magnitude of effects? For example, a percentile–percentile plot comparing treatment and control groups would allow the researcher to fully characterize how the two distributions compare without relying on a particular scale, though in many cases the groups will not be ordered without scaling restrictions. While we understand this inclination, we are inclined toward the approach outlined by Nielsen (2015a), who seeks to narrow the class of scale transformations that are considered reasonable. At a minimum, we recommend that researchers make greater effort to test the robustness of their results to changes in the test score scale. For example, researchers might test their sensitivity to modest scale transformations such as the log or exponential of the reported scale score.

The common practice of standardizing reported scores also raises concerns. Secondary researchers should standardize based on the broadest possible population, even if their study focuses on a subpopulation. Comparisons of standardized effect sizes across studies should account for differences in the norming populations. Moreover, in most cases the true (net of measurement error) standard deviation should be used for standardization; in cases where this cannot be computed from measured scores, sometimes it can be backed out from information in the assessment's technical documentation, such as estimates of the test–retest reliability.

Measurement

Scaling involves the conversion of some initial ability measure into scores with a desired distribution. In this section, we discuss how test-makers obtain those initial ability estimates. The simplest estimate of ability is the fraction of items an individual answers correctly, often referred to as the "raw" score. But this approach has several limitations. First, a student's performance on a test, typically with relatively few items, measures student ability with error. Second, raw scores obtained from different tests or even from different test forms are not comparable. This is a particular issue for "adaptive" testing, where the student's performance on early items determines the difficulty of the items presented later. Third, even within the same form, items of moderate difficulty provide more information about a student's proficiency than do items that are very easy or very hard for that student; holding constant the difficulty of questions, some items may be better or worse at discriminating between more- and less-able individuals. For example, a test item about baseball statistics may measure knowledge of the sport better than it does statistical proficiency. These considerations motivate use of more complex performance measures, typically based on what is known as "item response theory."

Item Response Theory

An Item Response Theory (IRT) model specifies the probability that a student will answer each test item correctly as a function of a latent parameter representing the student's ability and of parameters relating to the item (van der Linden and Hambleton 1997). In one of the simplest specifications, the probability of a correct answer is a logit function of the difference between the student's ability and the item's difficulty. This is known as the "1 parameter logistic" (1PL) model. The implicit assumption here is straightforward: higher-ability students are more likely to answer each item correctly than are lower-ability students; all students are more likely to correctly answer simple than difficult items; and both relationships follow a simple, parametric form. (Some psychometricians argue that the implicit scale assigned to student ability by this model should be treated as interval, and refer to it as the "Rasch Model.")

Most item response theory models are more complex. The most common is the "three parameter logistic" (3PL) model, which adds two item parameters to the 1PL. One parameter represents a test item's "discrimination" between high- and low-ability students. The more discriminating an item, the steeper the relationship between the student's ability and the probability of a correct answer, and the less overlap there is in ability between those who answer it correctly and those who do not. The second is "guessability"—the probability that even a very low ability student will guess the correct answer. There are also item response theory models for essay questions or multiple-correct-answer questions that are scored in ways other than simply right or wrong.[3]

[3] For a more complete discussion of item response theory models, see van der Linden and Hambleton (1997). Embretson and Reise (2000) provide a readable introduction to the field for nonpsychometricians.

Measuring Student Ability in Test Scoring

After a particular item response theory specification—say, the three-parameter logistic version—is chosen, the next steps are to estimate the test item parameters (for example, difficulty, discrimination, and guessability), and then to use a student's particular combination of right and wrong answers, along with the item parameters, to generate a measure of the student's latent ability. The item parameters are well-identified as the number of tested students gets large, and can typically be estimated with relatively little error. But the typical test has relatively few items, so that the ability of an individual student is not precisely identified.[4] Modern assessment systems vary in the way they handle this.

Some testing systems, including most state tests used for accountability purposes, treat student ability as a fixed effect to be estimated directly via maximum likelihood methods or some variant thereof, applied to the sequence of right and wrong answers. The resulting estimate is (approximately) unbiased in most cases, but can be very noisy. Moreover, when a student gets all questions incorrect or all correct, a maximum likelihood estimate does not exist. A former state-mandated test in Michigan (the Michigan Educational Assessment Program) assigned students who answered all items correctly a score 10 percent higher than what was otherwise possible, and conversely assigned students who answered all items incorrectly a score 10 percent lower than what was otherwise possible. Other tests simply set minimum and maximum scores, and assign students with perfect scores to the endpoints.

Other testing systems estimate student ability using random effects models, which generate posterior distributions for each student's ability (including for those who answer all items correctly or all incorrectly). To assign a single score to a student, some tests report the mean or mode of this posterior distribution. Posterior mean scores can be seen as Empirical Bayes estimates of students' latent ability (Morris 1983), which "shrink" the individual's own score (roughly, the maximum likelihood estimate) toward the population mean in proportion to the noisiness of the maximum likelihood score. In item response theory models, ability is estimated most precisely for individuals near the middle of the measured ability distribution. This is because test items are most "discriminating," in the sense that a right or wrong answer provides the most information about the student's ability, when the probability of a correct answer is close to 50 percent.[5] For this reason, the reported ability measure in this framework will be shrunk more towards the mean for students who score extremely high or low on the exam.

Importantly, neither posterior means nor posterior modes are unbiased *estimates* of student ability. Recall, an unbiased estimate is one in which the estimation error

[4] The problem is similar to that which arises in many panel data models in econometrics, with the individual effect being the object of interest rather than a nuisance parameter.

[5] Interestingly, the standard errors of raw scores are largest at this point, and smaller in the tails: The variance of the fraction correct, p, is $p(1 - p)$, and this is highest when p is close to 0.5. Intuitively, logistic item response theory models stretch out the tails of the ability scale relative to raw scores, even as test performance provides relatively little information to discriminate amongst students who do very well or very poorly on the test.

(that is, the difference between the estimate and the individual's true ability) is zero in expectation and is uncorrelated with the true ability. As noted above, a student's posterior mean is "shrunk" toward the population mean. So, we would expect that the individual's posterior mean is on average smaller (in absolute value) than his or her true ability—it is a biased estimate.

On the other hand, posterior means are unbiased *predictors* of true latent ability. In other words, the prediction error (that is, the difference between a student's true latent ability and that student's posterior mean score) is mean zero in expectation, and is uncorrelated with the posterior mean score. This difference stems from the fact that when predicting how an individual will do in another context, it is optimal to adjust your prediction to account for the measurement error inherent in the student ability measure generated from a prior assessment. This insight has important consequences for secondary analysis of the scores, which we discuss below.

Most of the longitudinal databases created and distributed by the National Center for Education Statistics, including the Early Childhood Longitudinal Study, the National Educational Longitudinal Study of 1988 (NELS:88), the Educational Longitudinal Study (ELS), and the High School Longitudinal Study (HSLS), report scores constructed from posterior means. The Armed Services Vocational Aptitude Battery (ASVAB) scores reported in the 1997 wave of the National Longitudinal Study of Youth (NLSY97) are posterior modes.

Several major assessments, including the National Assessment of Educational Progress, attempt to provide more information about the posterior distribution than a single mean or mode by reporting several "plausible values," which are random draws from the examinee's posterior distribution. Plausible values are closely related to multiple imputation for missing data, and derive from Rubin's (1987; 1996) work on the topic. For excellent summaries of plausible values, including guidance on how to use them properly in secondary analyses, useful starting points are von Davier, Gonzalez, Mislevy (2009) and Carstens and Hastedt (2010).

Plausible values are neither unbiased estimators (like maximum likelihood estimates) nor unbiased predictors (like posterior means) of individual ability. Their primary benefit is that the variance of plausible values across students equals (in a large sample) the variance of latent ability, which allows one to calculate population variances. In contrast, the variance of an unbiased estimator (like maximum likelihood) will overstate the population variance while the variance of an unbiased predictor (like posterior means) will understate it. On the other hand, as we discuss below, while maximum likelihood and posterior mean ability estimates can each support some secondary analyses without further adjustment, there is essentially no multivariate secondary analysis that would be of interest to economists for which plausible values will yield unbiased estimates.

Incorporating Conditioning Variables into the Generation of Latent Student Ability Measures

To minimize examinee burden, tests are often kept short. Tests with relatively few questions will not provide precise (posterior) estimates of individual ability. To increase precision, some assessments—including the premier US and international assessment systems, the NAEP, and the Program on International Student Assessment (PISA), respectively—use priors that vary with student background characteristics. In this approach, a "conditioning model" relates performance on the exam to students' background characteristics (for example, race, gender, and family income), and then the prior that is used for computation of each student's posterior ability distribution is centered at the predicted values from this conditioning model.

As with random effects approaches described above, the posterior distribution from a conditioning model can be summarized by its mean or by several plausible values. The posterior mean still can be viewed as an Empirical Bayes or shrinkage estimator, but instead of being shrunk toward the unconditional mean, a student's performance is shrunk toward the predicted performance of students with similar background characteristics.

While the conditioning approach permits more precise estimates of students' ability (that is, the posterior distributions are tighter), it means that a student's measured score depends on personal background characteristics, even conditional on that student's test responses. Suppose, for example, that race is one of the background variables (as it is in National Assessment of Educational Progress tests), and that on average black students perform less well on the assessment than white students. Now consider two students, one black and one white but otherwise identical in their background characteristics and in their test item responses. Our two students' performance, initially identical, is "shrunken" toward different group averages. As a result, the white student's posterior distribution will stochastically dominate that of the black student, leading to gaps in their posterior means and plausible values. This does not bias the average black–white test score gap. The average score of all black students remains the same because the scores of high-performing black students are pushed down just as the scores of low-performing black students are pushed up, and the same for white students (with each pushed toward a group-specific mean). However, individual scores are affected. Scores generated in this way are at odds with the expectations of many data users, and as we discuss below can create important biases in more complex secondary analyses.

Recent administrations of the National Assessment of Educational Progress use hundreds of student and school characteristics in the conditioning model, including student demographics (like race, gender, and age), family background characteristics (like parental employment and parental education), school characteristics (including racial composition of the school and whether a school location is in an urban location), student self-reports of study habits and school performance (including overall grades, expected educational attainment, and time spent on homework), and teacher reports of aspects of the curriculum and of school policies. The model contains few variables that are likely to be of interest for policy evaluations, however. For example, it does not include measures of whether the school offers performance pay to its teachers, the type of school accountability system in place in the state, or the form of the state school finance formula. Moreover, none of these policy variables are likely to be well-proxied by the student-level characteristics that are included. As we discuss below, this may mean that program evaluations using NAEP scores as outcomes will understate programs' true effects.

Secondary Analysis with Latent Ability Measures

In this section, we discuss how the scaling and measurement issues described above can influence secondary analyses using test scores. For simplicity, we focus on ordinary least squares regressions and refer to the regression of interest as the "research model," distinguishing this from the "measurement model" (that is, the item response theory specification) and the "conditioning model" sometimes used to construct the test scores. For example, if one is interested in estimating the poverty achievement gap, the research model might be a regression of student ability on a binary measure of being above or below the poverty line. To focus specifically on issues arising from scaling and measurement, we ignore both sampling variability (essentially assuming that the number of examinees is large) and omitted variable bias.

In many cases, simple estimation of the research model using the test performance measure provided by a test-maker will lead to biased estimates of the relationships of interest. It is important for secondary researchers and consumers of this research to be aware of these biases. Their existence and magnitude depend on the type of ability measure used—that is, whether it is a fixed effects approach to student ability based on a direct maximum likelihood estimate, a posterior mean, or a plausible value, and in the latter cases whether the prior distribution is unconditional or conditional on background characteristics—and also on whether the ability measure is a dependent or independent variable in the research model. An important question is whether there are options available to the secondary researcher that permit unbiased estimation. Fortunately, there are options in many cases; unfortunately, all require access to additional information beyond the reported test score itself—often, but not always, item-level test data that can be hard to acquire.

While the sign of the bias arising in various scenarios is clear, the magnitude of the bias is not. We present illustrative evidence from two studies that assess the magnitude of biases that arise, one regarding racial and ethnic test score gaps and the other examining a "Mincerian" wage regression—that is, a regression that uses schooling and experience as explanatory variables—that also includes measures of ability derived from test scores as an explanatory variable.

Ability as the Dependent Variable

Measures of student ability based on test score data are common outcomes in both descriptive analyses (for example, of disparities across demographic groups) and evaluations of the causal effects of education programs or policies. A lesson of basic statistics is that classical measurement error in a regression's dependent variable will not lead to biased coefficient estimates, though it may reduce the precision of such estimates. When the available test score is of the fixed effects (maximum likelihood) type, this result is likely to apply. But none of the random effects approaches discussed above yield test scores that can be approximated as true ability plus classical measurement error, and regressions that use these type of estimates as dependent variables are likely to be biased. Consider, for example, estimation of the test score gap between poor and nonpoor students. If poor children have lower ability on average, then the poor/nonpoor gap in posterior mean scores (without conditioning) will understate the poverty achievement gap. The same is true when the available scores are plausible values, which are merely the sum of the posterior mean and a random component uncorrelated with student ability.

How potentially important is this bias? To measure this, we need to compare the biased estimates to unbiased results from the same test. Few databases of test scores report both random effects and fixed effects ability estimates. Some testing systems, however, report individual item responses. With these data, it is possible to obtain unbiased estimates by estimating a system of equations combining the item response theory measurement model together with the research model. This system specifies the likelihood for the observed item responses in terms of the item parameters and the research model coefficients, in essence using the research model covariates as the conditioning set. This approach, known as Marginal Maximum Likelihood (MML), is described in seminal articles by Mislevy (Mislevy 1991; Mislevy, Beaton, Kaplan, and Sheehan 1992).[6]

Briggs (2008) assesses the extent of bias in estimates of racial and ethnic gaps in student achievement that rely on posterior mean scores without conditioning variables. He uses a sample of 10th graders in 1999 who were administered the Partnership for the Assessment of Standards-based Science (PASS) test. Table 1 reproduces his estimates. Column 1 shows gaps (relative to whites) in scaled posterior mean scores. These indicate that the black–white achievement gap is –0.61 scale points. Column 2 shows unbiased Marginal Maximum Likelihood (MML) estimates. These indicate a black–white gap of –0.77 scale points in the same sample. Columns 3 and 4 report estimates for z-scores, created by dividing the scale scores by the standard deviation of these scores (column 3) or by the estimated standard deviation of latent proficiency (column 4). Again, the two sets of estimates give notably different answers: A black–white gap of –0.87 standard deviation units when posterior means are used, or –0.95 when computed via MML. Elsewhere, Briggs shows that the biases are even larger when considering subdomains within the larger test.

Another application where this issue has arisen is in the examination of teacher value-added, which is often computed via Empirical Bayes procedures. Chetty, Friedman, and Rockoff (2014a, Appendix Table 2) assess inequities in access to good teachers by regressing teacher value-added on observable student

[6] Implementing the model requires that the researcher invest some time in coding and in computational techniques (like Markov Chain Monte Carlo). The National Center for Education Statistics once contracted with the American Institutes of Research to develop software intended to estimate such models (at http://am.air.org/contact2.asp), although this software is now dated.

Table 1
Biases When Using Posterior Mean Test Scores as a Dependent Variable

	Logit units		*Z-scores*	
	Posterior mean (1)	*Marginal Maximum Likelihood* (2)	*Posterior mean* (3)	*Marginal Maximum Likelihood* (4)
Intercept	0.90	0.96	1.29	1.19
Black	–0.61	–0.77	–0.87	–0.95
Hispanic	–0.52	–0.67	–0.75	–0.83
Asian	–0.10	–0.12	–0.14	–0.14
Other	–0.30	–0.37	–0.43	–0.46
N	420	433	420	433
SD of test score	0.7	0.81	1	1

Source: Estimates are reproduced from Briggs (2008, Tables 4 and 6).
Notes: Data pertain to performance on a 10th grade science assessment. Columns 1 and 3 report estimates when posterior means (without conditioning variables) are used as the dependent variable in an ordinary least squares regression; Columns 2 and 4 report estimates obtained via the Marginal Maximum Likelihood method discussed in the text. In Columns 1–2, scores use the scale of a logit index (so that the probability of a correct answer equals the logit function applied to the scaled score with an additive adjustment); in Columns 3–4, these are divided by their estimated standard deviation. Briggs does not report standard errors, but all Intercept, Black, and Hispanic coefficients are significantly different from zero at the 1 percent level, while none of the Asian or Other coefficients are reported to be significant at the 5 percent level.

characteristics. They estimate that the value-added scores are shrunken by 36 percent, on average, and attempt to undo this by multiplying their estimated coefficients by 1.56 = 1/(1 – 0.36.) But this is only an approximation. Because the Empirical Bayes estimates are shrunken differently for each student, the bias need not be uniform.

The above discussion applies to random effects estimates of ability without conditioning variables. When a conditioning model is used, the potential biases become more complicated. As described above, the inclusion of conditioning variables can be thought of as shrinking a student's individual performance toward the group-specific mean for those sharing the characteristics of that student. Thus, only the portion of achievement that is not predicted by the conditioning variables is shrunken. An implication is that the coefficients in the research model are unbiased if all of the explanatory variables in the research model were also included in the conditioning model (Mislevy 1991). However, this is unlikely to be the case in many applications. Recall that the conditioning model used in the National Assessment of Educational Progress includes many student background characteristics but few, if any, variables that relate to education policies or programs. So if one regressed test scores on student background characteristics and, say, an indicator for whether the school had a high-stakes teacher evaluation system, the coefficient on the teacher evaluation system would be likely to be attenuated, and the student background coefficients might also be biased if the background and policy measures were correlated. Mislevy (1991) reanalyzes data from the 1984 NAEP Long-Term Trend reading assessment. He finds that biases in coefficients on variables not included in the conditioning model can be substantial.

But modern assessment systems typically include hundreds of variables in the conditioning model, many more than were used in the 1984 National Assessment of Educational Progress. It is not clear how important this type of bias is in today's NAEP. We have investigated this as it applies to two specific examples: an evaluation of the federal school accountability policy No Child Left Behind (Dee and Jacob 2011) and an assessment of the effects of school finance reform on inequalities in spending and achievement across districts (Lafortune, Rothstein, and Schanzenbach 2016), each of which relies on difference-in-differences regressions using state-by-year panels. We found that cross-sectional regressions of NAEP performance on either the state's accountability rule or the district's funding were insensitive to the use of plausible values versus a Marginal Maximum Likelihood (MML) approach. However, other studies that examine different policies or programs may show greater bias. We view this as an important subject for future research.

Ability as an Independent Variable

Now consider a research model in which ability is an independent variable: for example, a regression of wages on education, family background, and an ability measure (for example, Neal and Johnson 1996). Again, economists are generally familiar with the idea that classical measurement error in an explanatory variable leads to an attenuated coefficient on that variable—in this case, the ability measure—and to biases of predictable sign and magnitude in other coefficients. Once again, however, this result applies only to test scores generated by a fixed-effects method. By contrast, when test scores are posterior means or plausible values, measurement error in these scores is correlated (generally negatively) with the student's true ability. Intuitively, "shrinkage" estimators pull an examinee's reported score more toward the mean the further that person's true score is from the mean. Hence, classical measurement error results do not apply. In this setting, ordinary least squares coefficients are unbiased only in restrictive circumstances: for example when the test score is a posterior mean and the conditioning model includes the covariates from the research model but no other variables that are correlated with the research model outcome. Unlike in the dependent variable case, likely biases are quite different for posterior means than for plausible values, though as before they depend importantly on the presence and form of the conditioning model.

Schofield, Junker, Taylor, and Black (2015) model the likelihood of the outcome variable jointly with that for item responses. The resulting estimator, the "Mixed Effects Structural Equations" (MESE) model, is similar in spirit to the Marginal Maximum Likelihood (MML) approach discussed above, and permits unbiased

Table 2
Biases When Using Estimates of Latent Ability as an Independent Variable

	Dependent variable = log(weekly wage) *Estimate of literacy skill used in model*			
	No skill control *(1)*	*Maximum Likelihood Estimate* *(2)*	*Mixed Effects Structural Equation* *(3)*	*Plausible Values* *(4)*
Black	–0.366	–0.144	–0.094	–0.121
	(0.033)	(0.033)	(0.033)	(0.041)
Literacy skill		0.151	0.191	0.221
		(0.008)	(0.010)	(0.015)
Effect of a one SD change in literacy skill		0.19	0.218	0.221

Source: Estimates reproduced from Junker, Schofield, and Taylor (2012).
Notes: N = 3,267. In Columns 2 and 4, Maximum Likelihood Estimate and Plausible Values test scores (respectively) are entered as regressors in ordinary least squares regressions. Column 3 applies the Mixed Effects Structural Equation system-of-equations method. The research model in each column includes controls for a quartic in potential experience as well as indicators for urban status and census region.

estimation. As with MML, this requires both access to item responses and bespoke programming and computational methods.[7]

Junker, Schofield, and Taylor (2012) use this approach to assess the bias in a simple wage regression using data from the National Adult Literacy Survey, a nationally representative sample of US adults in 1992 that contains information on cognitive ability along with survey information on a variety of demographic and socioeconomic outcomes such as educational attainment and earnings. They focus on a subsample of 25–55 year-old men and women who work full-time, answered at least one item on the literacy test, report a weekly wage and self-report as black or non-Hispanic white. Their research model specifies log weekly wages as a linear function of race, a quartic in potential experience, indicators for urban status and census region, and the literacy test score. Table 2 reproduces their results for their sample of 3,267 men. Column 1 shows that the racial gap in wages is 36.6 log points (30.6 percent) without ability controls. Column 2 adds a maximum likelihood estimate of individual literacy, generated from a standard item response theory model. The implied black–white wage gap in this model drops dramatically to 14.4 log points (13.4 percent). However, recall from above that the literacy coefficient is attenuated due to classical measurement error in the maximum likelihood score, implying that the racial gap is overstated here. Column 3 presents unbiased Mixed

[7] Another approach, not pursued in the literature to our knowledge, would be to instrument for a noisy measure of ability with a second, independent, measure, if available. For example, the National Assessment of Educational Progress test consists of two separate blocks of items; one could use the fraction correct from the first block as an instrument for the fraction correct on the second.

Effects Structural Equations (MESE) estimates. As expected, the literacy coefficient increases and the implied black–white wage gap drops to 9.4 log points (9 percent). This finding suggests that latent ability accounts for 74 percent of the unconditional black–white log wage gap (= 1 – (–0.094/–0.366)) when properly controlled, but that a naive estimator would indicate that it accounts for only 61 percent of the gap.

As it happens, the National Adult Literacy Survey data report ability as a set of plausible values, based on a conditioning model that includes several hundred main effects and interactions of background variables collected in the survey. Importantly, the conditioning set includes measures of individual wages (the outcome variable in the research model above) as well as other measures highly related to wages such as family income and occupation, though the complex conditioning procedure makes it difficult to understand the functional form assigned to the relationship between ability and wages. Schofield et al. (2015) demonstrate that this sort of ability measure typically will result in bias. Indeed, the race coefficient when controlling for the plausible value scores (column 4) is –0.121, overstating the unbiased estimate by roughly 33 percent.

Again, this example makes clear the importance of sometimes obscure measurement choices in the construction of test scores to the substantive conclusions from secondary analysis of these scores. Regressions with test scores as dependent variables are plausibly unbiased when the score is constructed as a fixed effects estimate or as a random effects estimate with a sufficiently large conditioning set, but in nearly all other cases bias is likely. The most likely result is that the coefficients on key policy variables (which are unlikely to be included in conditioning models) will be attenuated, while those on demographic covariates will be overstated. When the test score is an independent variable, in the most common case using plausible values and conditioning on a wide range of predictors of individual ability (but not the dependent variable itself), we are aware of no general results on the sign or magnitude of bias. Information provided by the assessment—namely, the reliability of the ability measure—can in some cases be used to generate consistent estimates. Or even better, the reported scores can be discarded in favor of analyses that draw directly on examinees' item responses. However, it remains unusual for analytic samples from test score data to include item-level responses; in any event, few secondary analysts are likely to be willing to invest in the appropriate analysis of these responses, which remains tedious.

Conclusions

Modern psychometrics utilizes a variety of sophisticated models and techniques to develop cognitive assessments and produce individual ability scores. The applied researcher who does not possess at least a rudimentary understanding of these methods is liable to misuse test scores in a way that can lead to serious biases. These biases have not been widely recognized in the literature to date, and may be important to our understanding of key issues in education and labor economics.

In this concluding section, we discuss their implications for several of the running examples discussed throughout this article.

The black–white test score gap is a commonly cited statistic, used by educators and policymakers not only to judge specific schools or districts but also to evaluate the effectiveness of reform efforts. Recent studies using the nationally representative Early Childhood Longitudinal Survey (ECLS) have received considerable attention (for example, Fryer and Levitt 2004, 2006, 2013). As Bond and Lang (2013) illustrate, this statistic is quite sensitive to arbitrary decisions about how to scale test scores, and changes in the gap are particularly unstable depending on such choices. A less recognized concern, but perhaps as important, is that the ECLS test scores are posterior means generated without any conditioning variables—that is, individual ability measures in ECLS are shrunk toward the population mean. This almost certainly implies that the black–white scale score gap in ECLS is attenuated in the cross section, although we are not aware of any research that seeks to assess the magnitude of this bias.

Value-added measures are becoming increasingly common in education, health care, and other fields. Indeed, value-added measures of teacher effectiveness are currently used to evaluate teachers in many states, and value-added indicators of quality and cost effectiveness are used to reward hospitals as part of Medicare reforms in the recent Affordable Care Act. Choices about how to scale the outcome measure can have substantial impacts on the resulting statistic, and possibly important policy implications depending on exactly how such measures are used. As discussed above, we recommend that researchers assess the sensitivity of value-added measures by comparing the results of models that use scale scores with those that rely only on percentile ranks. We also caution researchers and policymakers to match more carefully the calculation of value-added (particularly the choice between fixed-effects-style estimators and random-effects-style predictors) to the use to which the scores will be put, as mismatches create biases of the forms discussed above.

Regressions that control for some measure of human capital are common in labor economics (for example, Neal and Johnson 1996). While measures of cognitive ability can be powerful controls in many models, estimated coefficients will be biased under typical conditions. If a maximum likelihood-based estimate of cognitive ability based on underlying test scores is used as a predictor, the coefficient on ability will likely be attenuated, and its relationship with other covariates will be under-controlled. In this case, if the test-maker reports the reliability of the test score, standard errors-in-variables results allow unbiased coefficients to be reverse-engineered. If the test score is instead derived from a random effects framework, either a posterior mean or a plausible value, the nature of the bias is much harder to determine as it depends on the other covariates in the model and the correlation between these covariates and ability. There are no simple fixes, other than to be cautious in interpreting results.

Finally, policy evaluations that use aggregate panel data (at the state-by-year level, for example) on student outcomes may be biased due to inappropriate construction of the underlying test scores. While the few cases that we have explored (specifications like those used in Dee and Jacob 2011 and Lafortune, Rothstein, and Schanzenbach 2016) do not seem to suffer from important biases, there is no guarantee of the same result in other contexts. In such cases, researchers must first make sure that they understand the cognitive ability data they are using well enough to recognize what biases might be relevant. Also, we suggest that researchers test the sensitivity of their results as much as possible. For example, with surveys such as the National Assessment of Educational Progress it is possible to obtain item-level data, with which one can either implement one of the more sophisticated approaches, such as the one suggested by Schofield et al. (2015), or a quick-and-dirty check such as testing robustness of results to using the fraction of items correct for each student as an alternative outcome.

The issues that arise in quantitative analysis of cognitive traits are only becoming more salient. The landscape of testing in US schools is changing rapidly, driven by the widespread adoption of the Common Core State Standards for K–12 education.[8] In spring 2015, more than half the states introduced new assessments to match the Common Core standards. These assessments all rely on sophisticated item response theory models to generate the exams and to calculate estimates of individual proficiency. One of the two new assessments (Smarter Balance) is computer-adaptive, so that a student who does well on early items is routed to hard items later in the test. This method can allow for more efficient estimation of student proficiency by ensuring that students are given many items that are appropriately difficult for them, but makes the resulting scores more sensitive to the underlying item response theory specification and measurement model.

There is some discussion of developing standardized assessments aimed at college students, too. Moreover, psychometric methods are spreading beyond cognitive skill assessment. Common measures of "noncognitive" traits such as persistence, self-esteem, and socio-emotional regulation, as well as of more cognitive traits such as working memory, rely on the same item response theory–based measurement models discussed above, typically applied to batteries of very few survey items (Schofield 2015). Test-score–like measures are also being used in health, as health care reform has encouraged increased emphasis on quantitative measurement. Across all of these domains, secondary researchers will need to account more carefully for scaling and measurement issues.

[8] The Common Core standards have been developed by a consortium of states, with strong encouragement from the federal government. They articulate in some detail what students should know and be able to do in each grade and subject in elementary and secondary school. A running theme is a reduced emphasis on memorization and rote computation, in favor of more problem-solving and higher-order thinking. Despite considerable controversy, as of August 2015, 42 states and the District of Columbia had adopted the Common Core standards in English/language arts and math.

References

Barlevy, Gadi and Derek Neal. 2012. "Pay for Percentile." *American Economic Review* 102(5): 1805–21.

Bond, Timothy N., and Kevin Lang. 2013. "The Evolution of the Black–White Test Score Gap in Grades K–3: The Fragility of Results." *Review of Economics and Statistics* 95(5): 1468–79.

Bond, Timothy N., and Kevin Lang. 2015. "The Black–White Education-Scaled Test-Score Gap in Grades K–7." October 18. http://people.bu.edu/lang/testgap-2.pdf.

Briggs, Derek C. 2008. "Using Explanatory Item Response Models to Analyze Group Differences in Science Achievement." *Applied Measurement in Education* 21(2): 89–118.

Carstens, Ralph, and Dirk Hastedt. 2010. "The Effect of Not Using Plausible Values When They Should Be: An Illustration Using TIMSS 2007 Grade 8 Mathematics Data." Presented at the 4th IEA International Research Conference (IRC-2010), July 1–3, 2010, at the University of Gothenburg, Sweden. Available at: http://www.iea.nl/irc-2010.html.

Cascio, Elizabeth U., and Ethan G. Lewis. 2006. "Schooling and the Armed Forces Qualifying Test: Evidence from School-Entry Laws." *Journal of Human Resources* 41(2): 294–318.

Cascio, Elizabeth U., and Douglas O. Staiger. 2012. "Knowledge, Tests, and Fadeout in Educational Interventions." NBER Working Paper 18038.

Chetty, Raj, John N. Friedman, and Jonah E. Rockoff. 2014a. "Measuring the Impacts of Teachers I: Evaluating Bias in Teacher Value-Added Estimates." *American Economic Review* 104(9): 2593–2632.

Chetty, Raj, John N. Friedman, and Jonah E. Rockoff. 2014b. "Measuring the Impacts of Teachers II: Teacher Value-Added and Student Outcomes in Adulthood." *American Economic Review* 104(9): 2633–79.

Cunha, Flavio and James J. Heckman. 2006. "Formulating, Identifying and Estimating the Technology of Cognitive and Noncognitive Skill Formation." *Journal of Human Resources* 43(4): 738–82.

Cunha, Flavio, James J. Heckman, and Susanne M. Schennach. 2010. "Estimating the Technology of Cognitive and Noncognitive Skill Formation." *Econometrica* 78(3): 883–931.

Dee, Thomas S., and Brian Jacob. 2011."The Impact of No Child Left Behind on Student Achievement." *Journal of Policy Analysis and Management* 30(3): 418–46.

Embretson, Susan E., and Steven P. Reise. 2000. *Item Response Theory for Psychologists.* Multivariate Applications Series. Lawrence Erlbaum Associates, Inc.

Fryer, Roland G., Jr., and Steven D. Levitt. 2004. "Understanding the Black–White Test Score Gap in the First Two Years of School." *Review of Economics and Statistics* 86(2): 447–64.

Fryer, Roland G., Jr., and Steven D. Levitt. 2006. "The Black–White Test Score Gap Through Third Grade." *American Law and Economics Review* 8(2): 249–81.

Fryer, Roland G., Jr., and Steven D. Levitt. 2013. "Testing for Racial Differences in the Mental Ability of Young Children." *American Economic Review* 103(2): 981–1005.

Hart, Betty, and Todd R. Risley. 1995. *Meaningful Differences in the Everyday Experience of Young American Children.* Paul H. Brookes Publishing.

Heckman, James, Seong Hyeok Moon, Rodrigo Pinto, Peter Savelyev, and Adam Yavitz. 2010. "Analyzing Social Experiments as Implemented: A Reexamination of the Evidence from the HighScope Perry Preschool Program." *Quantitative Economics* 1(1): 1–46.

Ho, Andrew Dean. 2009. "A Nonparametric Framework for Comparing Trends and Gaps Across Tests." *Journal of Educational and Behavioral Statistics* 34(2): 201–228.

Ho, Andrew D., and Edward H. Haertel. 2006. "Metric-Free Measures of Test Score Trends and Gaps with Policy-Relevant Examples." CSE Report 665. Technical Report, Graduate School of Education & Information Studies University of California, Los Angeles.

Junker, Brian, Lynne Steuerle Schofield, and Lowell J. Taylor. 2012. "The Use of Cognitive Ability Measures as Explanatory Variables in Regression Analysis." *IZA Journal of Labor Economics* 1(1): Article 4

Lafortune, Julien, Jesse Rothstein, and Diane Whitmore Schanzenbach. 2016. "School Finance Reform and the Distribution of Student Achievement." NBER Working Paper 22011.

Mislevy, Robert J. 1991. "Randomization-Based Inference about Latent Variables from Complex Samples." *Psychometrika* 56(2): 177–196.

Mislevy, Robert J., Albert E. Beaton, Bruce Kaplan, and Kathleen M. Sheehan. 1992. "Estimating Population Characteristics from Sparse Matrix Samples of Item Responses." *Journal of Educational Measurement* 29(2): 133–61.

Morris, Carl N. 1983. "Parametric Empirical Bayes Inference: Theory and Applications." *Journal of the American Statistical Association* 78(381): 47–55.

Neal, Derek A., and William R. Johnson. 1996. "The Role of Premarket Factors in Black–White Wage Differences." *Journal of Political Economy* 104(5): 869–95.

Nielsen, Eric R. 2015a. "Achievement Gap Estimates and Deviations from Cardinal Comparability." Finance and Economics Discussion Series Paper 2015-40, Board of Governors of the Federal Reserve System.

Nielsen, Eric R. 2015b. "The Income–Achievement Gap and Adult Outcome Inequality." Finance and Economics Discussion Series Paper 2015-041, Board of Governors of the Federal Reserve System.

Reardon, Sean. 2008. "Differential Growth in the Black–White Achievement Gap During Elementary School among Initially High- and Low-Scoring Students." Working Paper No. 2008-07, Institute for Research on Education Policy & Practice. http://web.stanford.edu/group/cepa/workingpapers/WORKING_PAPER_2008_07.pdf.

Rothstein, Jesse. 2010. "Teacher Quality in Educational Production: Tracking, Decay, and Student Achievement." *Quarterly Journal of Economics* 125(1): 175–214.

Rothstein, Jesse. 2016. "Revisiting the Impact of Teachers." Working paper. http://eml.berkeley.edu/~jrothst/CFR/rothstein_cfr.pdf.

Rothstein, Jesse, and Nathan Wozny. 2013. "Permanent Income and the Black–White Test Score Gap." *Journal of Human Resources* 48(3): 510–44.

Rubin, Donald B. 1987. *Multiple Imputation for Nonresponse in Surveys.* New York: John Wiley.

Rubin, Donald B. 1996. "Multiple Imputation After 18+ Years." *Journal of the American Statistical Association* 91(434): 473–89.

Schofield, Lynne Steuerle. 2015. "Correcting for Measurement Error in Latent Variables Used as Predictors." *Annals of Applied Statistics* 9(4): 2133–52.

Schofield, Lynne Steuerle, Brian Junker, Lowell J. Taylor, and Dan A. Black. 2015. "Predictive Inference Using Latent Variables with Covariates." *Psychometrika* 80(3): 727–47.

Stevens, S. S. 1946. "On the Theory of Scales of Measurement." *Science,* June 7, 103(2684): 677–80.

Thorndike, Robert L. 1966. "Intellectual Status and Intellectual Growth." *Journal of Educational Psychology* 57(3): 121–27.

van der Linden, Wim J., and Ronald K. Hambleton, eds. 1997. *Handbook of Modern Item Response Theory*. Springer.

von Davier, Matthias, Eugenio Gonzalez, and Robert J. Mislevy. 2009. "What Are Plausible Values and Why Are They Useful?" IERI Monograph Series. In *Issues and Methodologies in Large-Scale Assessments*, vol. 2, edited by Matthais von Davier and Dirk Hastedt, p.9–36.

Journal of Economic Perspectives—Volume 30, Number 3—Summer 2016—Pages 109–132

The Need for Accountability in Education in Developing Countries

Isaac M. Mbiti

Over the past two decades, developing countries have invested a considerable and rising portion of their GDP on education. A UNESCO (2011) report found that real education expenditures in a sample of 26 African countries grew by an average of 6 percent annually from 2000 to 2009. Similar patterns of education expenditure growth can be observed in South Asia, where the total education budget in India doubled between 2004 and 2009 (Muralidharan, Das, Holla, and Mophal 2016). As a result of this increased investment, countries in the sub-Saharan Africa and Latin America and the Caribbean regions spend about 5 and 4.6 percent of GDP on education, respectively, which compares favorably to North American and European countries that spend about 5.3 percent of GDP on education. However, south Asian countries such as India lag behind their African and Latin American counterparts by spending only 3.3 percent of GDP on education (UNESCO 2011). This rise in education spending in developing countries has mostly been channeled towards initiatives that improve schooling access, and school inputs such as classrooms, textbooks, and teachers. As a result, the global proportion of primary students who were out of school fell from 19 percent in 1999 to 11 percent in 2013 (UNESCO Institute of Statistics Database). Although enrollment rates in sub-Saharan Africa lag behind other regions, enrollment rates in primary

■ *Isaac M. Mbiti is Assistant Professor of Public Policy and Economics, Frank Batten School of Leadership and Public Policy, University of Virginia, Charlottesville, Virginia. He is also an Affiliate, Abdul Latif Jameel Poverty Action Lab (J-PAL), Cambridge, Massachusetts, and Research Fellow, Institute for Study of Labor (IZA), Bonn, Germany. His email address is imbiti@virginia.edu.*

† For supplementary materials such as appendices, datasets, and author disclosure statements, see the article page at
http://dx.doi.org/10.1257/jep.30.3.109 doi=10.1257/jep.30.3.109

school have risen from just 55 percent in the mid-1990s to almost 80 percent at present (based on the UNESCO Institute for Statistics database).

While education spending levels and enrollment rates in schools have increased across the developing world, a variety of research studies and datasets show that learning levels remain low. Roughly 50 percent of fifth-grade students could not read a second grade text in rural India, and only about 45 percent could correctly compute a two-digit second grade subtraction problem (Pratham 2014). In East African countries, only about 50 percent of fifth graders could read at a second-grade level in English, while only about 60 percent had attained basic second-grade numeracy and a slightly higher proportion could attain second-grade literacy in Kiswahili (Uwezo 2013). These data also show that these low learning levels have persisted over some time and are especially dire in rural areas, highlighting some of the pressing challenges facing many developing countries.

In addition to the low levels of learning, education systems (especially public systems) in developing countries are plagued by high rates of teacher absenteeism, leakages of financial transfers to schools, ineffective school monitoring systems, and poor parental engagement, which are all symptomatic of low levels of accountability in the system (according to the World Bank Service Delivery Indicators database; see also World Bank 2003). These low levels of accountability could dampen the effect of increased resource investment, which could help to explain why learning levels have been unresponsive to increased educational investment. In principle, education systems should be accountable to parents (and children). However, due to the centralized structure of the (public) education system and the nature of the political economy in developing countries, it is difficult for parents to hold education systems accountable through voting (the long route) or through direct action against public education service providers (the short route) (World Bank 2003). There is some optimism that the growth of the private education sector may increase accountability; more than 10 percent of students in sub-Saharan Africa and 20 percent in south Asia now attend private schools, which often cater to the poor (World Bank EdStats Database; Heyneman and Stern 2013). A growing body of research has shown that private schools employing lower-paid teachers, who face different incentives compared to their public school counterparts and in some cases are provided with improved technological support, are often able to deliver similar or better student results at markedly lower costs (for example, Muralidharan and Sundararaman 2015). However, the potential for the private school market to improve educational accountability depends on a number of factors, including the thickness of the market, the quality of information available to parents, and government policy and regulation.

A growing body of literature has used empirical methods such as randomized control trials and regression discontinuity designs to examine the effectiveness of various interventions in the education systems of developing countries on student outcomes. There are now multiple review papers and meta-analyses of this literature, including Conn (2014); McEwan (2015); Glewwe Hanushek, Humpage, and Ravina (2014); Glewwe and Muralidharan (2015); Kremer, Brannen and

Glennester (2013); and Murnane and Ganimiaan (2014). There has even been a systematic review of reviews by Evans and Popova (2015), which sheds light on some of the divergent findings and recommendations put forward in the aforementioned reviews. Across a variety of contexts, these reviews generally show that input-based policies on their own are largely ineffective in increasing learning outcomes in the absence of complementary initiatives to improve accountability or pedagogy.

However, shifting the focus of education systems in developing countries from primarily input-based policy towards policies that focus on outcomes such as learning is extremely challenging due to the political economy of education service delivery. There is evidence that curriculums often focus on the needs of the top-performing children and the children of elites rather than the median child (Glewwe, Kremer, and Moulin 2009; Banerjee and Duflo 2010). Relative to developed countries, per pupil spending in developing countries is heavily skewed toward tertiary education, which only a select few can access, rather than primary education (as shown in the World Bank EdStats database). In addition, a number of authors have documented examples of elite capture of education resources such as new school construction in Kenya (Kramon and Posner 2016), and school finances in Uganda (Reinikka and Svensson 2004).

Education-related visions, plans, and promises often occupy prominent positions in public debates and the promises of politicians in developing countries. However, there is little overlap between the campaign promises and the policies shown to be effective in the research literature. Across a number of countries, these promises (and the resulting policies) typically focused on highly visible education inputs such as building schools, reducing school fees, offering more loans and scholarships, purchasing computers, raising teacher pay, and reducing class size, rather than less visible but more effective reforms that increase learning through improved accountability and pedagogy. As Rukmini Banerji (2014), who directs the Indian nongovernment organization Pratham, has noted: "Parents can easily discuss issues of access to schooling and debate and argue about inputs and entitlements that their children are supposed to receive as a result of going to school. But discussions focused on learning are neither easy nor automatic." Her assertion is corroborated by Harding and Stasavage (2014), who use data from a number of African countries to show that policies that improve school quality do not affect electoral support, whereas policies that reduce school fees, especially primary fees, resonate with voters (see also Stasavage 2005)

In this paper, I first review some evidence on the effects of inputs to education in developing countries, such as teachers and textbooks. I then examine the need for accountability across different areas of the education system. I further examine potential pathways to improving accountability among teachers, school management, and parents. Because many developing countries have experienced a dramatic rise in private school enrollments, I discuss the potential for the market to improve accountability in developing countries, highlighting the emergence of low-cost private schools and the innovations and controversies surrounding their business models. Given the political economy challenges of reforming the system, I will at various points seek to assess the potential for education policy to be reoriented towards learning outcomes.

The Impact of Increasing Education Inputs in Low Accountability Contexts

Classroom Inputs

Despite the increases in education investment, many classrooms in developing countries continue to face real resource constraints. Average pupil-teacher ratios in Malawi, Chad and Rwanda were at least 60:1, while Pakistan, Cambodia, Uganda, Tanzania, Burkina Faso all had ratios over 40:1 (from the World Bank World Development Indicators database). The average pupil-teacher ratio in India was approximately 35:1 in 2011 (according to the World Development Indicators), however, these ratios could reach as high as 90:1 in rural areas (Pratham 2013). Some schools operate in two shifts—one morning, one afternoon—which reduces scheduled classroom instructional time for students to approximately three hours per day (World Bank Service Delivery Indicators database). Only 25 percent of primary schools in sub-Saharan Africa had electricity, while 68 percent had toilets, and approximately 50 percent had access to potable water. When textbooks are available, they are often shared by two, three, or more students (World Bank EdStats database).

But perhaps surprisingly, given the low levels of resources found in schools in developing countries, interventions that provide inputs or resources such as school grants, flipcharts, or textbooks rarely improve learning outcomes. A variety of randomized experimental studies have reached this conclusion. For example, Glewwe, Kremer, Moulin, and Zitzewitz (2004) find that providing flip charts in rural Kenyan schools did not improve student outcomes. Randomized evaluations of textbook provision programs also find limited increases in learning outcomes. For example, Glewwe, Kremer, and Moulin (2009) argue that the English language used in the textbooks was not appropriate for most children in rural Kenya who tend to have limited exposure to English at home. Sabarwal, Evans, and Marshak (2014) argue that the uncertainty of future resource (or input) flows encourages schools in Senegal to engage in a type of precautionary savings behavior where they store the books for future use rather than distribute them to students.

In an experiment in 350 Tanzanian primary schools, Mbiti, Muralidharan, Romero, Schipper, Rajani, and Manda (2016) found that school grants that doubled per pupil spending were ineffective in increasing learning outcomes, unless the grants were coupled with teacher incentives. Das, Dercon, Habryarimana, Krishnan, Muralidhanan, and Sundararaman (2013) found that school grants given to schools in Zambia and India were completely offset by reductions in parental education expenditures. Experimental studies also show that computer resources that fail to target instruction also rarely boost learning outcomes, although they promote familiarity with computers (Cristia, Ibarraran, Cueto, Santiago, and Severin 2012; Kremer, Brannen, and Glennerster 2013). Overall, these studies suggest that the effectiveness of increased inputs may be hampered by behavioral responses by parents or head-teachers, and the lack of accountability. Political pressure to institute visible education policies may also lead education systems to invest in less effective inputs.

Another reason why overall increases in education spending by the central government have had limited impact on student outcomes is that often a substantial share of the earmarked funds does not reach schools. An extreme case of leakage was documented in Uganda in the mid-1990s, when only about 22 percent of allocated funds reached schools after local politicians diverted the funds to their election campaigns (Reinikka and Svensson 2004). The capture of education funds by local politicians again highlights the importance of political economy and accountability concerns in these settings. In short, these data suggest that there is very limited accountability in the management of education resources. Moreover, preventing such leakage would involve improved transparency coupled with reforms that strengthen the monitoring capacity and governance ability of key stakeholders including central government, local government, school committees, principals, and parents.

Pupil/Teacher Ratios and Teacher Pay

Given the large pupil/teacher ratios found in developing countries, there is often pressure on governments to hire more teachers to reduce class sizes. In an experiment in western Kenya, Duflo, Dupas, and Kremer (2012) find that lowering class size by adding more centrally hired civil service teachers did not improve student learning outcomes. Instead, existing teachers reduced their effort in response to the new hires, and helped to get their relatives hired into a significant portion of these new teaching slots. Even though the bulk of primary and secondary education spending in developing countries is allocated to teacher and staff payroll, governments often face pressure to increase teacher remuneration. Almost 90 percent of the education budget in India, Jamaica, Pakistan and Togo was devoted to teachers and staff (World Bank EdStats database; Muralidharan, Das, Holla, and Mophal 2016). On a per person basis, primary school teacher salaries in sub-Saharan Africa were on average four times per capita GDP, whereas the OECD average teacher salary was at most 1.3 times per capita GDP (author's calculations using data from UNESCO 2011 and OECD Online Education Database).

Teachers' unions often argue that teachers need better pay to be more effective, but there is limited evidence to support this claim. A policy change in Indonesia permanently doubled salaries for teachers who met certain certification criteria, and de Ree, Muralidharan, Pradhan and Rogers (2015) use a randomized phase-in design across a large sample of teachers to evaluate the program. They find teacher satisfaction increased, but there was no discernable impact on teacher effort or student learning two to three years after the reform. Using a regression discontinuity design based on geographic boundaries, Pugatch and Schroeder (2014a, b) examine the effect of hardship allowances which increased teacher pay by 30–40 percent in remote areas in Gambia. While the pay increase increased the number (and proportion) of qualified teachers in remote areas, there was no resulting increase in average test scores (although they do find evidence that the program benefited the top-performing students). Given that many developing country education systems lack accountability and teachers are unlikely to be dismissed for poor performance, it seems plausible that pay increases are mostly a transfer to teachers, because they do not lead to increases in teacher effort or performance. However, increases in remuneration could yield some improvements in the long-run if they attract more able and potentially more motivated individuals to the teaching profession.

The Need for Accountability Among Teachers

In developing countries, teachers are typically civil service workers, often unionized, who are hired and paid directly by a central authority which has ultimate authority on teacher staffing. This centralized system makes it very difficult for parents and even school principals to hold teachers accountable. Consequently, documented measures of quality teaching are quite low across many countries. Teacher absence is a pervasive issue in many developing countries. Almost one-quarter of teachers were absent from schools on a given day in India, Tanzania, and Uganda, while just over 15 percent were absent from schools in Senegal and Kenya (Muralidharan, Das, Holla, and Mophal 2016; World Bank Service Delivery Indicators database).

Even teachers who are on the school grounds school seem to spend considerable time in the staff room drinking tea or conversing with each other (or visitors), rather than in the classroom. Approximately 50 percent of Tanzanian and Ugandan teachers were not in the classroom (as reported in the World Bank Service Delivery Indicators database). As a result of these high rates of absence, the actual average instructional time in schools was limited, ranging from two hours per day in Tanzania, to about three hours and 15 minutes per day in Uganda and Senegal. Teacher absence also imposes negative externalities on other teachers and students. Nearby teachers are often obligated to check in on the unattended classroom or integrate the unattended students into their classrooms (sometimes resulting in multi-grade classrooms). Despite the high levels of teacher absence, not a single teacher in a sample of Indian public schools had been dismissed during the tenure of the principal (as shown in data from the Young Lives India study at http://www.younglives-india.org).

In addition, school inspectors who monitor schools to ensure compliance with education standards and regulations rarely seem to focus on the most pressing issues. For example, schools in Tanzania were visited about twice a year by ministry of education officials. These visits were mainly administrative, often to collect information such as enrollment or to deliver exams. Only 30 percent of schools report that the most recent inspection visit focused on teaching and learning. During a recent visit to a school in Tanzania, I was accompanied by a quality assurance officer. Although several teachers were absent from the school, the officer did not report this fact, but rather complained to the principal that the students were speaking a local language rather than Kiswahili, the official language of instruction. Given the high levels of expenditures on teachers in developing countries, Muralidharan, Das, Holla, and Mohpal (2016) argue that investing in more effective teacher monitoring and accountability systems could significantly increase the productivity of the education budget by reducing the high levels of teacher absenteeism and encouraging greater teacher effort. They argue that absenteeism costs Indian taxpayers the equivalent of over US$1.5 billion per year.

Several studies show that teacher absenteeism responds to incentives—although not always in the desired manner. For example, evidence from Kenya and India shows that when there are more teachers, or a lower pupil-teacher ratio, absence rates are typically higher (Duflo, Dupas, and Kremer 2011; Muralidharan, Das, Holla, and Mohpal 2016). This finding may help to explain why simply adding more teachers without changes in the accountability structure has such a disappointingly small effect on student outcomes.

Randomized experiments in India and Kenya have demonstrated that teachers who are hired directly by the school on short-term contracts can improve student test score outcomes (Duflo, Dupas, and Kremer 2011, 2012; Muralidharan and Sundararaman 2013). Because contract teachers face stronger incentives to deliver quality teaching relative to their civil service counterparts, they are more likely to be at school, to be in the classroom teaching, and to deliver better or a least similar learning outcomes compared to civil service teachers, all while being paid between one-fifth to one-third the salary of their government counterparts. However, proposals to formalize policies around greater use of contract teachers have met heavy opposition from teachers' unions. There are additional concerns that scaling up such a program through the "business as usual" government procedures may undermine its effectiveness. Building on the experiment by Duflo, Dupas, and Kremer (2012) in Western Kenya, Bold, Kimenyi, Mwabu, Ng'ang'a and Sandefur (2013) evaluate a larger experiment in Kenya which scaled up the contract teacher program to nearly 200 schools across all provinces of Kenya. The study compared the effectiveness of the program when it was administered by the government rather than a non-government organization (as was the case in Duflo, Dupas, and Kremer 2012). They find that the benefits of the program completely disappeared when administered by the government rather than a non-government organization, highlighting the challenge of scaling up promising interventions through government systems that lack accountability and (in this case) implementation fidelity.

While improving incentives by altering the contractual structure of teachers is politically difficult, a growing body of experimental research has demonstrated the potential for providing teachers with financial incentives to improve learning outcomes. Muralidharan and Sundararaman (2011) found gains in student outcomes in an experiment in rural primary schools in the Indian state of Andhra Pradesh, where teachers were awarded bonus payments based on the improvement of their students' test scores. Loyalka, Sylvia, Liu, Chu, and Shi (2016) also found student gains from an experiment tying teacher pay to student performance in 216 schools in western China, using a variety of incentive designs. Duflo, Hanna, and Ryan (2012) carried out a randomized study in India, where some teachers were given a digital camera, and received a financial incentive for taking a time-stamped picture of themselves with their class at the beginning and end of the school day. The incentives, coupled with monitoring by the camera, reduced teacher absenteeism and improved student outcomes.

However, while teacher incentive schemes can increase accountability by aligning teacher effort with student outcomes, they are often insufficient in raising learning outcomes when they are introduced as stand-alone interventions, as there may be additional binding constraints. For example, teachers' incentives may be complementary to other classroom inputs (as found in the experiment of Mbiti, Muralidharan, Romero, Schipper, Rajani, and Manda 2016) or to student effort (Behrman, Parker, Todd, and Wolpin 2012). In addition, the design of the incentive scheme is an important factor in determining the effectiveness of such schemes. Economic theory suggests the most effective schemes will feature individual incentives and payoffs that are based on student growth and elicit effort across the entire student distribution, such as the "pay for percentile" scheme described by Barlevy and Neal (2012), and experimentally evaluated in Chinese villages by Loyalka, Sylvia, Liu, Chu, and Shi (2016). However, in practice there may be a tradeoff between the transparency and ease of comprehension of the incentive design on one hand, and the power of the incentive on the other.

Yet another difficulty with plans to link teacher pay to student performance is that many teachers may be limited by their knowledge of their subject(s) and pedagogical techniques. Consequently, teacher incentive programs may not be sufficient to improve learning outcomes as the increased effort by teachers may not be directed towards effective activities. Using linked teacher-student databases from Peru (Meltzer and Woessman 2012), and from 13 different sub-Saharan African countries (Bietenbeck, Piopiunik, and Weiderhold 2015), the authors find that teacher subject knowledge is correlated with student learning outcomes. However, data from a variety of settings suggest that teacher subject knowledge is quite limited. In Kenya, sixth grade math teachers scored about 50 percent on an externally administered grade appropriate math exam (Ngware, Ciera, Musyoka, and Oketch 2015). About 40 percent of teachers in Kenya, 20 percent of teachers in Uganda, 5 percent of teachers in Senegal, and 1.2 percent of teachers in Tanzania had the "minimum knowledge needed to be effective" (data for 2012 from the World Bank Service Delivery Indicators).

Lessons by teachers are generally not interactive—and this lack of interaction may be more common among teachers who are not as comfortable with the material. I have observed teachers spending close to 30 minutes drawing science diagrams on the board, with absolutely no interaction with the class. Much of the time students are asked to solve problems, while the teachers sit at the front of the room without interacting with the class.[1]

[1] Detailed micro-data on teaching practices and teacher knowledge are available through the World Bank Service Delivery Indicators data set for Kenya, Tanzania, Uganda, and Senegal. For example, there

Corporal punishment is common. When I observed classes in Kenya, teachers were often seen walking around with intimidating foot-long PVC pipes which they use as a pointer on the blackboard but also to cane students. Tabulations from the Young Lives database (at http://www.younglives-india.org) show that almost one-half of the students surveyed in India had been beaten in the week prior to the survey, while one-third of Ethiopian students, just over one-quarter of Peruvian students, and around one in six students in Vietnam had been punished in a similar time frame. Taken together, these data suggest that there a number of ways in which teachers could alter their actions to improve the learning environment.

Teacher training programs are an obvious approach to address teachers' inadequate knowledge of their subjects and instructional methods. Research on teacher training in developing countries is limited, but there is a growing body of literature on "scaffolding" instruction programs. These programs provide step-by step instructional methods for teachers, and in some cases even include daily lesson plans. Well-designed scaffolding programs are a generally a cost-effective approach to improving learning outcomes as they mitigate limited teacher subject knowledge and pedagogical skills. For example, Lucas, McEwan, Ngware, and Oketch (2014) show gains to student learning in Uganda from a randomized evaluation of the "Reading to Learn" curriculum, which takes a scaffolding approach to teaching literacy, and ongoing teacher support. Piper, Zuilkowski, and Mugenda (2014) use a randomized controlled trial in over 500 schools in Kenya to evaluate a scaffolding-style program of teacher training for early grade learning called PRIMR. The results on early grade reading and numeracy were so promising that the Kenyan government implemented the reading program in all public primary schools. Critics argue that scaffolding can be too restrictive or constraining, especially for effective teachers. But the approach need not be mandatory for all to be useful for many.

Since education systems are often oriented toward top-performing students, interventions that support the teacher's ability to adapt to their students' level of preparation across the range of performance may be complementary to accountability programs. In an experiment in schools in an urban setting in India, Banerjee, Cole, Duflo, and Linden (2007) find that hiring young women as tutors in literacy for students who had fallen behind or using computer-aided adaptive learning for math are cost-effective ways of raising student outcomes. Duflo, Dupas, and Kremer (2011) conducted an experiment in 121 schools in western Kenya, where students were tracked based on their past performance. They find that tracking helped lower-performing students in particular, because it gave teachers a rationale for teaching them at their own level. This change is more significant than it may sound, as the norm among many teachers in developing countries is to finish the syllabus, regardless of the actual learning progression of students. When this practice is combined with the automatic grade-to-grade promotion rules that have been implemented in many countries, a significant portion of students end up leaving primary school

are comparisons of specific teaching practices between civil service teachers and contract teachers, as well as between teachers in public and private schools.

without acquiring basic competencies in numeracy and literacy. For example, data from Tanzania show that across all subjects approximately 83 percent of Tanzanian teachers in first, second and third grade covered the entire material in the syllabus in a year, yet 25 percent, 47 percent, and 17 percent of seventh-grade students failed a second-grade exam in Kiswahili, English, and Math respectively (Twaweza 2013; Uwezo 2013).

The recent scale-up of the PRIMR program in Kenya provides an illustration of how learning-centered education reforms can be enacted. In this case, the program had support from teachers' unions, government, nongovernment organizations, and donors such as USAID and the UK Department for International Development (DFID). The program likely garnered broad support because it provided a combination of visible inputs such as new student textbooks and instructional materials for teachers, as well as less visible changes in pedagogy and ongoing teacher support. Future research should focus on evaluating the complementarities between teacher incentive programs (broadly defined) and interventions that support teachers' ability to teach all students, accounting for the various political economy and accountability challenges that may continue to bind. Research that illuminates the challenges of scaling up programs and potential solutions for addressing those challenges is especially important.

The Need for Improved Accountability and Resource Management in Schools

Schools in developing countries are usually managed by principals in conjunction with local school management committees which consist of teachers, parents, and community members. Principals are generally more educated than teachers. For instance, almost 45 percent of principals in the Young Lives sample of Indian schools for 2012 had a master's degree and 43 percent had a college degree, whereas only 19 percent of teachers had a master's degree and 58 percent had a college degree (at http://www.younglives-india.org). But despite the higher education level of principals, school management capacity is relatively weak. Two-thirds of principals in the Young Lives India sample utilized in-person meetings with teachers as their primary method of monitoring. In this data, principals in India believe that the most important indicators of good schools are observable inputs such as buildings, geographical accessibility, and the availability of teaching materials. Only 11 percent of principals believe that learning outcomes (or exam results) are the most important indicator of a good school. Further, only 13 percent of public school principals in the survey in 2012 conducted unannounced teaching observations, while only 8 percent report using student learning outcomes to monitor teacher performance.

Such skewed perceptions of quality suggest that effective management training for principals could have large impacts on schools. However, data from Tanzania show that only 22 percent of principals attended a school management training in the past five years, while in 2012 just over 67 percent of principals in a Peru survey and 78 percent of principals in an Indian survey had attended school management trainings (Twaweza 2013; Young Lives Database at http://www.younglives.org.uk). In an experiment in Senegal that provided schools with grants, Carneiro, Koussihouede, Lahire, Meghir, and Mommaerts (2015) find that schools that invested in materials saw limited improvements in learning outcomes, whereas schools that invested in programs that increased management and teacher productivity through training programs saw improvements in learning. Such training programs may also be more effective if coupled with reforms that incentivize increased oversight effort among principals. However, given the mixed evidence on training programs for schools, more research is needed to enhance our understanding of how to design these programs.

Principals and school committees are jointly responsible for managing school finances. Following the reduction or elimination of school fees in public primary schools in many African countries, governments instituted capitation grants to replace the previously collected school fees (Lucas and Mbiti 2012). These grants are transferred from the central government to schools, although sometimes they are routed through intermediary institutions such as local governments or ministry of education departments. Coupled with the irregularity and uncertainty about the flow of funds, there was considerable confusion about the funding policies in many contexts. Almost 60 percent of principals in the Tanzanian survey did not know how much they were eligible to receive from the government, while 35 percent of Kenyan principals did not know the size of the capitation grant for nonteaching expenses (Twaweza 2013; World Bank Service Delivery Indicators database for 2012). In Tanzania, only 55 percent of principals had a manual that explained the capitation grant policy, and 64 percent kept organized financial records (Twaweza 2013).

This financing structure does little to encourage quality teaching, because better-performing schools are unlikely to receive additional resources given the uncertainty and irregularity of resource flows from the government such as grants and additional teachers. Also, as Kremer, Moulin, and Namuyu (2003) argue, efforts to improve school performance may be undermined if they are offset by increased student enrollment. In addition, schools have limited discretion on spending, and so may not be able to channel their resources efficiently. For instance, almost 95 percent of schools in Kenya are given specific instructions on what materials to purchase from government officials, and 86 percent report having no discretionary funds at all (World Bank Service Delivery Indicators data for 2013). Pairing school finances with head teacher incentives may be a promising approach to encourage the more efficient use of school resources. In a randomized study in Tanzanian primary schools, Mbiti, Muralidharan, Romero, Schipper, Rajani, and Manda (2016) find that school grants were quite effective at improving learning outcomes when paired with teacher and head-teacher incentives. They argue that the combination of incentives and resources encouraged schools to invest their available resources more efficiently.

Training, empowering, and funding school committees are potential approaches to improving school management practices. However, most evaluations of school management training have found that they are generally ineffective, at least as standalone interventions. For example, Blimpo, Evans, and Lahire (2015) conduct an experiment in 273 Gambian primary schools where school management committees in the treatment group received additional funds, training, or both interventions. There was some effect in reducing student and teacher absence, but no effect on student outcomes. There is some evidence that empowering school management committees may help student performance. In a multi-treatment experiment in 520 Indonesian public schools, Pradhan, Suryadarma, Beaty, Wong, Gaduh, Alisjahbana and Artha (2014) evaluate the effectiveness of increasing the legitimacy of the school committee through elections. They find that elections for school committees (coupled with school grants) improved teacher effort and parental engagement, but did not raise learning outcomes. However, they find that building linkages from the school committee to the powerful village council improved learning outcomes. Their study suggests that policies that solely increase the accountability of school committees may not be sufficient to improve learning, as school committees have limited power to enact change without additional support. Decentralization is often proposed as a solution to improve accountability. However, the evidence from the randomized studies discussed above show that decentralization initiatives, such as providing school committees with more funding, would need to be coupled with additional programs to facilitate effective and accountable local management. This is another area for future research. Such studies should also examine how to best empower and support school principals. To the extent possible, these studies should also be conducted at scale to facilitate the examination of market-level responses.

Accountability through Parents

Parental engagement can play a large complementary role in education production of children. Parents can hold schools and teachers accountable by voicing concerns, or even by moving their children to another school. They can support the school's fundraising efforts, and can also support their children directly at home.

However, many parents do not seem to be well-informed. A survey in Tanzania found that only 20 percent of parents knew what their child had scored on their last math, English, or Swahili test. Only 48 percent of parents received a report from the school about their child's performance. Enrollments per grade were around 110, but 45 percent of parents reported that their child was in the ranked among the top ten children in the grade, which suggests that most parents were overestimating their child's performance (Twaweza 2013). Parents were also not well-informed about education finance policy at schools. Tanzanian primary schools are supposed to receive capitation grants worth 10,000 shillings per child from the central government to cover the school's (non-teacher-related) operating expenses such as administration, minor repairs, and input purchases such as textbooks. However, only 13 percent of parents knew what a capitation grant was and only 3 percent of parents knew the amount of money that schools were meant to receive. Moreover, parents had limited interactions with schools. About two-thirds of households had no discussions with teachers in the previous year. Just over one-half of parents in Tanzania attended a meeting at the school in the previous year, but the main topics of discussion were academic performance (usually about the national exams in fourth and seventh grade) and fundraising. Almost 70 percent of parents contributed to schools by donating either financially, in-kind, or with their labor. Overall, these levels of interaction are higher than those documented in India by Banerjee, Banerji, Duflo, Glennerster, and Khemani (2010). In their study in 280 villages in Uttar Pradesh, they find that only 6 percent of households donated to schools, 8 percent volunteered at school, and 28 percent visited the school to complain or monitor.

Increased parental (or community) involvement in school management could potentially improve accountability. A common low-cost approach is to provide parents with information about the school, usually through some form of report card. However, there is limited evidence on the effectiveness of providing such information. For example, Banerjee, Banerji, Duflo, Glennerster, and Khemani (2010) carried out an experiment in India where parents were provided with information about learning outcomes, and community members were trained on a testing tool for children. They find that the information intervention did not improve student learning. Lieberman, Posner, and Tsai (2014) carried out an experiment in 26 Kenyan villages where parents received information about their child's performance and materials about how to help, but found no effect on student outcomes. In contrast, Reinikka and Svensson (2005) studied a newspaper campaign in Uganda that provided schools and parents with information so that they could to monitor how local officials were managing a large education grant, and argue that it reduced the capture of these funds and measurably improved student enrollment and learning outcomes.

One reason that providing information may be insufficient to affect outcomes is that parents may have limited avenues to affect the education system. The low levels of parental engagement, and the general ineffectiveness of information campaigns could be a rational response by parents, who, perhaps correctly, surmise that their voice, pressure, and engagement will have little impact as they have limited avenues to hold public schools accountable. Indeed, the Banerjee, Banerji, Duflo, Glennerster, and Khemani (2010) study in India found that report cards paired with a training program on how to conduct summer reading camps did lead to improved learning outcomes among camp attendees as it provided parents with a specific course of action to address the issues raised in the report card.

Collective action problems are also important barriers to parental action, and these may be amplified by ethnic and social divisions within the community. Focusing on a sample of schools in western Kenya, Miguel and Gugerty (2005) find that as the community diversity increased, parental contributions to schools decreased as it was harder to coordinate in order to impose social sanctions on parents that did not contribute. However, the relationship between ethnicity and parental contributions was relatively muted in Tanzania, as ethnicity is less salient there relative to neighboring Kenya. There is growing evidence that collective action problems can be overcome. Barr, Mugisha, Seernels, and Zietlin (2012) analyze an experiment involving 100 primary schools in Uganda, where parents played an active role in deciding on their own objectives, roles, and indicators of progress for monitoring schools, and found that this process was associated with improved student outcomes as it alleviated collective action problems. Studies that shed light on potential pathways to reduce collective action problems and which provide parents with specific avenues to effect changes in schools would be productive avenues for future research, especially if conducted at scale.

The Potential of the Private Schools and Market Competition to Provide Accountability

Private school enrollment rates have been growing slowly, but steadily, in many developing countries. In the South Asia region, private schools account for around one-fifth of all primary school enrollment (according to the World Bank EdStats database). Andrabi, Das, and Khwaja (2008) show that the number of private schools in Pakistan increased by a factor of ten in less than two decades, with most of the growth in the 1990s. The share of primary school students in private schools is more than 15 percent in Latin America and exceeds 10 percent in the sub-Saharan Africa region. While the share of primary school student in private schools is only about 7–8 percent in the Middle East and East Asia/Pacific reasons, this level is double what it was 25 years ago (again, according to World Bank EdStats).

The rise of private schools is partly driven by parental beliefs about the relative quality of private schools, which may be a consequence of the low accountability in public schools. The shift toward reducing school fees for public education, along with rising enrollments, caused some parents to seek private schools instead. Lucas and Mbiti (2012) show that the introduction of free primary education in Kenya increased the demand for private schooling, especially in districts with higher levels of economic inequality, which is perhaps suggestive of parental preferences for peer groups.

Given the myriad of challenges faced by public schools in developing countries, a key policy question is the extent to which the private sector can provide more accountability in the education system. By relying on school fees, private schools are possibly more accountable to parents. In addition, private schools may be better placed to deliver better quality education, as measured by learning outcomes, and could generate positive (or negative) spillovers to the public sector through greater competition. The potential effects of private schools depend critically on factors such as the market structure, information constraints, parental preferences, and government policy.

There is considerable heterogeneity in private schools in developing countries, ranging from elite institutions that cater the richest households to low-cost private schools that operate in disadvantaged areas such as urban slums which are typically underserved by public schools and other public services. There is also substantial product differentiation in this sector. For example, private schools in Pakistan and India offered different languages of instruction and different subjects, suggesting that they are responsive to market demand (Muralidharan and Sundararaman 2015; Andrabi, Das, Khwaja, Vishwanath, and Zajonc 2008). A disproportionate share of the recent growth in private school enrollment has actually been in private schools that cater to the poor, as discussed in the Heyneman and Stern (2013) case studies of low-fee private schools in Jamaica, Kenya, Tanzania, Ghana, Indonesia, and Pakistan.[2] These schools are typically located in lower-income, densely populated urban areas—even in slums—but were also prevalent in peri-urban and more rural settings. For instance, Muralidharan and Sundararaman (2015) find that 35 percent of students in rural Andra Pradesh (the fifth-largest state in India) attended a private school. A multi-country school census in low-income areas conducted by Tooley and Dixon (2005) found that 65 percent of schools in Hyderabad in India and the state of Lagos in Nigeria were privately run, while 75 percent of schools were private in Ga district, a peri-urban and somewhat rural district in Ghana. Oketch, Mutisya, Ngware, and Ezeh (2010) show that over 90 percent of schools in two slums of Nairobi, Kenya, were private. Because many of these schools were not formally registered (or recognized) by the government, official statistics may underestimate private sector enrollment rates.

Private school fees vary but were often modest, with the unregistered schools charging less than registered private schools. In the Ga district in Ghana, unregistered private schools charged US$14 per term on average (roughly $5 per month), while registered schools charged US$24 per term on average (roughly $8 per month). Using a comprehensive school census from Pakistan, Andrabi, Das, and Khwaja (2008) find that rural private schools charged an average of US$17 per year in fees, while urban schools charged US$27 per year.

Although these fees seem modest, there are concerns that the growth of private schools may exacerbate social inequalities (even in rural areas or slums) by excluding the very poorest households, girls, and disadvantaged groups such as ethnic minorities or lower-caste groups. Across different contexts, the data generally show that students who attend private schools come from relatively wealthier households, with better-educated parents (for example, Andrabi, Das, and Khwaja 2008; Muralidharan and Sundararaman 2015; Singh 2015). However, digging deeper into the data, access to private schools among the poorest is clearly quite high. Survey data from Lahore, Pakistan and two slums in Nairobi, Kenya show that 37 percent of children from households at or below the 15th percentile of the

[2] In general, these areas are not well served by public services such as education or sanitation; often a consequence of the limited or nebulous property rights in informal settlements (Marx, Stoker, and Suri 2013). Because a public school has to be set up on land with a title deed, and has to fulfill various rules (say, having sufficient acreage for a playground), the limited presence of public schools in low-income areas has created an opportunity for the private sector.

wealth distribution and 43 percent of children from the poorest quintile of households attended private schools, respectively (Alderman, Orazem, and Paterno 2001; Oketch, Mutisya, Ngware, and Ezeh 2010). This relatively high rate of private enrollment among the poor may in part reflect the lack of government school options in urban slums and other disadvantaged areas. There is less consistency regarding gender patterns in enrollment. Using the Young Lives data from India, Singh (2015) find that girls are less likely to be enrolled in private school. However, using data from five states in North India, Pal (2010) finds the reverse pattern. Using an experiment in Pakistan, Alderman, Orazem, and Paterno (2001) find that private school entry can actually help close the gender gaps in enrollment. Andrabi, Das, and Khwaja (2008) reach a similar conclusion using a rich set of panel data from Pakistani villages. Because distance to schools is a major barrier to enrollment, especially for girls, both sets of authors argue that policies that induce the expansion of private schools into underserved areas may be effective at closing gender gaps in enrollment. Such expansions could also close enrollment gaps by caste. However, the challenge is to design policies that sufficiently entice private schools to locate in underserved and disadvantaged areas, rather than to cluster around other private schools or relatively richer households (Andrabi, Das, Khwaja, Vishwanath, and Zajonc 2008).

Can Private Schools Deliver Better Outcomes?

Across various settings, there is growing evidence that private schools are finding ways of using their resources more effectively. In India and Pakistan, the operating costs for private schools are one-half to one-fourth that of government schools (Muralidharan and Sundararaman 2015 in India; Andrabi, Das, Khwaja, Vishwanath, and Zajonc 2008 in Pakistan). Most of the cost savings comes from differences in teacher hiring and remuneration. Private school teachers are younger, less educated, less likely to be formally trained, less experienced, and paid roughly one-third to one-fifth of their public school counterparts (based on World Bank Service Delivery Indicators data for 2012; Muralidharan and Sundararaman 2015; Andrabi, Das and Khwaja 2008). However, private school teachers display better attendance and effort, as measured by the proportion of time teachers are actually in class (World Bank Service Delivery Indicators for 2012). In addition, evidence from Pakistan suggests that teacher pay is negatively correlated with absence rates in the private sector, but positively correlated in the public sector, where older, more experienced higher paid teachers are more likely to be absent. The high rate of teacher turnover in Pakistani private schools, at over 25 percent per year, may be one mechanism that private schools employ to hold teachers accountable (Andrabi, Das, Khwaja, Vishwanath, and Zajonc 2008).

Most low-cost private schools are owned by sole proprietors, especially in Ghana and Nigeria (Tooley and Dixon 2005). These schools were often unable to expand to take advantage of any potential economies of scale. Because private schools tend to locate in clusters, they are often quite competitive, which drives down their profits. Using the data from Pakistan, Andrabi, Das, Khwaja, Vishwanath, and Zajonc (2008) show that the average profits of private schools are low, on par with the salary of teacher in a private school, which is the likely outside option of the school owner.

There has been a recent emergence of chains of for-profit low-cost private schools which are leveraging technology to deliver lessons and to manage teachers more effectively. Examples include Bridge International Academies in Kenya and the Omega Schools in Ghana (owned in part by James Tooley, author of numerous studies on low-cost schools). Bridge International Academies opened its first school in a Nairobi slum in January 2009. By November 2014, it had opened nearly 400 schools across Kenya and had enrolled over 100,000 students (see http://www.bridgeinternationalacademies.com/company/history). Bridge has now expanded into Nigeria and Uganda and is preparing to launch in India and Liberia.[3] Bridge employs curriculum development specialists who create scripted lessons. Each teacher is given a tablet and delivers extremely detailed scripted content to the classroom: for example, the scripts even include prompts to call on students. Bridge hires individuals who are not necessarily trained as teachers and pays them less than teachers in government schools. However, the tablets enable Bridge to monitor both teacher attendance and what material has been delivered in the classroom. Bridge also uses a database to track student learning outcomes. Teacher absence is less than 2 percent compared to over 16 percent in government schools, and teachers also spend more time in class (for more details about Bridge schools, including common critiques about their model see Rosenberg 2013, 2016).[4]

Simple comparisons of survey data across several contexts suggests that learning outcomes are generally higher in private schools (as shown by the World Bank Service Delivery Indicators for 2012; the Young Lives dataset at http://www.younglives.org.uk; Andrabi, Das, Khwaja, Vishwanath, and Zajonc 2008 in Pakistan). With respect to Bridge schools, an internal Bridge study focusing on grades 1, 2, and 3 found that students in a Bridge schools saw greater increases in learning relative to students in government schools (Bridge International Academies 2015). At the upper primary level, Bridge students did better than students in public schools in the Keynan national primary school exit exam. Bridge students scored between 0.2 to 0.3 standard deviations more than their government counterparts (author's calculations using Kenyan examinations data). However, it is likely that a substantial portion of the learning differences are driven by selection, given the differences

[3] The Liberian government has invited a number of private operators including Bridge to manage and operate around 100 public schools. These schools will be free to the families of the students, and the government will pay the operators a fixed fee per student. More details are reported by Rosenberg (2016).

[4] Bridge has attracted investors such as Mark Zuckerberg, Bill Gates, the Omidyar Network, the International Finance Corporation (part of the World Bank Group) and the UK Department of International Development (DFID). The full list of investors can be found at http://www.bridgeinternationalacademies.com/company/investors/. There are concerns in some circles about international development agencies financing or subsidizing a for-profit entity (Rosenberg 2013, 2016; Das 2016).

in observable characteristics such as parental education across school types, and in particular the probable differences in unobservable factors such as parental motivation or child ability.

Rigorous evaluations of private schools in Pakistan, India, and Colombia show that private schools deliver outcomes that are at least as good as public schools. Using student-level panel data and value-added approaches in Pakistan, Andrabi, Das, Khwaja, and Zajonc (2011) show that private schools raise learning between 0.19 to 0.3 standard deviations across English, math, and Urdu. Using a similar approach in India, Singh (2015) finds a large effect of private schools on English (over 0.6 standard deviations), but limited effects on math and Telegu (the local language) for younger students, and modest effects in both subjects for older students. Muralidharan and Sundararaman (2015) examine a program that randomly allocated vouchers to private schools among a pool of applicants from 180 villages in Andra Pradesh. Four years after the launch of the program, they find no impacts of private schools on math or Telegu, but do find significant impacts on English (0.12 standard deviations) and Hindi (0.55 standard deviations). Angrist, Bettinger, Bloom, King, and Kremer (2002) examine a low- to medium-cost private school voucher lottery that targeted low-income students in Colombia. Focusing on a sample of applicants from Bogota, they find a moderate effect of the program on test scores (0.2 standard deviations). Given that private schools generally operate with far fewer resources compared to public schools, these results suggest that private schools are much more productive, because they can deliver learning outcomes that are comparable or better than public schools at a much lower cost. Muralidharan and Sundararaman (2015) also show that private schools devote less time to certain subjects such as math, yet deliver outcomes that are at least as good as public schools in those subjects. This finding provides additional evidence of the relative productivity of private schools.

Policies to Leverage the Private Sector

There are a variety of policy options that could potentially leverage the productivity of the private sector. Some possibilities include using a voucher scheme in which students could choose their own low-cost private school; public-private partnerships in which the government uses private schools to expand enrollment; and encouraging competition between public and private schools.

Voucher programs are often touted as a mechanism to improve the productivity of the entire education system by promoting competition among schools. By allowing parents to vote with their feet, vouchers could promote accountability throughout the education system. However, there are concerns that such programs would lead to increased sorting and cause harm to public schools. Hsieh and Urquiola (2006) show that Chile's voucher program increased socioeconomic stratification, but had limited impact on learning outcomes. However, Muralidharan and Sundararaman (2015) find there were no negative spillovers of the voucher program on public school students in India. They also find suggestive evidence that the vouchers were more effective in markets with greater school competition.

Information constraints could also limit the effectiveness of vouchers or other school choice mechanisms. Andrabi, Das, and Khwaja (2015) show that providing information about the market for schooling, through village report cards, can increase both attendance and learning outcomes. Using a randomized experiment in 112 Pakistani villages that had a combination of public and private schools, they show that the provision of both school-level and student-level report cards in treatment villages increased the competitive pressures on both types of school to perform.

Because school choice is only feasible if there are a sufficient number of schools, policies that encourage the expansion of the supply of private schools could be cost-effective options to provide quality schooling to underserved locations or populations. In an early study along these lines, Kim, Alderman, and Orazem (1999) look at a program to stimulate girls' schooling by subsidizing the creation of private schools in poor urban neighborhoods of a city in Pakistan. Not only did enrollments rise for girls, but for boys, too. More recently, Barrera-Osorio, Blakeslee, Hoover, Linder, Raju, and Ryan (2013) examine an experiment in a sample of 199 villages in underserved rural districts in Pakistan where the government funded low-cost private schools. They show that the program both increased enrollment and led to a dramatic rise in test scores (compared with control villages with limited schooling options). Barrera-Osorio, de Galbert, Habyarimana, and Sarbarwal (2015) also find positive enrollment and test score effects when they examine a government program that subsidizes students to attend low-cost private schools in Uganda. The program was implemented with a randomized phase-in, thus allowing an experimental evaluation. Given the thin profit-margins generated by private schools, designing these subsidy programs to ensure the sustainability and survival of private schools that are induced to open in new locations is very challenging. Alderman, Kim, and Orazem (2003) and Andrabi, Das, Khwaja, Vishwanath, and Zajonc (2008) argue that these programs may not be well-suited to serve rural areas, which are typically less dense, poorer, and harder to staff.

Credit constraints are generally binding in the private school education sector, given the small scale of most private school operators. Such constraints could limit school investment, hindering the potential benefits of school choice. Andrabi, Das, Khwaja, and Singh (2015) examine a randomized experiment that provided unconditional grants to low-cost private schools. If only one school (or a few schools) in the market receives a grant, they find that the school is more likely to invest in expanding access rather than quality; however, when all schools in a market are provided finances, schools are more likely to compete on quality. They also show that labor constraints can make it difficult for private schools to enter or expand (Andrabi, Das, and Khwaja 2013). Private schools in Pakistan rely on female high school graduates to serve as teachers. They show that areas of Pakistan which had higher rates of female secondary school enrollment, due to the presence of a public girls' secondary school, are now seeing higher growth rates of private schools. Thus, an expansion of schooling also creates a larger labor pool of future teachers and benefits future schooling.

Despite the growing evidence on the effects of (low-cost) private schools, teachers' unions in developing countries have been very vocal in opposing these schools. They argue that private schools exploit parents by providing low quality education, due to their use of unqualified teachers. For example, the teachers' union in Kenya is demanding that Bridge schools be shut down (as reported in Wanzala 2016). To the extent that teachers' preferences are at odds with parental preferences, the growing political clout of teachers' unions in many developing countries may tilt education reforms towards policies that favor teachers. However, Davies (2015) suggests that parental support for (low-cost) private schools may increase as they gain greater familiarity with these schools. The greater exposure of parents (and their children) to private schools could be a necessary condition for parents to lobby for school choice, or other policies that generally support private schools.

Conclusion

The education system is of central importance to the economic future of developing countries, both because of the important role of education in economic growth and because of the limited ability of parents in many countries—given their own limited education levels—to provide home inputs to education. Developing countries as a group have made substantial steps in raising enrollment and committing more resources to education. Subsequent reforms need to focus on initiatives that increase accountability and incentives across the education system, improve the effort and pedagogical practice of teachers, support the more efficient use of the existing resources, and leverage the growing private sector.

Recent research, including a number of randomized control trials, has shed light on possible interventions and policies that could be employed to address the accountability and incentive problems facing schools in developing countries. Much of this research so far has focused on using teachers to deliver primary education. Future research seems likely to move toward using technology to deliver content, as well as to monitor teachers, students, and funding. In particular, finding ways for technology to allow instruction to be tailored to the student's level could dramatically improve the productivity of the education system. Also, as many countries have adopted free primary education, future research seems likely to turn to secondary school and other post-primary education options. Finally, there is limited research on early childhood education in developing countries, especially in African contexts.

Translating the emerging research findings into actual changes in public policy always faces problems of implementation and political economy. Small-scale experiments run by credible non-government organizations may not scale up so well if financed and administered at large scale by governments. Additionally, it is a practical challenge to find ways to focus the attention of parents and voters on effective policies that address learning, rather than visible inputs, and then seek to build coalitions for promoting effective educational reforms in developing countries. There is optimism that increased adoption of results-based financing schemes can help shift the focus of entire education systems towards learning. The World Bank announced in May 2015 it would double the amount devoted to results-based financing in education to over US$5 billion over the next five years (see http://www.worldbank.org/en/news/press-release/2015/05/18/world-bank-group-doubles-results-based-financing-for-education-to-us5-billion-over-next-5-years) By paying for (pre-agreed) results, the hope is that these schemes can help increase accountability from the Ministry of Finance to the Ministry of Education all the way down schools, teachers, students, and parents. As failure to meet a specified target will be extremely visible, results-based financing could potentially change the political salience of learning outcomes. However, the effectiveness of such schemes will depend critically on their design and implementation.

■ *I thank Angela Ambroz, Allison Bigelow, Andrew Brudevold, Jishnu Das, Paul Glewwe, Adrienne Lucas, Anthony Keats, Michael Kremer, Shawn Powers, Youdi Schipper, Jay Shimshack, Tavneet Suri, Caitlyn Tulloch, Craig Volden, Andrew Zeitlin for their comments and suggestions. Taylor Salisbury, Ramsey Stewart, and Morgan Falls provided super research assistance. I would also like to thank the JEP Editors for their support, encouragement, and feedback.*

References

Alderman, Harold, Peter F. Orazem, and Elizabeth M. Paterno. 2001. "School Quality, School Cost, and the Public/Private School Choices of Low-Income Households in Pakistan." *Journal of Human Resources* 36(2): 304–26.

Alderman, Harold, Jooseop Kim, and Peter F. Orazem. 2003. "Design, Evaluation, and Sustainability of Private Schools for the Poor: The Pakistan Urban and Rural Fellowship School Experiments." *Economics of Education Review* 22(3): 265–74.

Andrabi, Tahir, Jishnu Das, and Asim Ijaz Khwaja. 2008. "A Dime a Day: The Possibilities and Limits of Private Schooling in Pakistan." *Comparative Education Review* 52(3): 329–55.

Andrabi, Tahir, Jishnu Das, and Asim I. Khwaja. 2013. "Students Today, Teachers Tomorrow? Identifying Constraints on the Provision of Education." *Journal of Public Economics*, 100: 1-14.

Andrabi, Tahir, Jishnu Das, and Asim Ijaz Khwaja. 2015. "Report cards: the impact of providing school and child test scores on educational markets." World Bank Policy Research Working Paper 7226.

Andrabi, Tahir, Jishnu Das, Asim Ijaz Khwaja, and Niharika Singh. 2015. "Upping the Ante: The Equilibrium Effects of Unconditional Grants to Private Schools." Unpublished paper.

Andrabi, Tahir, Jishnu Das, Asim Ijaz Khwaja, and Tristan Zajonc. 2011. "Do Value-Added Estimates Add Value? Accounting for Learning Dynamics." *American Economic Journal: Applied Economics* 3(3): 29-54.

Andrabi, Tahir, Jishnu Das, Asim Ijaz Khwaja, Tara Vishwanath, and Tristan Zajonc. 2008. "PAKISTAN: Learning and Educational Achievements in Punjab Schools (LEAPS): Insights to Inform the Education Policy Debate." World Bank Report.

Angrist, Joshua, Eric Bettinger, Erik Bloom, Elizabeth King, and Michael Kremer. 2002. "Vouchers for Private Schooling in Colombia: Evidence from a Randomized Natural Experiment." *American Economic Review* 92(5): 1535–58.

Banerjee, Abhijit V., Rukmini Banerji, Esther Duflo, Rachel Glennerster, and Stuti Khemani. 2010. "Pitfalls of Participatory Programs: Evidence from a Randomized Evaluation in Education in India." *American Economic Journal: Economic Policy*

2(1): 1–30.

Banerjee, Abhijit V., Shawn Cole, Esther Duflo, and Leigh Linden. 2007. "Remedying Education: Evidence from Two Randomized Experiments in India." *Quarterly Journal of Economics* 122(3): 1235–64.

Banerjee, Abhijit V., and Esther Duflo. 2010. *Poor Economics: A Radical Rethinking of the Way to Fight Global Poverty*. Public Affairs.

Banerji, Rukmini. 2014. "From Invisible to Visible: Being Able To 'See' the Crisis in Learning." Brookings Institution Education Plus Development blog." May 22. http://www.brookings.edu/blogs/education-plus-development/posts/2014/05/22-india-learning-banerji.

Banerji, Rukmini, James Berry, and Marc Shotland. 2013. "The Impact of Mother Literacy and Participation Programs on Child Learning: Evidence from a Randomized Evaluation in India." J-PAL Working Paper.

Barlevy, Gadi, and Derek Neal. 2012. "Pay for Percentile." *American Economic Review* 102(5): 1805–31.

Barr, Abigail, Frederick Mugisha, Pieter Serneels, and Andrew Zeitlin. 2012. "Information and Collective Action in Community-based Monitoring of Schools: Field and Lab Experimental Evidence from Uganda." Unpublished paper, Georgetown University.

Barrera-Osorio, Felipe, David S. Blakeslee, Matthew Hoover, Leigh L. Linden, Dhushyanth Raju, and Stephen Ryan. 2013. "Leveraging the Private Sector to Improve Primary School Enrolment: Evidence from a Randomized Controlled Trial in Pakistan." Unpublished paper, Harvard Graduate School of Education.

Barrera-Osorio, Felipe, Pierre de Galbert, James Habyarimana, and Shwetlena Sabarwal. 2015. "The Impact of Public-Private Partnerships on Private School Performance: Evidence from a Randomized Controlled Trial in Uganda." Unpublished paper, Harvard Graduate School of Education.

Behrman, Jere H., Susan W. Parker, Petra E. Todd, and Kenneth I. Wolpin. 2012. "Aligning Learning Incentives of Students and Teachers: Results from a Social Experiment in Mexican High Schools." PIER Working Papers 13-004, Penn Institute for Economic Research.

Bietenbeck, Jan, Marc Piopiunik, and Simon Wiederhold. 2015. "Africa's Skill Tragedy: Does Teachers' Lack of Knowledge Lead to Low Student Performance?" CESifo Working Paper Series 5470.

Blimpo, Moussa Pouguinimpo, David K. Evans, and Nathalie Lahire. 2015. "Parental Human Capital and Effective School Management: Evidence from the Gambia." World Bank Policy Research Working Paper 7238.

Bold, Tessa, Mwangi Kimenyi, Germano Mwabu, Alice Ng'ang'a, and Justin Sandefur. 2013. "Scaling Up What Works: Experimental Evidence on External Validity in Kenyan Education." Center for Global Development Working Paper 321.

Bridge International Academies. 2015. "The Bridge Effect: A Comparison of Early Grade Learning Gains in English and Maths." Bridge International Working Paper. http://www.bridgeinternationalacademies.com/wp-content/uploads/2016/07/The-Bridge-Effect_Working-Paper-Draft-V4_Website2.pdf.

Carneiro, Pedro Manuel, Oswald Koussihouede, Nathalie Lahire, Costas Meghir, and Corina Mommaerts. 2015. "Decentralizing Education Resources: School Grants in Senegal." CEPR Discussion Paper DP10527.

Conn, Katherine. 2014. "Identifying Effective Education Interventions in Sub-Saharan Africa: A meta-analysis of rigorous impact evaluations." Unpublished paper, Columbia University.

Cristia, Julián P., Pablo Ibarrarán, Santiago Cueto, Ana Santiago, and Eugenio Severín. 2012. "Technology and Child Development: Evidence from the One Laptop Per Child Program IZA Discussion Paper 6401, March.

Das, Jishnu. 2016. "Foreign aid should support private schooling, not private schools." Future Development Blog. Brookings Institution. June. http://www.brookings.edu/blogs/future-development/posts/2016/06/29-foreign-aid-private-schooling-education-das.

Das, Jishnu, Stefan Dercon, James Habyarimana, Pramila Krishnan, Karthik Muralidharan, and Venkatesh Sundararaman. 2013. "School Inputs, Household Substitution, and Test Scores." *American Economic Journal: Applied Economics* 5(2): 29–57.

Davies, Emmerich. 2015. "The Lessons Private Schools Teach: Using a Downstream Experiment to Understand the Effects of Private Schools on Political Behavior." Unpublished paper.

De Ree, Joppe, Karthik Muralidharan, Menno Pradhan, and Halsey Rogers. 2015. "Double for Nothing? Experimental Evidence on the Impact of an Unconditional Teacher Salary Increase on Student Performance in Indonesia." NBER Working Paper 21806.

Duflo, Esther, Pascaline Dupas, and Michael Kremer. 2011. "Peer Effects, Teacher Incentives, and the Impact of Tracking: Evidence from a Randomized Evaluation in Kenya." *American Economic Review* 101(5): 1739–74

Duflo, Esther, Pascaline Dupas, and Michael Kremer. 2012. "School Governance, Teacher Incentives, and Pupil-Teacher Ratios: Experimental

Evidence from Kenyan Primary Schools." *Journal of Public Economics* 123: 92–110.

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

Evans, David K., and Anna Popova. 2015. "What Really Works to Improve Learning in Developing Countries? An Analysis of Divergent Findings in Systematic Reviews." World Bank Policy Research Working Paper 7203.

Glewwe, Paul. W., Eric A. Hanushek, Sarah D. Humpage, and Renato Ravina. 2014. "School resources and educational outcomes in developing countries: a review of the literature from 1990 to 2010." In Glewwe, Paul, ed., *Education Policy in Developing Countries*. University of Chicago Press.

Glewwe, Paul, Michael Kremer, and Sylvie Moulin. 2009. "Many Children Left Behind? Textbooks and Test Scores in Kenya." *American Economic Journal: Applied Economics* 1(1): 112–35.

Glewwe, Paul, Michael Kremer, Sylvie Moulin, and Eric Zitzewitz. 2004. "Retrospective vs. Prospective Analyses of School Inputs: The Case of Flip Charts in Kenya." *Journal of Development Economics* 74(1): 251–68.

Glewwe, Paul, and Karthik Muralidharan. 2015. "Improving School Education Outcomes in Developing Countries: Evidence, Knowledge Gaps, and Policy Implications." Research on Improving Systems of Education Working Paper 15/001. October 25.

Harding, Robin, and David Stasavage. 2014. "What Democracy Does (and Doesn't Do) for Basic Services: School Fees, School Inputs, and African Elections." *Journal of Politics* 76(1): 229–45.

Heyneman, Stephen P., and Jonathan M. B. Stern. 2013. "Low Cost Private Schools for the Poor: What Public Policy is Appropriate?" *International Journal of Educational Development* 35: 3–15.

Hsieh, Chang-Tai, and Miguel Urquiola. 2006. "The Effects of Generalized School Choice on Achievement and Stratification: Evidence from Chile's Voucher Program." *Journal of Public Economics* 90(8–9): 1477–1503.

Kim, Jooseop, Harold Alderman, and Peter F. Orazem. 1999. "Can Private School Subsidies Increase Enrollment for the Poor? The Quetta Urban Fellowship Program." *World Bank Economic Review* 13(3): 443–65.

Kramon, Eric, and Daniel N. Posner. 2016. "Ethnic Favoritism in Education in Kenya." *Quarterly Journal of Political Science* 11: 1–58.

Kremer, Michael, Connor Brannen, and Rachel Glennerster. 2013. "The challenge of education and learning in the developing world." *Science* 340(6130), 297–300.

Kremer, Michael, Sylvie Moulin, and Robert Namuyu. 2003. "Decentralization: A Cautionary Tale." Unpublished Paper.

Lieberman, Evan S., Daniel N. Posner, and Lily L. Tsai. 2014. "Does Information Lead to More Active Citizenship? Evidence from an Education Intervention in Rural Kenya." *World Development* 60, 69–83.

Loyalka, Prashant, Sean Sylvia, Chengfang Liu, James Chu, and Yaojiang Shi. 2016. "Pay by Design: Teacher Performance Pay Design and the Distribution of Student Achievement." http://ssrn.com/abstract=2775461.

Lucas, Adrienne M., and Isaac M. Mbiti. 2012. "Access, Sorting, and Achievement: The Short-Run Effects of Free Primary Education in Kenya." *American Economic Journal: Applied Economics* 4(4): 226–53.

Lucas, Adrienne M., Patrick McEwan, Moses Ngware, and Moses Oketch. 2014. "Improving Early-Grade Literacy in East Africa: Experimental Evidence from Kenya and Uganda." *Journal of Policy Analysis and Management* 33(4): 950–76.

Marx, Benjamin, Thomas Stoker, and Tavneet Suri. 2013. "The Economics of Slums in the Developing World." *Journal of Economic Perspectives* 27(4): 187–210.

Mbiti, Isaac, Karthik Muralidharan, Mauricio Romero, Youdi Schipper, Rakesh Rajani, and Constantine Manda. 2016. "Inputs, Incentive and Complementarities in Primary Education: Experimental Evidence from Tanzania." Unpublished Paper.

McEwan, Patrick J. 2015. "Improving learning in primary schools of developing countries: A metaanalysis of randomized experiments." *Review of Educational Research* 85: 353-394.

Metzler, Johannes, and Ludger Woessmann. 2012. "The Impact of Teacher Subject Knowledge on Student Achievement: Evidence from Within-Teacher Within-Student Variation." *Journal of Development Economics* 99(2): 486–96.

Miguel, Edward. 2004. "Tribe or Nation? Nation-Building and Public Goods in Kenya versus Tanzania." *World Politics* 56(3): 327–62.

Miguel, Edward, and Mary Kay Gugerty. 2005. "Ethnic Diversity, Social Sanctions, and Public Goods in Kenya." *Journal of Public Economics* 89(11–12): 2325–68.

Muralidharan, Karthik, Jishnu Das, Alaka Holla, and Aakash Mohpal. 2016. "The Fiscal Cost of Weak Governance: Evidence from Teacher Absence in India." World Bank Policy Research Working Paper 7579. February.

Muralidharan, Karthik, and Venkatesh Sundararaman. 2015. "The Aggregate Effect of School Choice: Evidence from A Two-Stage Experiment in India." *Quarterly Journal of Economics* 130(3): 1011–66.

Muralidharan, Karthik, and Venkatesh Sundararaman. 2011. "Teacher Performance Pay: Experimental Evidence from India." *Journal of Political Economy* 119(1): 39–77.

Muralidharan, Karthik, and Venkatesh Sundararaman. 2013. "Contract Teachers: Experimental Evidence from India." NBER Working Paper 19440.

Murnane, Richard J., and Alejandro J. Ganimian. 2014. "Improving Educational Outcomes in Developing Countries: Lessons from Rigorous Evaluations." NBER Working Paper 20284.

Ngware, Moses W., James Ciera, Peter K. Musyoka and Moses Oketch. 2015. "Quality of teaching mathematics and learning achievement gains: evidence from primary schools in Kenya." *Educational Studies in Mathematics* 89(1): 111–131.

Oketch, Moses, Maurice Mutisya, Moses Ngware, Alex C. Ezeh. 2010. "Why Are There Proportionately More Poor Pupils Enrolled in Non-state Schools in Urban Kenya in Spite of FPE Policy?" *International Journal of Educational Development* 30(1): 23–32.

Pal, Sarmistha. 2010. "Public Infrastructure, Location of Private Schools and Primary Attainment in an Emerging Economy." *Economics of Education Review* 29(5): 783–94.

Piper, Benjamin, Stephanie Simmons Zuilkowski, and Abel Mugenda. 2014. "Improving Reading Outcomes in Kenya: First-Year Effects of the PRIMR Initiative." *International Journal of Educational Development,* 37, November.

Pradhan, Menno, Daniel Suryadarma, Amanda Beatty, Maisy Wong, Arya Gaduh, Armida Alisjahbana, and Rima Prama Artha. 2014. "Improving Educational Quality through Enhancing Community Participation: Results from a Randomized Field Experiment in Indonesia." *American Economic Journal: Applied Economics* 6(2): 105–26.

Pratham. 2013. *Annual Status of Education Report (ASER).* New Delhi, India: Pratham.

Pratham. 2014. *Annual Status of Education Report (Rural)* 2013. New Delhi, India: Pratham.

Pugatch, Todd, and Elizabeth Schroeder. 2014a. Incentives for Teacher Relocation: Evidence from the Gambian Hardship Allowance." *Economics of Education Review,* 41: 120–136.

Pugatch, Todd and Elizabeth Schroeder. 2014b. "Teacher Pay and Student Performance: Evidence from the Gambian Hardship Allowance." IZA Discussion Paper 8621.

Reinikka, Ritva, and Jakob Svensson. 2004. "Local Capture: Evidence From A Central Government Transfer Program In Uganda." *Quarterly Journal of Economics* 119(2): 679-705.

Reinikka, Ritva, and Jakob Svensson. 2005. "Fighting Corruption to Improve Schooling: Evidence From a Newspaper Campaign in Uganda." *Journal of the European Economic Association* 3: 259–267.

Rosenberg, Tina. 2013. "A By-the-E-Book Education, for $5 a Month." *New York Times* May 22, 2013. http://opinionator.blogs.nytimes.com/2013/05/22/a-by-the-e-book-education-for-5-a-month/?_r=0.

Rosenberg, Tina. 2016. "Liberia, Desperate to Educate, Turns to Charter Schools." *New York Times.* June 14. http://www.nytimes.com/2016/06/14/opinion/liberia-desperate-to-educate-turns-to-charter-schools.html.

Sabarwal, Shwetlena, David Evans, and Anastasia Marshak. 2014. "The Permanent Input Hypothesis: School Inputs and (No) Student Learning in Sierra Leone." World Bank Policy Research Working Paper 7021. September.

Singh, Abhijeet. 2015. "Private School Effects in Urban and Rural India: Panel Estimates at Primary and Secondary Ages." *Journal of Development Economics* 113: 16–32.

Stasavage, David. 2005. "Democracy and Education Spending in Africa." *American Journal of Political Science* 49, 343–358.

Tooley, James, and Pauline Dixon. 2005. "Private Education is Good for the Poor: A Study of Private Schools Serving the Poor in Low-Income Countries." Cato Institute Report.

Tooley, James, and Pauline Dixon. 2007. "Private Schooling for Low-Income Families: A Census and Comparative Survey in East Delhi, India." *International Journal of Educational Development* 27(2): 205–19.

Twaweza. 2013. "KiuFunza: The Thirst to Learn Baseline Database." Twaweza, Dar Es Salaam. http://twaweza.org/go/kiufunza-launch1. Public access to data is forthcoming.

UNESCO. 2011. *Financing Education in Sub-Saharan Africa Meeting the Challenges of Expansion, Equity and Quality.* UNESCO Institute of Statistics, Montreal.

Uwezo. 2013. "Are Our Children Learning? Literacy and Numeracy across East Africa." Uwezo, Nairobi, Kenya.

Wanzala, Ouma. 2016. "Teachers' Unions Want Low-Cost Slum Schools Closed." *Daily Nation* (Kenya), January 26. http://www.nation.co.ke/news/Teachers-unions-want-low-cost-slum-schools-closed/-/1056/3049538/-/edyuix/-/index.html.

World Bank. 2003. *World Development Report 2004: Making Services Work for Poor People.* Washington, DC: World Bank.

Journal of Economic Perspectives—Volume 30, Number 3—Summer 2016—Pages 133–140

The Mechanics of Motivated Reasoning

Nicholas Epley and Thomas Gilovich

Whenever we see voters explain away their preferred candidate's weaknesses, dieters assert that a couple scoops of ice cream won't *really* hurt their weight loss goals, or parents maintain that their children are unusually gifted, we are reminded that people's preferences can affect their beliefs. This idea is captured in the common saying, "People believe what they want to believe."

But people don't *simply* believe what they want to believe. The psychological mechanisms that produce motivated beliefs are much more complicated than that. Personally, we'd like to believe that our contributions to the psychological literature might someday rival those of Daniel Kahneman, but, try as we might, the disparity in citations, prizes, invitations—you name it—makes holding such a belief impossible. People generally *reason* their way to conclusions they favor, with their preferences influencing the way evidence is gathered, arguments are processed, and memories of past experience are recalled. Each of these processes can be affected in subtle ways by people's motivations, leading to biased beliefs that feel objective (Gilovich and Ross 2015; Pronin, Gilovich, and Ross 2004). As Kunda (1990) put it, "people motivated to arrive at a particular conclusion attempt to be rational and to construct a justification of their desired conclusion that would persuade a dispassionate observer. They draw the desired conclusion only if they can muster up the evidence necessary to support it" (p. 482–83). Motivated reasoning is constrained.

■ *Nicholas Epley is the John T. Keller Professor of Behavioral Science, University of Chicago, Booth School of Business, Chicago, Illinois. Thomas Gilovich is the Irene Blecker Rosenfeld Professor of Psychology, Cornell University, Ithaca, New York. Their email addresses are epley@chicagobooth.edu and tdg1@cornell.edu.*

† For supplementary materials such as appendices, datasets, and author disclosure statements, see the article page at
http://dx.doi.org/10.1257/jep.30.3.133 doi=10.1257/jep.30.3.133

Psychological research makes it clear, in other words, that "motivated beliefs" are guided by motivated reasoning—reasoning in the service of some self-interest, to be sure, but reasoning nonetheless. We hope that being explicit about what psychologists have learned about motivated reasoning will help clarify the types of motivated beliefs that people are most likely to hold, specify when such beliefs are likely to be strong and when they are likely to be relatively weak or fragile, and illuminate when they are likely to guide people's behavior.

In this introduction, we set the stage for the discussion of motivated beliefs in the papers that follow by providing more detail about the underlying psychological processes that guide motivated reasoning, including a discussion of the varied motives that drive motivated reasoning and a description of how goals can direct motivated reasoning to produce systematically biased beliefs. The first paper in this symposium, by Roland Bénabou and Jean Tirole, presents a theoretical framework for how motives might influence behavior in several important domains; two additional papers focus on specific motives that can guide motivated reasoning: Russell Golman, George Loewenstein, Karl Ove Moene, and Luca Zarri discuss how a "preference for belief consonance" leads people to try to reduce the gap between their beliefs and those of relevant others, and Francesca Gino, Michael Norton, and Roberto Weber consider how people engage in motivated reasoning to feel as if they are acting morally, even while acting egoistically.

A more detailed understanding of motivated beliefs and motivated reasoning yields a middle-ground view of the quality of human judgment and decision-making. It is now abundantly clear that people are not as smart and sophisticated as rational agent models assert (Kahneman and Tversky 2000; Thaler 1991; Simon 1956), in the sense that people do not process information in unbiased ways. But people are also not as simple-minded, naïve, and prone to simply ignoring unpalatable information as a shallow understanding (or reporting) of motivated beliefs might suggest.

Motives for Reasoning

People reason to prepare for action, and so reasoning is motivated by the goals people are trying to achieve. A coach trying to win a game thinks about an opponent's likely moves more intensely than a cheerleader trying to energize the crowd. A lawyer trying to defend a client looks for evidence of innocence, whereas a lawyer seeking to convict tries to construct a chain of reasoning that will lead to a guilty verdict. A person feeling guilty about harming another focuses on ways to assuage the guilt, while the person harmed is likely to focus on the nature and extent of the harm. As the great psychologist and philosopher William James (1890, p. 333) wrote more than a century ago: "My thinking, is first and last and always for the sake of my doing, and I can only do one thing at a time."

One of the complexities in understanding motivated reasoning is that people have many goals, ranging from the fundamental imperatives of survival and reproduction to the more proximate goals that help us survive and reproduce, such as achieving social status, maintaining cooperative social relationships, holding accurate beliefs and expectations, and having consistent beliefs that enable effective action. Sometimes reasoning directed at one goal undermines another. A person trying to persuade others about a particular point is likely to focus on reasons why his arguments are valid and decisive—an attentional focus that could make the person more compelling in the eyes of others but also undermine the accuracy of his assessments (Anderson, Brion, Moore, and Kennedy 2012). A person who recognizes that a set of beliefs is strongly held by a group of peers is likely to seek out and welcome information supporting those beliefs, while maintaining a much higher level of skepticism about contradictory information (as Golman, Loewenstein, Moene, and Zarri discuss in this symposium). A company manager narrowly focused on the bottom line may find ways to rationalize or disregard the ethical implications of actions that advance short-term profitability (as Gino, Norton, and Weber discuss in this symposium).

The crucial point is that the process of gathering and processing information can systematically depart from accepted rational standards because one goal—desire to persuade, agreement with a peer group, self-image, self-preservation—can commandeer attention and guide reasoning at the expense of accuracy. Economists are well aware of crowding-out effects in markets. For psychologists, motivated reasoning represents an example of crowding-out in attention.

In any given instance, it can be a challenge to figure out which goals are guiding reasoning. Consider the often-cited examples of "above-average" effects in self-evaluation: on almost any desirable human trait, from kindness to trustworthiness to the ability to get along with others, the average person consistently rates him- or herself above average (Alicke and Govorun 2005; Dunning, Meyerowitz, and Holzberg 1989; Klar and Giladi 1997). An obvious explanation for this result is that people's reasoning is guided by egoism, or the goal to think well of oneself. Indeed, a certain percentage of above-average effects can be explained by egoism because unrelated threats to people's self-image tend to increase the tendency for people to think they are better than others, in an apparent effort to bolster their self-image (as in Beauregard and Dunning 1998).

But above-average effects also reflect people's sincere attempts to assess accurately their standing in the world. For instance, many traits are ambiguous and hard to define, such as leadership or creativity. When people try to understand where they stand relative to their peers on a given trait, people quite naturally focus on what they know best about that trait—and what they know best are the personal strengths that guide their own lives. As Thomas Schelling (1978, pp. 64–65) put it, "Careful drivers give weight to care, skillful drivers give weight to skill, and those who think that, whatever else they are not, at least they are polite, give weight to courtesy, and come out high on their own scale. This is the way that every child has the best dog on the block." The above-average effect, in other words, can result from a self-enhancement goal, or from a non-motivated tendency to define traits egocentrically. Supporting Schelling's analysis, the above-average effect is significantly reduced when traits are given precise definitions, or when the traits are inherently less ambiguous such as "punctual" or "tall" (Dunning, Meyerowitz, and Holzberg 1989).

Knowing which goal is guiding reasoning is critical for predicting the influence of specific interventions. For example, economists routinely predict that biases in judgment will be reduced when the stakes for accurate responding are high. This prediction implicitly assumes that people are not trying to be accurate already. But in fact, many cognitive biases are not affected by increased incentives for accuracy because the individuals in question are already trying hard to be accurate (Camerer and Hogarth 1999). Increasing the incentive to achieve a goal should influence behavior only when people are not already trying to achieve that goal.

How Motives Influence Beliefs

Understanding that multiple goals can shape reasoning does not explain *how* reasoning can become systematically biased. Reasoning involves the recruitment and evaluation of evidence. Goals can distort both of these basic cognitive processes.

Recruiting Evidence

When recruiting evidence to evaluate the validity of a given belief, an impartial judge would consider all of the available evidence. Most people do not reason like impartial judges, but instead recruit evidence like attorneys, looking for evidence that supports a desired belief while trying to steer clear of evidence that refutes it. In one memorable example, essayist Johanna Gohmann (2015) describes her improbable teenage crush on the actor Jimmy Stewart, and her reaction as she learned more and more about Mr. Stewart: "As I flipped through the pages my eyes skimmed words like 'womanizer' and 'FBI informant,' and I slapped it shut, reading no further." If you avoid recruiting evidence that you would prefer not to believe, your beliefs will be based on only a comforting slice of the available facts. One prominent example of motivated avoidance comes from studies of people's reactions to the prospect of having Huntington's disease: few people who are at risk of getting the disease get tested before showing symptoms, and those with symptoms who avoid testing have beliefs that are just as optimistic as those who show no symptoms (Oster, Shoulson, and Dorsey 2013).

Even when people do not actively avoid information, psychological research consistently demonstrates that they have an easier time recruiting evidence supporting what they *want* to be true than evidence supporting what they want to be false. But even here, people are still responsive to reality and don't simply believe whatever they want to believe. Instead, they recruit subsets of the relevant evidence that are biased in favor of what they want to believe. Failing to recognize the biased nature of their information search leaves people feeling that their belief is firmly supported by the relevant evidence.

Biased information processing can be understood as a general tendency for people to ask themselves very different questions when evaluating propositions they favor versus oppose (Gilovich 1991). When considering propositions they would prefer to be true, people tend to ask themselves something like "Can I believe this?" This evidentiary standard is rather easy to meet; after all, *some* evidence can usually be found even for highly dubious propositions. Some patients will get better after undergoing even a worthless treatment; someone is bound to conform to even the most baseless stereotype; some fact can be found to support even the wackiest conspiracy theory.

In contrast, when considering propositions they would prefer *not* be true, people tend to ask themselves something like "*Must* I believe this?" This evidentiary standard is harder to meet; after all, *some* contradictory evidence can be found for almost any proposition. Not all patients benefit from demonstrably effective treatments; not all group members conform to the stereotypes of their group; even the most comprehensive web of evidence will have a few holes. More compelling evidence is therefore required to pass this "Must I?" standard. In this way, people can again end up believing what they want to believe, not through mindless wishful thinking but rather through genuine reasoning processes that seem sound to the person doing it.

In one study that supports this Can I?/Must I? distinction, students were told that they would be tested for an enzyme deficiency that would lead to pancreatic disorders later in life, even among those (like presumably all of them) who were not currently experiencing any symptoms (Ditto and Lopez 1992). The test consisted of depositing a small amount of saliva in a cup and then putting a piece of litmus paper into the saliva. Half the participants were told they would know they had the enzyme deficiency if the paper changed color; the other half were told they would know they had it if the paper *did not* change color. The paper was such that it did not change color for anyone.

Participants in these two conditions reacted very differently to the same result—the unchanged litmus paper. Those who thought it reflected good news were quick to accept that verdict and did not keep the paper in the cup very long. Those who thought the unchanged color reflected bad news, in contrast, tried to recruit more evidence. They kept the paper in the cup significantly longer, even trying out (as the investigators put it) "a variety of different testing behaviors, such as placing the test strip directly on their tongue, multiple redipping of the original test strip (up to 12 times), as well as shaking, wiping, blowing on, and in general quite carefully scrutinizing the recalcitrant . . . test strip." A signal that participants wanted to receive was quickly accepted; a signal they did not want to receive was subjected to more extensive testing.

People's motivations thus do not directly influence what they believe. Instead, their motivations guide what information they consider, resulting in favorable conclusions that seem mandated by the available evidence.

Evaluating Evidence

Of course, even when looking at the very same evidence, people with different goals can interpret it differently and come to different conclusions. In one telling experiment cited in this symposium, participants who were randomly assigned to play the role of a prosecuting attorney judged the evidence presented in trial to be more consistent with the defendant's guilt than did participants randomly assigned to play the role of the defense attorney (Babcock and Loewenstein 1997).

These distorting influences can take many forms, influencing the apparent meaning of the evidence before us. For instance, any given action can be thought of in multiple ways. A father lifting a child off the floor could be described as "picking up a child" or "caring for the child." The two equally apt descriptions have very different meanings. Caring for a child is a more significant, benevolent act than simply picking up the child. A person trying to extol a parent's character will be more likely to code the event in a higher-level term like "caring" than a person trying to demean a parent's character. Differences in how people construe the very same action can lead two people to observe the same event but "see" very different things (Maas, Salvi, Arcuri, and Semin 1989; Trope and Lieberman 2003; Vallacher and Wegner 1987).

Psychologists have examined a host of ways in which people's goals influence how they evaluate information, and we won't review that voluminous literature here. But it is worth noting that psychologists have been especially interested in the distortions that arise in the service of consistency. Leon Festinger's (1957) theory of cognitive dissonance has been particularly influential. The central idea is that people are motivated to reconcile any inconsistencies between their actions, attitudes, beliefs, or values. When two beliefs are in conflict, or when an action contradicts a personal value, the individual experiences an unpleasant state of arousal that leads to psychological efforts to dampen or erase the discrepancy, often by changing a belief or attitude.

Festinger's (1957) theory stemmed in part from his earlier work on group dynamics and what he called "pressures to uniformity" (Festinger 1950). When differences of opinion arise within a group, a palpable tension arises that group members try to resolve. That tension, he maintained, is diminished only when agreement is achieved, typically by the majority pressuring the minority to go along. Festinger's theory of cognitive dissonance essentially took what he had observed in groups and put it in the head of the individual: that is, what plays out interpersonally in group dynamics also takes place in individual psychodynamics. We all feel psychological discomfort when our actions, attitudes, beliefs, or values conflict, and that discomfort leads us to seek ways to reduce the dissonance.

By focusing on cognitive processes that occur in the head of the individual, Festinger (1957) helped to usher in a period in which social psychology became a lot less social. But dissonance reduction is often a group effort. We help one another feel better about potentially upsetting inconsistencies in our thoughts and deeds. Our friends reassure us that we chose the right job, the right house, or the right spouse. We console an acquaintance who's messed up by saying that "it's not so bad," "he had it coming," or "things would have turned out the same regardless of what you did." Indeed, whole societies help their members justify the ill-treatment of minorities, the skewed division of resources, or the degradation of the environment through a variety of mechanisms, including everyday discourse, mass media messages, the criminal code, and even how the physical environment is structured.

The social element of rationalization and dissonance reduction fits nicely with the insightful piece by Golman, Loewenstein, Moene, and Zarri on people's preference for belief consonance. Furthermore, by connecting the preference for belief consonance to the existing literature on dissonance reduction, a great body of empirical research can be tapped to advance our understanding of when and why people will have an easy time achieving the belief consonance they seek, and when and why they are likely to struggle.

Coda

The most memorable line from the classic film *Gone with the Wind*—indeed, the most memorable line in the history of American movies according to the American Film Institute—is Rhett Butler's dismissive comment, "Frankly Scarlett, I don't give a damn." But a different line from that film has attracted more interest from psychologists: Scarlett O'Hara's frequent lament, "I can't think about that right now. . . . I'll think about it tomorrow."

The comment captures people's intuitive understanding of how motivations and emotions influence our judgments and decisions. When Scarlett doesn't want to accept some unwelcome possibility, she willfully cuts herself off from the relevant evidence. She can continue to believe what she wants because she never consults evidence that would lead her to believe differently.

Scarlet's path is one way that people can end up believing what they want to believe. But as we have noted, there are many others. Furthermore, people's preferred beliefs, developed and sustained through whatever path, guide their behavior whenever they are called to mind as choices are made. The path from motives to beliefs to choices should not be a black box to be filled with analytically convenient assumptions. Different motives can guide reasoning in different ways on different occasions—altering how information is recruited and evaluated—depending on what a person is preparing to do. We are delighted to see a topic with such a long history in psychological science being taken seriously by economists.

■ *Thanks to George Loewenstein, who took the leading role in stimulating and organizing the papers that appear in this symposium.*

References

Alicke, Mark D., and Olesya Govorun. 2005. "The Better-than-Average Effect." In *The Self in Social Judgment*, edited by Mark D. Alicke, David A. Dunning, and Joachim I. Krueger, 85–106. New York: Psychology Press.

Anderson, Cameron, Sebastien Brion, Don A. Moore, and Jessica A. Kennedy. 2012. "A Status-Enhancement Account of Overconfidence." *Journal of Personality and Social Psychology* 103(4): 718–35.

Babcock, Linda, and George Loewenstein. 1997. "Explaining Bargaining Impasse: The Role of Self-Serving Biases." *Journal of Economic Perspectives* 11(1): 109–26.

Beauregard, Keith S., and David Dunning. 1998. "Turning Up the Contrast: Self-Enhancement Motives Prompt Egocentric Contrast Effects in Social Judgments." *Journal of Personality and Social Psychology* 74(3): 606–621.

Camerer, Colin F., and Robin M. Hogarth. 1999. "The Effects of Financial Incentives in Experiments: A Review and Capital-Labor-Production Framework." *Journal of Risk and Uncertainty* 19(1–3): 7–42.

Ditto, Peter H., and David F. Lopez. 1992. "Motivated Skepticism: Use of Differential Decision Criteria for Preferred and Nonpreferred Conclusions." *Journal of Personality and Social Psychology* 63(4): 568–84.

Dunning, David, Judith A. Meyerowitz, and Amy D. Holzberg. 1989. "Ambiguity and Self-Evaluation: The Role of Idiosyncratic Trait Definitions in Self-Serving Assessments of Others." *Journal of Personality and Social Psychology* 57(6): 1082–90.

Festinger, Leon. 1950. "Informal Social Communication." *Psychological Review* 57(5): 271–82.

Festinger, Leon. 1957. A Theory of Cognitive Dissonance. Stanford, CA: Stanford University Press.

Gigerenzer, Gerd. 2004. "Dread Risk, September 11, and Fatal Traffic Accidents." *Psychological Science* 15(4): 286–87.

Gilovich, Thomas. 1991. *How We Know What Isn't So: The Fallibility of Human Reason in Everyday Life.* New York, NY: Free Press.

Gilovich, Thomas, and Lee Ross. 2015. *The Wisest in the Room: How You Can Benefit from Social Psychology's Most Powerful Insights.* New York, NY: Free Press.

Gohmann, Johanna. 2015. "Jimmy Stewart Was My Teen Idol." Salon, December 24. http://www.salon.com/2015/12/24/jimmy_stewart_was_my_teen_idol/.

James, William. 1890. *Principles of Psychology*, vol. 2. New York, NY: Cosimo.

Kahneman, Daniel, and Amos Tversky (eds.) 2000. *Choices, Values, and Frames.* New York, NY: Cambridge University Press and the Russell Sage Foundation.

Klar, Yechiel, and Eilath E. Giladi. 1997. "No One in My Group Can Be Below the Group's Average: A Robust Positivity Bias in Favor of Anonymous Peers." *Journal of Personality and Social Psychology* 73(5): 885–901.

Kunda, Ziva. 1990. "The Case for Motivated Reasoning." *Psychological Bulletin* 108(3): 480–98.

Maass, Anne, Daniela Salvi, Luciano Arcuri, and Gún R. Semin. 1989. "Language Use in Intergroup Contexts: The Linguistic Intergroup Bias." *Journal of Personality and Social Psychology* 57(6): 981–93.

Oster, Emily, Ira Shoulson, and E. Ray Dorsey. 2013. "Limited Life Expectancy, Human Capital, and Health Investments." *American Economic Review* 103(5): 1977–2002.

Pronin, Emily, Thomas Gilovich, and Less Ross. 2004. "Objectivity in the Eye of the Beholder: Divergent Perceptions of Bias in Self versus Others." *Psychological Review* 111(3): 781–99.

Schelling, Thomas C. 1978. *Micromotives and Macrobehavior.* New York, NY: W.W. Norton.

Simon, Herbert A. 1956. "Rational Choice and the Structure of the Environment." *Psychological Review* 63(2): 129–138.

Thaler, Richard H. 1991. *Quasi-Rational Economics.* New York: Russell Sage Foundation.

Trope, Yaacpv, and Nira Liberman. 2003. "Temporal Construal." *Psychological Review* 110(3): 403–421.

Vallacher, Robin R., and Wegner, Daniel M. 1987. "What Do People Think They're Doing? Action Identification and Human Behavior." *Psychological Review* 94(1): 3–15.

Mindful Economics: The Production, Consumption, and Value of Beliefs

Roland Bénabou and Jean Tirole

In the economic models of old, agents had backward-looking expectations, arising from simple extrapolation or error-correction rules. Then came the rational-expectations revolution in macroeconomics, and in microeconomics the spread and increasing refinements of modern game theory. Agents were now highly sophisticated information processors, who could not be systematically fooled. This approach reigned for several decades until the pendulum swung back with the rise of behavioral economics and its emphasis on "heuristics and biases" (as in Tversky and Kahneman 1974). Overconfidence, confirmation bias, distorted probability weights, and a host of other "wired-in" cognitive mistakes are now common assumptions in many areas of economics. Over the last decade or so, the pendulum has started to swing again toward some form of adaptiveness, or at least implicit purposefulness, in human cognition.

In this paper, we provide a perspective into the main ideas and findings emerging from the growing literature on *motivated beliefs and reasoning*. This perspective emphasizes that beliefs often fulfill important psychological and functional needs of the individual. Economically relevant examples include confidence in ones' abilities, moral self-esteem, hope and anxiety reduction, social identity, political ideology and religious faith. People thus hold certain beliefs in part because

■ *Roland Bénabou is Theodore A. Wells '29 Professor of Economics and Public Affairs, Princeton University, Princeton, New Jersey and a Senior Fellow, Canadian Institute for Advanced Research, Toronto, Canada. Jean Tirole is Chairman, Toulouse School of Economics, and President of the Executive Committee, Institute for Advanced Study in Toulouse, both in Toulouse, France.*

† For supplementary materials such as appendices, datasets, and author disclosure statements, see the article page at
http://dx.doi.org/10.1257/jep.30.3.141 doi=10.1257/jep.30.3.141

they attach value to them, as a result of some (usually implicit) tradeoff between *accuracy* and *desirability*. Such beliefs will therefore be resistant to many forms of evidence, with individuals displaying non-Bayesian behaviors such as not wanting to know, wishful thinking, and reality denial. At the same time, motivated beliefs will respond to the costs, benefits, and stakes involved in maintaining different *self-views* and *world-views*. These tradeoffs can be influenced by experimenters, allowing for empirical tests, and by a person's social and economic environment, leading to the possibility of self-sustaining "social cognitions."[1]

At an individual level, overconfidence is perhaps the most common manifestation of the motivated-beliefs phenomenon. There is considerable evidence of overoptimistic tendencies on the part of consumers, investors, and top corporate executives (as discussed in a "Symposium on Overconfidence" in the Fall 2015 issue of this journal). While excessive overconfidence is quite dangerous, moderate amounts can be valuable: hope and confidence feel better than anguish and uncertainty, and they often also enhance an individual's ability to act successfully on their own behalf and interact productively with others. Using data from the Survey of Consumer Finances, Puri and Robinson (2007) thus find that more optimistic individuals work more, save more, expect to retire later, and are more likely to remarry after divorce. Alloy and Abrahamson (1979) and Korn et al. (2014) find that most psychologically "healthy" people display some degree of overoptimism and biased updating, while it is primarily depressed subjects who seem to be more objective. People thus find themselves motivated (often unconsciously) to achieve "positive" beliefs, and this typically occurs through a fundamental asymmetry in the *process* by which beliefs are revised in the face of new evidence: individuals update suitably when facing good news, but fail to properly account for bad news (Eil and Rao 2011; Möbius, Niederle, Niehaus, and Rosenblat 2011; Sharot and Garrett et al. 2016).

Although goal-directed, self-deception can nonetheless end up hurting the individual: since it is an informational game that people play with themselves, the outcome may be highly inefficient—a form of self-trap. When motivated thinking becomes a *social* phenomenon, consequences can be even more severe. Collectively shared belief distortions may amplify each other (an issue we shall address), so that entire firms, institutions, and polities end up locked in denial of unpleasant realities and blind to major risks: unsustainable fiscal imbalances or labor market policies, climate change, collapse of housing or financial markets, and so on. Case and Shiller (2003) surveyed the expectations of homeowners during the real-estate bubbles of 1988 and 2003. In both cases, 90 percent of respondents thought housing prices in their city would "increase over the next several years," with an average expected gain for their own property of 9 to 15 percent *per year* over the next ten years. In the political realm, examples of persistent ideological blind spots impeding reforms and of evidence-proof conspiracy theories are abundant.

[1] Parts of this paper draw substantially on Bénabou (2015), which also provides a more explicit treatment of the underlying formal framework.

We now turn to the sources, means, costs, and benefits of motivated cognition. In a sense, we propose to treat beliefs as regular economic goods and assets—which people *consume, invest in,* reap returns from, and *produce,* using the informational inputs they receive or have access to. We first highlight the theory's general principles, then turn to a number of empirical tests and specific applications.

Motivated Beliefs: Why and How?

Why?

For a standard economic agent, information is always valuable, whether the news is good or bad: more data helps make better choices, and if not, it can just be ignored. The value of information exactly equals the extent to which it improves decision-making, and it cannot be negative. Schelling (1988), in contrast, aptly described "the mind as a consuming organ," and indeed we are all familiar with beliefs that have a direct and powerful affective impact. These may be perceptions about ourselves, like self-esteem and self-disappointment (Smith 1759; Bénabou and Tirole 2002; Köszegi 2006), or about the broader environment we face and our prospects in it that evoke strong feelings of fear, anxiety, hope, excitement, and so on (Akerlof and Dickens 1982; Loewenstein 1987; Caplin and Leahy 2001; Brunnermeier and Parker 2005; Eliaz and Spiegler 2006; Bénabou and Tirole 2011). Such "consumable" beliefs can be represented as an argument directly entering the preferences of agents.

Subjective beliefs also often have an important instrumental value, enhancing "self-efficacy." First, confidence in one's ability and chances of success is a powerful motivator to undertake and persevere in long-term projects. This source of demand for "positive" thoughts is generally derived as arising from a self-control problem over effort or tempting consumptions (Carrillo and Mariotti 2000; Brocas and Carrillo 2001; Bénabou and Tirole 2002, 2004). Belief distortions can similarly serve as commitment devices in other settings involving a divergence between preferences that occur before or after a decision, as with an agent who fears "getting cold feet" when a risky decision becomes imminent (Epstein 2008; Eisenbach and Schmalz 2015) or succumbs to "excessive" empathy and generosity when confronted with human misery (Dillenberger and Sadowski 2012). Second, being convinced of one's strength, determination, talent, honesty, and even sincerity helps convince others. Trivers (2011) and Von Hippel and Trivers (2011) hypothesize that this signaling value is why humans may have evolved the capacity to self-deceive, which later on was coopted for other uses.

The framework sketched in the next section will incorporate both classes of motives underlying departures from objective cognition: *affective* (making oneself or one's future look better) and *functional* (helpful to achieve certain goals, internal or external). Religion, the number one form of valued beliefs, typically serves both purposes, simultaneously providing comfort/reassurance and self-discipline.

How?

A consumption or efficacy motive for holding certain self-views and world-views does not ensure that such views will arise and persist, given the constraints and feedback of reality. Because the activities of paying attention (or not), processing, encoding, and rehearsing data predate the stage where we retrieve and ultimately use these signals, however, they open the door to strategic manipulations of our own information, whether conscious or automatic, progressive or abrupt. The strategies of self-deception and dissonance-reduction used to protect valued beliefs are many and varied, but we can group them into three main types: *strategic ignorance, reality denial,* and *self-signaling*.

Strategic ignorance consists in avoiding information sources that *may* hold bad news, for fear that such news could demotivate us, induce distressing mental states, or both. For instance, many at-risk subjects refuse to be tested for Huntington's disease or HIV (Oster, Shoulson, and Dorsey 2013; Ganguly and Tasoff forthcoming) even though the test is free, accurate, and can be done anonymously.

Reality denial is the failure to update beliefs properly in response to bad news. When credible warning signs are received but the feared state of the world is not yet materially incontrovertible, these signals can be processed and encoded in a distorted or dampened manner. Thus, accumulating red flags may indicate an ever-rising probability of disease, or of a housing-market crash, yet agents find ways of not internalizing the data and rationalizing away the risks, as revealed by their unchanged life plans, failure to divest or diversify from risky investments, and so on.

Self-signaling refers to a set of strategies by which the agent manufactures "diagnostic" signals of the desired type, by making choices that he later interprets as impartial evidence concerning his own underlying preferences, abilities, or knowledge about the state of the world (Quattrone and Tversky 1984; Bodner and Prelec 2003; Bénabou and Tirole 2004, 2011). In the health domain, for instance, this corresponds to people who "push" themselves to overcome their symptoms, carrying out difficult or even dangerous activities not only for their own sake, but also as "proof" that everything is fine.

Three Telltale Markers

Three key features differentiate motivated thinking and cognitive tendencies from "mechanical failures" of inference due to bounded rationality or limited attention.

1. Endogenous directionality. In contrast to what are often referred to as "System I" biases, motivated beliefs are by definition directed toward some end, though generally not consciously so. As an example, consider the opposite predictions of *confirmation bias* and *self-enhancement* for how someone who is initially insecure about his skill, attractiveness, or health will respond to feedback about these qualities in himself. A "wired in" confirmation bias would lead him to read any ambiguous signals received as confirming and hardening his negative self-view. This type of response is quite rare and found primarily in clinically depressed individuals. The great majority of people, in contrast, find ways to interpret the same evidence positively, and even clearly bad news as "not that bad," irrelevant, or biased—in line with a self-esteem maintenance motive (Alloy and Abrahamson 1979; Korn et al. 2014). Where confirmation bias typically arises is with respect to external facts, and in such cases it can often be understood as a form of motivated cognition: "This event or data is consistent with what I thought, and it shows I was right." Indeed, an important talent is the ability to analyze situations correctly from the outset. This generates a strong "demand for consistency" in judgments and choices (for evidence, see Falk and Zimmermann 2011), as a positive self-assessment on that dimension increases confidence that someone's personal investments will pay off, thus generating anticipatory utility and motivating people to undertake these projects in the first place

Another example of endogenous directionality is that, in contrast to the case of "built-in" overconfidence, agents will either overestimate or underestimate their own abilities depending on which distortion is advantageous in the situation they expect to face. In particular, when effort and talent are substitutes, rather than complements, building motivation to train requires attributing one's past successes to luck more than talent. Thus, a successful student or athlete may try to think of previous exams and competitions as having been easy compared to the next one that will require additional effort—a form of "defensive pessimism."[2]

2. *Neither naiveté nor lack of attention.* The concept of bounded rationality almost necessarily implies that more analytically sophisticated and better-educated individuals should be less prone to mistakes and biases. Such is indeed the case for the endowment effect, loss aversion, hyperbolic discounting, and even visual illusions (Frederick 2005). However, when it comes to rationalizing away contradictory evidence, compartmentalizing knowledge, and deluding oneself, more educated, attentive, and analytically able people often display greater propensities toward such behaviors. Thus in Kahan (2013) and Kahan, Peters, Dawson, and Slovic (2014), subjects who scored highest on the Cognitive Reflection Test (which measures deliberate and reflective versus intuitive and heuristic thinking) and highest on numeracy tests were less likely to display self-serving failures to update and rationalizations when facing ideologically neutral questions, but *more* likely to do so for ideologically charged issues such as man-made climate change or gun control. In large representative US surveys, Oliver and Wood (2014) similarly find that while education is negatively associated with belief in political conspiracies, political knowledge and interest are not. Ortoleva and Snowberg (2015a) find that overconfidence (about current and future inflation and unemployment) is uncorrelated with education or income and *increases* systematically with media exposure, age, and partisanship.

3. *Heat versus light.* Finally, in "motivated" there is also *emotion.* Challenging cherished beliefs directly—like a person's religion, identity, morality, or politics—evokes strong emotional and even physical responses of anger, outrage, and disgust.

[2] See Bénabou and Tirole (2002) in a context of self-motivation and Charness, Rustichini, and van de Ven (2013) in a context of strategic interactions, where experimental subjects who know they will face a competitive task become overconfident only when such beliefs confer a strategic advantage.

Such pushback is a clear "signature" of protected beliefs: not only would a Bayesian always welcome more data, but so would any naïve boundedly rational thinker.[3] Our emphasis on the interplay of emotions and information-processing is consistent with a similar trend under way in psychology and neuroscience, sometimes referred to as the "affective revolution" or "second cognitive revolution."

A Portable Paradigm

Our conceptual framework for analyzing motivated cognition, both individual and social, draws most closely on Bénabou and Tirole (2002) and Bénabou (2013), but more generally synthesizes a number of ideas common to the large literature on beliefs as direct and/or instrumental sources of utility (Akerlof and Dickens 1982; Loewenstein 1987; Carrillo and Mariotti 2000; Caplin and Leahy 2001; Brocas and Carrillo 2001; Brunnermeier and Parker 2005; Köszegi 2006, 2010). For a formal exposition of this approach, see Bénabou (2015).

A risk-neutral agent has a time horizon of three periods. Period 0 is when information may be sought or avoided, received, and ultimately processed into the beliefs carried into period 1. In period 1, the agent's actions and wellbeing will reflect these posteriors. In period 2, all uncertainty is resolved and final payoffs are received. These depend on the realized state of the world, the action taken at date 1, and possibly an initial endowment such as wealth, human or social capital, genes, or other factors. For simplicity, there are just two possible date-0 signals about the state of the world, L (low) and H (high), corresponding respectively to bad and good news (or, alternatively, bad news and no news, which then constitutes good news) about the return to effort.

A first reason why an individual (Self 0) may want to distort his or her own (Self 1's) beliefs away from what objective information indicates is enhancing *self-efficacy*. If the date-1 decision is subject to a temptation or self-control problem, Self 0 may want to bias Self 1's beliefs about the return to effort, like a parent telling their child that crime *never* pays and homework *always* does. The cost of "maintained optimism" as an internal commitment device is that decisions at date-1 will sometimes be costly mistakes even from an ex-ante point of view, such as attempting a task that is infeasible or even dangerous for the agent. When self-control is enough of a concern, however, some degree of "positive thinking" can be advantageous.

A second class of motives for self-deception is *affective*. The basic framework remains unchanged, except that instead of facing a self-control problem at time 1, the agent derives a direct flow of utility (or disutility) from the beliefs he holds during this period. These hedonic beliefs may be about his own fixed traits: seeing oneself as smart, attractive, and good is intrinsically more satisfying than the reverse. Alternatively, the beliefs can operate through *anticipatory utility*, meaning that the

[3] Sophisticated individuals who anticipate that they might be subject to "cognitive overload" may decline to receive information, but without any hostility.

individual experiences pleasant or aversive emotions from thinking about future (date-2) welfare: health or serious disease, successful marriage or messy divorce, riches or bankruptcy, eternal life or nothingness. Hope, fear, anxiety, and related emotions are important determinants of well-being, both as pure "mental consumptions" and through the psychosomatic, substance-abuse, and relational problems they induce. Under this broad class of preferences, a tradeoff clearly arises at date 0: one can react to bad news objectively, which leads to better decisions but having to live with grim prospects for some time (and possibly a long time), or adopt a more "defensive" cognitive response that makes life easier until the day of reckoning, when mistakes will have to be paid for.

Complementing these two sources of demand for "good" beliefs, the most frequent supply-side building block in the motivated-thinking paradigm is *selective (or differential) updating,* namely processing good and bad signals asymmetrically in term of *attention, interpretation, memory, or awareness.* While psychologically and neurally quite distinct (as discussed in the next section), these mechanisms are *formally equivalent* in terms of updating and behavioral consequences. To avoid repetition, we will often use the selective-recall interpretation, but it should *not* be taken literally. Realism then corresponds to appropriately coding L as L in memory, and denial to miscoding L as H, recalling it as an ambiguous mixture of the two, or (closest to standard information economics) forgetting the news entirely.

A more roundabout "belief-production" process, also based on imperfect accessibility of past states, is *self-signaling*: using our own behaviors as diagnostic of who we are, and conversely making choices with an eye toward our longer-run sense of identity. Because material actions are more easily codified, recalled, and documented than the exact mix of motives that caused them (evaluation of a hard tradeoff, momentary urges, feelings of guilt and pride), our past conduct can be informative about our "deep" preferences and predictive of later behaviors; yet at the same time, our choosing these self-signals makes future beliefs malleable.

Of course, the process of manipulating one's own attention, memory, or awareness must not be too transparent. There must be some opaqueness as to what exactly one is failing to update to, some ambiguity as to why certain actions are taken or not taken—such as crossing the street when seeing a beggar, starting or avoiding a fight, helping someone, or getting drunk. The thinnest of veils will often suffice however, as demonstrated by a number of experiments on "moral wiggle room" in which subjects seize upon or even seek out threadbare excuses for dishonesty, try to ignore or delegate their harming of others, and so on. (Konow 2000; Dana, Weber, and Kuang 2007; Gneezy, Saccardo, Serra-Garcia, van Veldhuizen 2014; Hamman, Loewenstein, and Weber 2010; Di Tella, Perez-Truglia, Babino, and Sigman 2015; Grossman and van der Weele forthcoming).

Naïveté is not needed for the key results (it just makes them stronger). A sophisticated individual knows that he has a tendency toward selective, self-serving attention, recall, and rationalizations. Such "metacognition" leads him to discount somewhat the "absence" of bad news at date 1, but as long as the sophisticated individual cannot fully reconstruct the censored or distorted original information, his posterior will remain inflated. Where sophistication matters is in making possible different "cognitive styles" as alternative personal equilibria. An agent who is aware that his updating is very selective will discount the "news is good" states of awareness substantially, thereby making it safer to censor or misinterpret bad news when it occurs. Conversely, when someone tends to be "honest with himself," good news can be taken at face value, and this self-trust generates strong behavioral responses that make denial of bad news too dangerous.

First Implications and Evidence

From this framework, a number of predictions can be derived and confronted with data.

1. Information Avoidance and Asymmetric Updating

When asset-like beliefs are involved, people will tend to ignore, discount, rationalize away, or "put out of mind" news that conflicts with these ideas while welcoming data that supports them. Möbius et al. (2010) and Eil and Rao (2011) how show that subjects defend their beliefs concerning their IQ and (in the latter paper) also their attractiveness. The experimenters first elicit (in an incentive-compatible fashion) the prior distribution of the beliefs of every participant about being in each decile of the subject pool, then their updated beliefs following each of two rounds of objective feedback in which they learned whether they ranked above or below another, randomly drawn subject. Both studies find a statistically significant *good news/bad news asymmetry,* as predicted by our theory: subjects systematically under-update to negative signals, and are much closer to Bayesian updating for positive ones.[4]

Asymmetry also shows up in the demand for—or the avoidance of—information. In both studies, subjects' willingness-to-pay for learning their true IQ or/and beauty rank at the end of the experiment was positive for those who had arrived at "good" posteriors but *negative* for those who had arrived at "bad" ones, just as patients whose history and symptoms put them at high risk for some major disease often refuse to be tested. Similarly, investors studied in Karlsson, Loewenstein, and Seppi (2009) go online to look up the value of their portfolios much more on days when the market as a whole is up. Gottlieb (2014) formally shows how such conditional *informational preferences* arise from the general selective-recall model.

Wiswall and Zafar (2015) elicit college students' beliefs about their own future earnings and the average earnings in different majors. Then, they provide

[4] Bénabou (2013) shows how the model can generate strict under-updating (relative to Bayes' rule) to bad news and a lesser under-adjustment (possibly none) to good news. Gottlieb (2010) shows that the agents' (endogenous and motivated) failure to learn bad news persists even in an infinite-horizon setting, with signals or feedback received in every period.

the actual figures for each major, and elicit subjects' updated beliefs about their expected incomes. An underestimation of population earnings by $1,000 results in an upward revision in own earnings of $347 (significant at 1 percent), compared with a downward revision of just $159 for an overestimation (significant only at 10 percent).

Asymmetric responses to good and bad news, in turn, readily produce the so-called "Lake Wobegon" effect—that is, a distribution of posteriors where a very high fraction of people see themselves as above average. This holds true even for sophisticated agents, whose posterior beliefs must average back to the population mean, as Bayes' rule does not constrain skewness (Carrillo and Mariotti 2000; Bénabou and Tirole 2002).

2. The Role Of Memory and Other Neural Processes

Several complementary and *de facto* equivalent cognitive mechanisms can sustain motivated updating, but the simplest one is selective recall or accessibility of past signals, which is also relatively easy to test. The first experiment of that type in economics is Thompson and Loewenstein (1992), who show that: i) subjects assigned to represent different parties in a labor negotiation and given the same materials (from an actual case) recall, later on, more facts favoring their side than the other; and ii) these egocentric recall differences were associated with longer (hence costlier) delays during the negotiation phase.

More recently, Chew, Huang, and Zhao (2013) have subjects take four questions from an IQ test. Two months later they are shown the same four questions, plus two they had never seen, together with all the answers, and are *incentivized* to recall how they answered each one, or if they did not encounter it before. The probability of "remembering" having correctly answered a question which one actually failed is six times as high as the probability of the reverse error. The probability of not remembering one's answer, or whether one saw a question, is on average twice as high if the answer was wrong than if it was right. As for the questions they had never seen, 56 percent of subjects "remembered" answering them correctly versus 9 percent incorrectly. Furthermore, the three types of positive-attribution recall biases were highly correlated across subjects.

Work in neuroscience is starting to explore the deep mechanisms involved in differential recall and updating. Benoit and Anderson (2012) show that people are able to lower their later recall rates (for word pairs) by either blocking associations as they start to resurface or by focusing on different thoughts, and that different brain networks are involved in these two processes of *voluntary forgetting*. Sharot, Korn, and Dolan (2012) confirm the general finding of asymmetric updating to good and bad news and show that, while the "raw" data are well remembered by their subjects, distinct regions of the prefrontal cortex track and code for positively versus negatively valenced (more or less desirable) implied estimation errors. Furthermore, highly optimistic individuals consistently exhibit reduced tracking of negative estimation errors (which require updating in the direction of bad news).

3. Costs and Salience

Beliefs for which the individual cost of being wrong is small are more likely to be distorted by emotions, desires, and goals. An example often given is voting, as the cost of holding mistaken political opinions is usually said to be proportional to the probability of being pivotal, and hence extremely small (Caplan 2007). In reality, it need not be, due to social and self-signaling costs of political convictions. The more difficult question lies elsewhere, however. For the cognitive distortions of voters to have policy implications, it is necessary that a majority of them occur in the same direction. How such *ideological alignments* may occur and become dominant will be discussed in a later section, after extending the basic framework to social cognition.

4. Stakes-dependent Beliefs

Consider an agent with anticipatory utility who entered period 0 with some illiquid asset—housing, over-the-counter securities, specialized human or social capital, culture, or religion—that, at time 2, will be more valuable in state H than in state L. The incentive to self-deceive following bad news is clearly stronger, the greater is the amount of "sunk" capital with which the agent is initially endowed. This key implication of the motivated-cognition framework, which we term s*takes-dependent beliefs*, was first demonstrated in psychology by Kunda (1987).

Babcock, Loewenstein, Issachroff, and Camerer (1995) provide further evidence with an incentivized economic experiment. Pairs of subjects were given the same case file from a lawsuit over a traffic accident and were randomly assigned to be either the advocate for the plaintiff or for the defendant. They then bargained over a monetary settlement, with costs of delay. Based on the common materials they received, both sides also (independently) made incentivized predictions as to what outsiders would deem fair and how the judge ruled on the case. When roles were assigned *before* subjects saw the materials, they made highly divergent predictions of fairness and legal outcomes, and incompatible demands, leading to costly delays and breakdowns in bargaining. When roles were assigned *after* the information-processing stage, in contrast, there was far less asymmetry and delay.

In Mijovic-Prelec and Prelec (2010), subjects made incentivized predictions about a series of binary events, both before and after being (randomly) given stakes in the outcomes. When the stakes were such that their initial forecast corresponded to a low-payoff state, subjects showed a significant propensity to reverse their prediction. This is not just inconsistent with rational expectations, but also the *exact opposite* of confirmation bias. Mayraz (2011) has subjects randomly assigned to being "farmers" or "bakers" forecast the price at which they will later trade. Their predictions again vary systematically and optimistically with their positions, as well as with the size of the monetary stakes involved in facing favorable terms of trade. In the field, Di Tella, Galiant, and Schargrodsky (2007) document how land squatters randomly granted property rights adopted more "pro-market" beliefs (possibility of succeeding on one's own, importance of money for happiness), relative to their less-lucky neighbors.

5. Sunk-Cost Fallacy, Escalating Commitment, and the Hedonic Treadmill

A person who starts with enough of some illiquid or sunk asset, generating strong incentives to persuade oneself of its future value, experiences a type of endowment effect. Once persuaded, he will want to invest more in this capital, succumbing to a form of the sunk-cost fallacy that psychologists refer to as *escalating commitment*. Furthermore, although the agent is optimizing at every point in time given current preferences and beliefs, the ex-ante welfare implications of such ratcheting accumulation or specialization can be negative (Bénabou and Tirole 2011). The easiest way to understand this *hedonic-treadmill* result is to think of the case where self-signaling through personal actions is the way the agent tries to manipulate his own beliefs, and to recall that signaling usually involves a deadweight loss. More generally, censoring bad news or trying to offset it through identity-enhancing behaviors can prevent a deterioration of beliefs (like moral self-esteem) in bad states, but such censoring also reduces confidence that good states are really what they seem to be (creating self-doubt). When agents are sophisticated and beliefs enter preferences linearly, the two effects cancel out, leaving only the costs of generating, and then acting on, incorrect beliefs.[5]

Social and Organizational (Mis)Beliefs

Investigation reports following public-agency and corporate disasters commonly describe how willful blindness and reality denial spread within the organization, leading to systemic failures. A large literature in organizational psychology similarly emphasizes the key roles of moral self-deception and overoptimistic hubris in misconduct and financial fraud (Tenbrunsel and Messick 2004; Anand, Ashforth, and Mahendra 2004; Bazerman and Tenbrunsel 2011; Schrand and Zechman 2012). People engaging in reckless or dishonest behavior find ways to convince *themselves* that they are doing nothing wrong, so transgressions that typically start small gradually escalate through a series of rationalizations, which are then further insulated from reality by "echo chamber" group dynamics. For instance, the NASA (2003, vol. 1, pp. 196-199) investigations following the Challenger and Columbia space shuttle accidents found that:

> NASA appeared to be immersed in a culture of invincibility, in stark contradiction to post-accident reality. The Rogers Commission found a NASA blinded by its 'Can-Do' attitude … which bolstered administrators' belief in an achievable launch rate, the belief that they had an operational system, and an unwillingness to listen to outside experts … At every juncture, the Shuttle Program's

[5] The case of linear utility-from-beliefs is a useful benchmark. Clearly, if the functional is instead concave (respectively, convex) in beliefs, the agent will gain from achieving coarser (respectively, more dispersed) posteriors. The actual shape of self-esteem or anticipatory preferences is, ultimately, an empirical question.

> structure and processes, and therefore the managers in charge, resisted new information … [E]vidence that the design was not performing as expected was reinterpreted as acceptable and non-deviant, which diminished perceptions of risk throughout the agency … Engineers and managers incorporated worsening anomalies into the engineering experience base, which functioned as an elastic waistband, expanding to hold larger deviations from the original design.

Strikingly similar patterns recurred at companies like Enron and General Motors and, prior to the 2008 financial crisis, at major investment banks, the insurance company AIG, the Federal Reserve, and the Securities and Exchange Commission (Bénabou 2013, Appendix D).

How can such motivated thinking and reality denial become "contagious" and spread through an organization or some of its units? Consider a setting in which individuals are embedded in a firm, network, or other collective endeavor. To highlight the endogenous emergence of interdependence in how people *think* and perceive events, let us assume (without loss of generality) a simple, linear interaction structure: each agent's final payoff is a weighted average of his own action and the group's average action, all multiplied by a common (gross) return. In the good state of the world, both private and social (group-wide) net returns are positive. In the bad state, the (net) private return is always negative, but depending on the nature of spillovers, the public return could be positive or negative. This last factor turns out to be critical for how groups respond to bad news, and whether a collective failure to update represents beneficial *group morale* or harmful *group delusions.*

In the case of projects with no or little social downside, like team effort or mobilization for a good cause, blind perseverance in the face of bad news is individually suboptimal but constitutes a *public good.* The overoptimism of others thus makes the bad state more tolerable, and therefore each individual more willing to accept its reality: cognitive attitudes are thus *strategic substitutes* (they tend to dampen one another), and denial is self-limiting.[6]

The more interesting case is that of ventures with important downside risk, in which blind persistence can inflict further losses on others, such as capital and reputational losses, firm bankruptcy, layoffs, catastrophic accident, or prosecution. The more people fail to attend to bad news and continue doing "business as usual," the worse the bad state becomes, making it even harder to face the impending disaster. Perceptions of reality are now *strategic complements,* so delusions will spread.

This Mutually Assured Delusion (MAD) mechanism is rather perverse, as denial and reality avoidance become *contagious* when they are socially *harmful,* but *not* when they are *beneficial.* The underlying intuition is straightforward: we saw earlier how each individual tends to align their beliefs with the fixed stakes they have in different

[6] In a sufficiently asymmetric interaction structure, it can even be that some agent who can short-sell the project gains so much from others' denial of state L that he prefers it to H. In that case, he will have a tendency to believe in H rather than L. This strong cognitive substitutability can lead two (sets of) agents to take opposite sides of a bet on which state will realize, as in Brunnermeier and Parker (2005).

states of the world. In a group or network, these stakes now depend on what other people do (do they generate positive or negative spillovers?), and hence on what they *believe,* in those states. It follows that what is optimal for each agent to think *depends on what others think,* and vice versa. Furthermore, the nature and welfare consequences of these cognitive linkages depend quite simply on the sign of externalities in the network structure, rather than on any built-in nonlinearities in payoffs.

This "psychological multiplier" leads to the possibility of *multiple social cognitions:* fundamentally similar groups or organizations can operate either in a *realistic mode* where everyone faces the facts as they are, or in a *delusional mode* in which everyone engages in denial of bad news, which in turn makes those states even worse for everyone else. Bénabou (2013) shows that such "groupthink" is more likely: i) when codependency among group members is high, meaning that they share a largely common fate, with few exit options from the collateral damage inflicted by others' mistakes; and ii) when the adverse state of the world is relatively rare but, when it occurs, really bad—a so-called "black swan" event.

These ideas and results readily extend to *asymmetric networks* and organizations: an agent's propensity to realism or denial depends most on how the people whose decisions have the strongest impact on his fate respond to bad news themselves. Therefore, in a hierarchy, top management's (mis)perceptions of market prospects, legal liabilities, or odds of victory will tend to trickle down to middle echelons, and from there on to workers or troops.

We have emphasized in this section the importance of "bad beliefs" mechanisms of organization failure, which have so far received too little attention from economists relative to the standard "bad incentives" mechanisms. In practice, most failures have both channels at work, and how they feed into each other represents a promising avenue for future research.

Political Ideology

The study of political economy is also undergoing a pendulum swing of perspectives. An emphasis on the strategic choices of rational voters and pressure groups pursuing their material self-interest remains indispensable, but it is increasingly being complemented by "behavioral" considerations like the expression of identities and emotions, reference-dependent concerns such as fairness or loss aversion, biased attributions (like scapegoating), and ideological or wishful denials of reality. The ongoing political events and campaigns in the United States and a number of European countries should, if need be, dispel any remaining doubts about the relevance of such psychological factors in politics.

While each voter may choose to maintain beliefs which they value for affective or instrumental reasons, in order for this to have policy consequences these worldviews must somehow align within a country, while potentially diverging across borders. We now provide examples of how such complementarities in political beliefs can arise rather naturally, leading to the emergence and persistence of *dominant ideologies.*

Just-World Beliefs

A “just world” is one in which “people get what they deserve, and deserve what they get” (Lerner 1980). Do they? The World Values Survey reveals considerable differences in beliefs about the role of effort versus luck in life. In the United States, 60 percent of people believe that effort is key; in Western Europe, only 30 percent do on average, with major variations across countries. Moreover, these nationally dominant beliefs bear no relationship to the actual facts about social mobility or how much the poor are actually working, and yet they are strongly correlated with the share of social spending in GDP (Alesina, Glaeser, and Sacerdote 2001). At the individual level, similarly, voters’ perceptions of the extent to which people control their own fate and ultimately get their just desserts are first-order determinants of attitudes toward inequality and redistribution, swamping the effects of own income and education (Fong 2001).

In Bénabou and Tirole (2006), we describe how such diverse *politico-ideological equilibria* can emerge due to a natural complementarity between (self-)motivation concerns and marginal tax rates. When the safety net and redistribution are minimal, agents have strong incentives to maintain for themselves, and pass on to their children, beliefs that effort is more important than luck, as these will lead to working hard and persevering in the face of adversity. With high taxes and generous transfers, such beliefs are much less adaptive, so fewer people will maintain them. Thus, there can coexist: i) an “American Dream” equilibrium, with just-world beliefs about social mobility, and little redistribution; and ii) a “Euro-pessimistic” equilibrium, with more cynical beliefs and a large welfare state. In the latter, the poor are less (unjustly) stigmatized as lazy, while total effort (annual hours worked) and income are lower, than in the former. More generally, across all steady-states there is a negative correlation between just-world beliefs and the size and the welfare state, just as observed across countries.

Complementing this national-level evidence, Frank, Wertenbroch, and Maddux (2015) experimentally validate the role (and malleability) of just-world beliefs in determining distributional preferences. Using MBA participants from 30 countries, they find that: i) subjects’ priors on the effort-versus-luck question predict their preferences toward redistribution of earnings from a task performed in the lab; ii) aggregating these beliefs at the national level yields a predictor of preferences (from national surveys) for performance pay versus redistributive pay; and iii) “priming” just-world beliefs so that they are more prominently in mind at the start of the experiment has a causal effect on individuals’ choices over which pay system to impose in their session.

Statist and Laissez-Faire Ideologies

A similar international divergence is observed for beliefs in the merits of “the free enterprise system and free market economy.” The average degree of agreement that this is “the best system on which to base the future of the world” was 61 percent in the 2005 World Public Opinion Survey. Countries near the top include China at 74 percent, the United States at 71 percent, and Germany at 65 percent. Those at the bottom include Argentina at 42 percent, Russia at 43 percent, and France at 36 percent. Here again the objective facts belie these divergent worldviews; Germany and France, for instance, have very similar economic structures but an almost two to one divergence on this survey question. Yet again, these beliefs are highly predictive of the size of government, whether measured by the tax-to-GDP ratio or by indices of labor and product market regulation.

Bénabou (2008) documents this international divergence and shows how such ideological differences can be sustained. When people expect to be "living with" and paying for a large public sector, the psychological incentive is to view it as an important source of future benefits (anticipatory utility), which in turn makes one more willing to vote for it. Conversely, when people anticipate having to purchase these services in the market (as in the United States in the case of health insurance), the incentive is to think of the latter as efficient and see less need for public provision and funding. Individual voters' beliefs can thus again be mutually amplifying, leading to history-dependent dynamics and multiple steady-states: a "Statist" one featuring a large government and obstinate beliefs in its benevolent efficacy, a "Laissez-Faire" one with a small government and equally inflexible beliefs in the virtues of the invisible hand, or a "Realistic" one in which voters acknowledge both state and market failures.

Pandering Politicians

If voters have demands for rosy beliefs—say, painless solutions to economic and social problems, external scapegoats, or feel-good demonstrations of power—office-motivated politicians and profit-maximizing media will gladly oblige. As shown by Levy (2014), this feedback can lead to a "Soothing Politics" equilibrium, which features no reform even when needed, hence much pain down the line. This prospect increases voters' incentives to forget or rationalize bad news, and, in turn, their inattention and wishful thinking allow politicians whose interests are noncongruent with those of the electorate to indulge in the easy life of no reform. Conversely, a "*Realpolitik*" equilibrium can emerge in which voters remain aware of negative signals. In this case, politicians must follow up with reform, to avoid appearing noncongruent (lazy, incompetent, or captured by lobbies) and being voted out.

Overconfidence, Polarization, and Extremism

Within each polity there are also sharply divided beliefs, often with a tendency toward polarization rather than convergence. Here again, the perspective of voters bending their thought processes and worldviews to fit their needs and desires provides a useful explanatory framework. Agents whose material, social, human-capital, and cultural endowments give them different stakes in various states of the world (large or small role of effort; efficient or inefficient government; degree of trustworthiness of others) will process and interpret the same signals very differently. A greater divergence of beliefs even as more and more information becomes commonly available through the global media and internet is thus not really a puzzle.

This class of models also implies that people will seek interactions with those who think like them, and shun those whose words or actions provide signals and reminders that threaten valued "constructed realities" (Bénabou and Tirole 2011). There is an important role here for Cassandras, who speak unpleasant truths, to guard against the group falling prey to costly delusions (this also ensures that good news, or the absence of bad news, is genuine). Yet these truth-tellers will be cast away, or worse, once a bad state does occur, especially if some investments have already been sunk (Bénabou 2013). This *time-inconsistency* in attitudes toward dissent provides a new rationale for social commitment mechanisms such as constitutional rights to free speech, independence of the press, and so forth. Not only does the public presence of dissenting views help to ensure realism and confidence in the available information, but the anticipation that they will undermine wishful beliefs also lowers the return to engaging in motivated thinking in the first place.

Overconfidence and Ideology

The importance of overconfident beliefs in politics and their resistance to information, particularly among extremists, is documented in Ortoleva and Snowberg (2015a, b). A nationally representative sample of over 3,000 American adults was asked standard political-survey questions, and also to provide estimates and degrees of confidence for the current and next year's rates of inflation and unemployment. The study finds that: i) more overconfident agents have more extreme political views, and higher turnout rates in elections (thus cognitive distortions really matter); ii) overconfidence does not decrease with education, and it increases with both age and media exposure (polarization); and iii) it is found both on the Left and the Right of the political spectrum, though since 1980 more so on the Right. To explain these findings, Ortoleva and Snowberg (2015a) propose that voters suffer from heterogeneous degrees of "correlational neglect"—that is, they fail to take account that the observations they derive from their local environment, social network, and chosen information sources are largely redundant.

Such failures of Bayesian inference and biased information seeking and processing arise very naturally from a motivated-beliefs, identity-maintenance perspective. Thus, Le Yaouanq (2015) incorporates preference heterogeneity into the selective-awareness framework exposited above. He shows that the typical result of multiple dominant ideologies remains and, more interestingly, that: i) the more extremist agents in the political spectrum are the ones most prone to engage in reality denial; and ii) when agents can endogenously form networks within which political views will be exchanged or observed, *ideological homophily* will tend to prevail, making collective biases and polarization more likely (Della Vigna and Kaplan 2007; Gentzkow and Shapiro 2011).

Financial Bubbles and Crashes

The motivated-thinking framework also provides a psychologically grounded account for financial manias and crashes. Suppose that, following some initial good news, investors have accumulated stocks of some financial asset that is relatively illiquid. Next, good or bad news may be received about fundamentals, and investors can then keep investing or stop. The liquidation price at date 2 will reflect the total supply accumulated up to that time and the realized demand for the asset. Downward-sloping demand makes investment decisions strategic substitutes, and thus contagion harder to sustain. Nonetheless, investors' *cognitive responses* to bad news can be *strategic complements* (they reinforce one another), giving rise to an "irrationally exuberant" buildup that further amplifies the coming crash. Indeed, when illiquid initial positions are sufficiently large, realism would require *recognizing*—in both senses of the term—early on major capital losses, made all the worse by the blind overinvestment of others. This capital-loss externality is the Mutually Assured Delusion (MAD) multiplier at work again, and when it dominates demand substitutability the market can be seized by periodic waves of contagious overoptimism, overheating, and meltdowns (Shiller 2005; Akerlof and Shiller 2009).

Cheng, Raina, and Xiong (2014) provide evidence supporting the relevance of such a mechanism in the real-estate-based financial bubble of 2003–2005. They examine the personal housing transactions of Wall Street "insiders"—a sample of 400 mid-level managers in the mortgage-securitization industry, who had a close-up view of the toxic subprime loans. Compared to sophisticated "outsiders"—lawyers and financial analysts not specializing in the real-estate sector—the insiders were *more* likely to buy a first, second, or larger house at the peak of the bubble, and slower to divest as housing prices started falling. As a result, they had a lower overall return on their own real-estate portfolios. The fact that insiders bought high and sold low goes against standard, rational moral-hazard explanations of the crisis, but it is very consistent with the mechanisms of escalating commitment and groupthink in which beliefs about future housing prices become badly distorted by personal and industry-wide stakes.

Identity and Morality

Psychologists, sociologists, and more recently economists (starting with Akerlof and Kranton 2000) have emphasized the central role that identity plays in determining social behavior. Identity is in essence a set of beliefs: about one's character, preferences, moral or religious values, abilities, and prospects (personal identity); and about where one belongs—within a family, firm, network, culture, or nation (group identity). Identity pertains to beliefs that people *value*—and therefore defend. For example, beliefs in an afterlife clearly affect anticipatory utility, just-world beliefs provide both motivation to act and a sense that life is somewhat predictable, and trust in others makes us more optimistic about the society we live in.

Personal Identity

There are obvious reasons why people would want to be seen by others as honest and prosocial. But the desire to think of *oneself* as a moral person—and the concomitant monitoring and judgment of one's behavior—is more subtle to explain.

A first benefit of maintaining such self-respect is to help resist short-run temptations to act opportunistically (cheating, defecting) or rashly (impulses toward sex, anger, or violence) which are likely to have detrimental long-run consequences for one's social relationships. Another adaptive benefit is that deceiving others is not that easy (we are "programmed" to blush or more broadly to give ourselves away), so that believing one's own "line" helps one profess it more credibly in public. In an experiment with incentive-compatible belief elicitations, Schwardman and van der Weele (2016) find that after performing a cognitively challenging task, subjects were 50 percent more overconfident in their relative performance, and less responsive to objective feedback on it, if informed in advance that they could later earn money by convincing others, face-to-face, of having scored high. Furthermore, this belief-management strategy was effective, as subjects in this condition did receive better assessments (also incentivized) from evaluators.

On the other hand, any form of reality distortion or wasteful signaling has costs as well: for instance, forcing us to act more generously than we really would like to, even in anonymous settings. This explains why, as demonstrated by a series of "moral wiggle room" experiments, people often seek excuses and situational ambiguities that avoid putting their moral identities to an explicit and necessarily costly test (in this issue, Gino, Norton, and Weber discuss the evidence on this point). Di Tella et al. (2015) actually elicit the beliefs generated in the process using a variant of the "trust game."[7] When given the opportunity to unconditionally confiscate all of the trustee's earnings, without *any* knowledge of how he or she had actually behaved, trustors became significantly more likely to predict (with incentives) that the trustee had chosen a "corrupt" action and then "punish" them by taking their money.

Social Identity

Many core identities relate to "belonging," such as identifying with a town, ethnic or cultural group, profession, religion, political party, or public cause. A first explanation may be that we derive material benefits from being part of a community. Berman and Iannaccone (2006) thus note that a number of religious groups provide "club goods" such as insurance against economic or health shocks, help in finding a spouse or job, and assistance with raising children. However, it is not at all obvious why public-goods clubs would have to rely on supernatural beliefs (which are absent from these models), as opposed to membership fees in money or in kind, reputational enforcement, or other signaling devices.[8] It is also not clear why religious beliefs and identities should be shared and sometimes violently defended across the world, when the above public goods are in most cases extremely local.

[7] In the canonical trust game, one player chooses how much of their endowment to entrust to another one. This investment gets multiplied (say, by three), then the "trustee" decides how much to return to the "trustor."

[8] On the choice between costly beliefs and costly rituals as signals, see Levy and Razin (2014).

A second explanation for a general "desire to belong" is evolutionary: humans are a social species, deriving an intrinsic satisfaction from interactions with similar others and deep anxiety from isolation—emotional incentives that serve to promote the fitness benefits of community. Even today, the quasi-automatic nature of the in-group/out-group phenomenon, in which shared identities and beliefs form almost immediately on the basis of meaningless and random group assignments, speaks to our need not to be left in a social no-man's land.

Beliefs as Identity Capital

Although economic treatments of identity generally describe it as a set of beliefs, in practice they often model it as preferences, or meta-preferences over utility functions (Akerlof and Kranton 2000; Rabin 1994; Shayo 2010). Our approach is cognitive, in that it explicitly models identity as beliefs about one's deep values, emphasizing the self-inference process through which beliefs operate and the underlying needs that identity serves.

The individual starts with a stock of identity-relevant capital that will affect future welfare. This stock can be viewed as fixed (gender, race) or augmentable (friends, professional accomplishments, wealth, religious faith). Critically, the agent is uncertain about how much this capital will contribute to his own welfare over the long run. Thus, an immigrant may at times be unsure of how attached he is to his original culture relative to the benefits of integration; a professional, of whether more accomplishments and wealth or more time with the family will make for a happier life; and a religious person, of the true strength of his faith. As explained earlier for "stakes-dependent" beliefs more generally, a positive view of the future returns to these stocks directly raises anticipatory utility, and may also enhance self-regulation.

This self-reputational approach to identity sheds light on many otherwise puzzling aspects of behavior, particularly in the moral domain. First, identity-enhancing behaviors are more likely when objective information about deep preferences is scarce (like true generosity, loyalty, or faith), and they are easily affected by minor manipulations of salience such as cues, reminders, and semi-transparent excuses that can be used to muddle personal responsibility. Second, whereas challenges to a weakly held identity (prior belief) may elicit acquiescence, challenges to a strongly held identity generally elicit forceful counterreactions. Another form of history-dependence is "hedonic treadmill" effects: when agents are endowed with sufficient identity capital, they will tend to keep investing even in the face of negative net returns. In some cases, a person's conflicting identities can actually generate self-destructive behaviors: rejecting education and integration into the mainstream labor market, unwillingness to adapt to societal or economic change, going to fight in a faraway land, or suicide bombings. In a milder example of this phenomenon, Burzstyn et al. (2014) found that 25 percent of male experimental subjects in Pakistan chose to forego a participation bonus equivalent to one-fifth of a day's wage when receiving it required anonymously checking a box indicating gratitude toward the US government for the funds.

A beliefs-based model also naturally generates *sacred values* and mental *taboos* (not just social ones), characterized by a strong aversion to even *thinking* about violations: the mere contemplation of tradeoffs between some "higher" principles and self-interest suffices to cast lasting doubt on one's identity (Fiske and Tetlock 1997; Bénabou and Tirole 2011). This creates the potential for significant reality distortion in realms where identity concerns loom large, and can also account for the desire to ban "repugnant markets" where *others* might too visibly engage in certain taboo transactions: prostitution, organ sales, payments for adoption or surrogate pregnancy, and so on.

Ostracism is another natural implication. Since the preferences and prospects of similar individuals are often correlated, "deviant" behavior by peers—violating norms and taboos, fraternizing with outsiders—conveys bad news about the value of existing social assets (anticipatory-utility motive) or that of future investments in them (imperfect self-control motive). On the other hand, if the morally dubious action was one's own, it is good behavior by others that becomes threatening, as it takes away potential excuses involving situational factors. In *both* cases, ostracizing the deviators suppresses the undesirable *reminders* created by their presence. Thus, depending on who acted more (im)morally, the same person or group will act pro-socially and shun free riders, or act selfishly and shun moral exemplars (for empirical evidence, see Monin 2007, Herrmann, Thöni, and Gächter 2008; on the demand for belief consonance and the "tyranny of small differences," see also Golman et al. in this volume).

Identity investments also help explain inefficiencies in bargaining, contracting, and the functioning of organizations. The failure to reach efficient Coasian deals—leading to legal trials, divorces, strikes, scapegoating of minorities in hard times, wars, and so on—is usually explained by economists through informational asymmetries about gains from trade and outside options. Evidence is accumulating, however, that belief distortions also play a key role in those phenomena, with field studies such as Bewley (1999) and Krueger and Mas (2004) complementing the previously mentioned experiments of Babcock et al. (1995). Pride, dignity concerns, and wishful thinking commonly lead people or groups to walk away from "reasonable" offers, try to shift blame for failure onto others, and destroy surplus and seek refuge in political utopias, resulting in costly impasses and conflicts. (For a model of bargaining with motivated beliefs, see Bénabou and Tirole 2009.)

Conclusion

The basic utility function based on consumption and leisure (or even social payoffs) was always recognized as a simplification—defensible in many cases, less so in others. The theory of motivated cognition broadens the purposefulness of human behavior along a variety of dimensions. Some beliefs and emotions are affectively more pleasant than others, like hope and confidence over fear and anxiety. People receive utility from having a positive self-image, and from thinking of themselves as belonging to groups. Optimistic beliefs can also be valuable motivators to overcome self-control problems, as well as helpful in strategic interactions.

In such situations, people will tend to manipulate their collection and processing of information in ways that depart from strict Bayesian inference, trading off the affective or functional value of belief distortions against the costly mistakes they also induce. It may seem that they are just displaying limited cognitive abilities due to some the biases discussed in the large behavioral-economics and bounded-rationality literatures. Instead, the theory of motivated beliefs emphasizes that many observed departures from standard rationality are not hard-wired or mechanical but can instead be understood within a broadened context of goal-directed (but not necessarily efficient) individual behavior. This, in turn, leads to novel views of risk-taking, prosociality, identity, organizations, financial crises, and politics.

■ *We are grateful to S. Nageeb Ali, Rafael DiTella, Gordon Hanson, Ulrike Malmendier, Enrico Moretti, and Timothy Taylor for valuable comments. Bénabou gratefully acknowledges financial support from the Canadian Institute for Advanced Research, and Bénabou and Tirole from the European Research Council (Grants No. FP7/2007-2013-249429 and No. 669217).*

References

Akerlof, George, and William T. Dickens. 1982. "The Economic Consequences of Cognitive Dissonance." *American Economic Review* 72(3): 307–19.

Akerlof, George A., and Rachel E. Kranton. 2000. "Economics and Identity." *Quarterly Journal of Economics* 115(3): 716–753.

Akerlof, George A., and Robert J. Shiller. 2009. *Animal Spirits: How Human Psychology Drives the Economy, and Why It Matters for Global Capitalism.* Princeton University Press.

Alesina, Alberto, Edward Glaeser, and Bruce Sacerdote. 2001. "Why Doesn't the United States Have a European-Type Welfare State?" *Brookings Papers on Economic Activity,* no. 2, pp. 187–277.

Alloy, L. B., and L. Y. Abrahamson. 1979. "Judgment of Contingency in Depressed and Nondepressed Students: Sadder but Wiser?" *Journal of Experimental Psychology: General* 108(4): 441–85.

Anand, Vikas, Blake E. Ashforth, and Joshi Mahendra. 2004. "Business as Usual: The Acceptance and Perpetuation of Corruption in Organizations." *Academy of Management Perpectives* 18(2): 39–53.

Babcock, Linda, George Loewenstein, Samuel Issacharoff, and Colin Camerer. 1995. "Biased Judgments of Fairness in Bargaining." *American Economic Review* 85(5): 1337–43.

Bazerman, Max H., and Ann E. Tenbrunsel. 2011. *Blind Spots: Why We Fail to Do What's Right and What to Do About It.* Princeton University Press.

Bénabou, Roland. 2008. "Ideology." *Journal of the European Economic Association* 6(2–3): 321–52.

Bénabou, Roland. 2013. "Groupthink: Collective Delusions in Organizations and Markets." *Review of Economic Studies* 80(2): 429–62.

Bénabou, Roland. 2015. "The Economics of Motivated Beliefs." Jean-Jacques Laffont Lecture, *Revue d'economie politique* 125(5): 665–85.

Bénabou, Roland, and Jean Tirole. 2002. "Self-Confidence and Personal Motivation." *Quarterly Journal of Economics* 117(3): 871–915.

Bénabou, Roland, and Jean Tirole. 2004. "Willpower and Personal Rules." *Journal of Political Economy* 112(4): 848–87.

Bénabou, Roland, and Jean Tirole. 2006. "Belief in a Just World and Redistributive Politics." *Quarterly Journal of Economics* 121(2): 699–746.

Bénabou, Roland, and Jean Tirole. 2009. "Over My Dead Body: Bargaining and the Price of Dignity." *American Economic Review* 99(2): 459–65.

Bénabou, Roland, and Jean Tirole. 2011. "Identity, Morals and Taboos: Beliefs as Assets." *Quarterly Journal of Economics* 126(2): 805–55.

Benoit, Roland G., and Michael C. Anderson. 2012. "Opposing Mechanisms Support the Voluntary Forgetting of Unwanted Memories." *Neuron* 76(2): 450–60.

Berman, Eli, and Laurence Iannaccone. 2006. "Religious Extremism: The Good, the Bad, and the Deadly." *Public Choice* 128(1): 109–129.

Bewley, Truman F. 1999. *Why Wages Don't Fall During a Recession.* Cambridge, MA: Harvard University Press

Bodner, Ronit, and Drazen Prelec. 2003. "Self-Signaling and Diagnostic Utility in Everyday Decision Making." In *The Psychology of Economic Decisions*, vol. 1, edited by Isabelle Brocas and Juan D. Carrillo. Oxford University Press.

Brocas, Isabelle, and Juan D. Carrillo. 2001. "Rush and Procrastination under Hyperbolic Discounting and Interdependent Activities." *Journal of Risk and Uncertainty* 22(2): 141–64.

Brunnermeier, Markus K., and Jonathan A. Parker. 2005. "Optimal Expectations." *American Economic Review* 95(4): 1092–1118.

Bursztyn, Leonardo, Michael Callen, Bruno Fermanx, Saad Gulzar, Syed Ali Hasanain, and Noam Yuchtman. 2014. "Identifying Ideology: Experimental Evidence on Anti-Americanism in Pakistan." NBER Working Paper 20153.

Caplan, Bryan. 2007. *The Myth of the Rational Voter: Why Democracies Choose Bad Policies.* Princeton University Press.

Caplin, Andrew, and John Leahy. 2001. "Psychological Expected Utility Theory and Anticipatory Feelings." *Quarterly Journal of Economics* 116(1): 55–80.

Carrillo, Juan D., and Thomas Mariotti. 2000. "Strategic Ignorance as a Self-Disciplining Device." *Review of Economic Studies* 67(3): 529–44.

Case, Karl E., and Robert J. Shiller. 2003 "Is There a Bubble in the Housing Market?" *Brookings Papers on Economic Activity* no. 2, pp. 299–342.

Charness, Gary, Aldo Rustichini, and Jeroen van de Ven. 2013. "Self Confidence and Strategic Behavior." CESifo Working Paper 4517, December.

Cheng, Ing-Haw, Sahil Raina, and Wei Xiong. 2014. "Wall Street and the Housing Bubble." *American Economic Review* 104(9): 2797–2829.

Chew, Soo Hong, Wei Huang, and Xiaojian Zhao. 2013. "Selective Memory and Motivated Delusion: Theory and Experiment." Unpublished paper, National University of Singapore, February.

Dana, Jason, Roberto A. Weber, and Jason Xi Kuang. 2007. "Exploiting Moral Wriggle Room: Experiments Demonstrating an Illusory Preference for Fairness." *Economic Theory* 33(1): 67–80.

Dillenberger, David, and Philipp Sadowski. 2012. "Ashamed to Be Selfish." *Theoretical Economics* 7(1): 99–124.

DellaVigna, Stefano, and Ethan Kaplan. 2007. "The Fox News Effect: Media Bias and Voting." *Quarterly Journal of Economics* 122(3): 1187–1234.

Di Tella, Rafael, Sebastian Galiant, and Ernesto Schargrodsky. 2007. "The Formation of Beliefs: Evidence from the Allocation of Land Titles to Squatters." *Quarterly Journal of Economics* 122(1): 209–41.

Di Tella, Rafael, Ricardo Perez-Truglia, Andres Babino, and Mariano Sigman. 2015. "Conveniently Upset: Avoiding Altruism by Distorting Beliefs about Others' Altruism." *American Economic Review* 105(11): 3416–42.

Eliaz, Kfir, and Ran Spiegler. 2006. "Can Anticipatory Feelings Explain Anomalous Choices of Information Sources?" *Games and Economic Behavior* 56(1): 87–104.

Eil, David, and Justin M. Rao. 2011. "The Good News–Bad News Effect: Asymmetric Processing of Objective Information about Yourself." *American Economic Journal: Microeconomics* 3(2): 114–38.

Eisenbach, Thomas M., and Martin C. Schmalz. 2015. "Anxiety, Overconfidence, and Excessive Risk-Taking." Staff Report 811, Federal Reserve Bank of New York.

Epstein, Larry G. 2008. "Living with Risk." *Review of Economic Studies* 75(4): 1121–41.

Falk, Armin, and Florian Zimmermann. 2011. "Preferences for Consistency." IZA Discussion Paper 5840, Institute for the Study of Labor.

Festinger, Leon. 1957. *A Theory of Cognitive Dissonance.* Stanford University Press.

Fiske, Alan Page, and Philip E. Tetlock. 1997. "Taboo Trade-offs: Reaction to Transactions that Transgress the Spheres of Justice." *Political Psychology* 18(2): 255–97.

Fong, Christina M. 2001. "Social Preferences, Self-Interest, and the Demand for Redistribution." *Journal of Public Economics* 82(2): 225–46.

Frank, Douglas H., Klaus Wertenbroch, and William W. Maddux. 2015. "Performance Pay or Redistribution? Cultural Differences in Just-World Beliefs and Preferences for Wage Inequality." *Organizational Behavior and Human Decision Processes* 130: 160–70.

Frederick, Shane. 2005. "Cognitive Reflection and Decision Making." *Journal of Economic Perspectives* 19(4): 25–42.

Ganguly, Ananda, and Joshua Tasoff. Forthcoming. "Fantasy and Dread: The Demand for Information and the Consumption Utility of the Future." *Management Science.*

Gentzkow, Matthew, and Jesse M. Shapiro. 2011. "Ideological Segregation Online and Offline." *Quarterly Journal of Economics* 126(4): 1799–1839.

Gneezy, Uri, Silvia Saccardo, Marta Serra-Garcia, and Roel van Veldhuizen. 2014. "Motivated Self-Deception and Unethical Behavior." Unpublished paper.

Gottlieb, Daniel. 2010. "Will You Never Learn? Self-Deception and Biases in Information Processing." Unpublished paper, Wharton School, University of Pennsylvania.

Gottlieb, Daniel. 2014. "Imperfect Memory and Choice under Risk." *Games and Economic Behavior* 85: 127–58.

Grossman, Zachary, and Joël van der Weele. Forthcoming. "Self-Image and Willful Ignorance in Social Decisions." *Journal of the European Economic Association.*

Hamman, John R., George Loewenstein, and Roberto A. Weber. 2010. "Self-Interest through Delegation: An Additional Rationale for the Principal–Agent Relationship." *American Economic Review* 100(4): 1826–46.

Herrmann, Benedikt, Christian Thöni, and Simon Gächter. 2008. "Antisocial Punishment across Societies." *Science*, March 7, 319(5868): 1362–67.

Kahan, Dan M. 2013. "Ideology, Motivated Reasoning, and Cognitive Reflection." *Judgment and Decision Making* 8(4): 407–24.

Kahan, Dan M., Ellen Peters, Erica Cantrell Dawson, and Paul Slovic. 2014. "Motivated Numeracy and Enlightened Self-Government." Cultural Cognition Project Working Paper 116, Yale Law School.

Karlsson, Niklas, George Loewenstein, and Duane Seppi. 2009. "The Ostrich Effect: Selective Avoidance of Information." *Journal of Risk and Uncertainty* 38(2): 95–115.

Konow James. 2000. "Fair Shares: Accountability and Cognitive Dissonance in Allocation Decisions." *American Economic Review* 90(2): 1072–91.

Korn, C. W., T. Sharot, H. Walter, H. R. Heekeren, and R. J. Dolan. 2014. "Depression Is Related to an Absence of Optimistically Biased Belief Updating about Future Life Events." *Psychological Medicine* 44: 579–92.

Köszegi, Botond. 2006. "Ego Utility, Overconfidence, and Task Choice." *Journal of the European Economic Association* 4(4): 673–707.

Köszegi, Botond. 2010. "Utility from Anticipation and Personal Equilibrium." *Economic Theory* 44(3): 415–44.

Krueger, Alan B., and Alexandre Mas. 2004. "Strikes, Scabs, and Tread Separations: Labor Strife and the Production of Defective Bridgestone/Firestone Tires." *Journal of Political Economy* 112(2): 253–89.

Kunda, Ziva. 1987. "Motivated Inference: Self-Serving Generation and Evaluation of Causal Theories." *Journal of Personality and Social Psychology* 53(4): 636–47.

Lerner, Melvin J. 1980. *The Belief in a Just World: A Fundamental Delusion.* New York, NY: Plenum Press.

Levy, Raphaël. 2014. "Soothing Politics." *Journal of Public Economics* 120: 126–33.

Levy, Gilat, and Ronny Razin. 2014. "Rituals or Good Works: Social Signaling in Religious Organizations." *Journal of the European Economic Association* 12(5): 1317–60.

Le Yaouanq, Yves. 2015. "Political Values and the Polarization of Beliefs." Unpublished paper, Toulouse School of Economics, October.

Loewenstein, George. 1987. "Anticipation and the Valuation of Delayed Consumption." *Economic Journal* 97(387): 666–84.

Mayraz, Guy. 2011. "Wishful Thinking." Available at SSRN: http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1955644.

Merkle, C., and M. Weber. 2011. "True Overconfidence: The Inability of Rational Information Processing to Account for Apparent Overconfidence." *Organizational Behavior and Human Decision Processes* 116(2): 262–71.

Mijović-Prelec, Danica, and Dražen Prelec. 2010. "Self-Deception as Self-Signaling: A Model and Experimental Evidence." *Philosophical Transactions of the Royal Society,* B, 365: 227–40.

Möbius, Markus M., Muriel Niederle, Paul Niehaus, and Tanya Rosenblat. 2011. "Managing Self-Confidence: Theory and Experimental Evidence." NBER Working Paper 17014.

Monin, Benoît. 2007. "Holier Than Me? Threatening Social Comparison in the Moral Domain." *International Review of Social Psychology* 20(1): 53–68.

NASA. 2003. *Report of Columbia Accident Investigation Board.* http://www.nasa.gov/columbia/home/CAIB_Vol1.html.

Oliver, J. Eric, and Thomas J. Wood. 2014. "Conspiracy Theories and the Paranoid Style(s) of Mass Opinion." *American Journal of Political Science* 58(4): 952–66.

Ortoleva, Pietro, and Erik Snowberg. 2015a. "Overconfidence in Political Behavior." *American Economic Review* 105(2): 504–35.

Ortoleva, Pietro, And Erik Snowbert. 2015b.

"Are Conservatives Overconfident?" *European Journal of Political Economy* 40(Part B): 333–44.

Oster, Emily, Ira Shoulson, and E. Ray Dorsey. 2013. "Optimal Expectations and Limited Medical Testing: Evidence from Huntington Disease." *American Economic Review* 103(2): 804–830.

Puri, Manju, and David T. Robinson. 2007. "Optimism and Economic Choice." *Journal of Financial Economics* 86(1): 71–99.

Quattrone, George A., and Amos Tversky. 1984. "Causal versus Diagnostic Contingencies: On Self-Deception and on the Voter's Illusion." *Journal of Personality and Social Psychology* 46(2): 237–48.

Rabin, Matthew. 1994. "Cognitive Dissonance and Social Change." *Journal of Economic Behavior and Organization* 23(2): 177–94.

Schelling, T. C. 1988. "The Mind as a Consuming Organ." Chap. 15 in *Decision Making: Descriptive, Normative, and Prescriptive Interactions,* edited by David E. Bell, Howard Raiffa, and Amos Tversky. Cambridge University Press.

Schrand, Catherine M., and Sarah L. C. Zechman. 2012. "Executive Overconfidence and the Slippery Slope to Financial Misreporting." *Journal of Accounting and Economics* 53(1–2): 311–29.

Schwardman, Peter, and Joël van der Weele. 2016. "Deception and Self-Deception." Tinbergen Institute Discussion Papers 16-012/I.

Sharot, Tali, Christoph W. Korn, and Raymond Dolan. 2012. "How Unrealistic Optimism is Maintained in the Face of Reality." *Nature Neuroscience* 14(11): 1475–79.

Sharot, Tali, and Neil Garrett. 2016. "Forming Beliefs: Why Valence Matters." *Trends in Cognitive Science* 20(1): 25–33.

Shayo, Moses. 2009. "A Model of Social Identity with an Application to Political Economy: Nation, Class and Redistribution." *American Political Science Review* 103(2): 147–74.

Shiller, Robert J. 2005. *Irrational Exuberance,* 2nd Edition. Princeton University Press.

Smith, Adam. 1759. *The Theory of Moral Sentiments.*

Tenbrunsel, Ann E., and David M. Messick. 2004. "Ethical Fading: The Role of Self-Deception in Unethical Behavior." *Social Justice Research* 17(2): 223–62.

Thompson, Leigh, and George Loewenstein. 1992. "Egocentric Interpretations of Fairness and Interpersonal Conflict." *Organizational Behavior and Human Decision Processes* 51(2): 176–197.

Trivers, Robert. 2011. The Folly of Fools: The Logic of Deceit and Self-Deception in Human Life. Basic Books.

Tversky, Amos, and Daniel Kahneman. 1974. "Judgment under Uncertainty: Heuristics and Biases." *Science,* New Series, 185(4157): 1124–31.

Wiswall, Matthew, and Basit Zafar. 2015. "How Do College Students Respond to Public Information about Earnings?" *Journal of Human Capital* 9(2): 117–169

Van den Steen, Eric. 2004. "Rational Overoptimism (and Other Biases)." *American Economic Review* 94(4): 1141–51.

Van den Steen, Eric. 2010. "On the Origins of Shared Beliefs (and Corporate Culture)." *RAND Journal of Economics* 41(4): 617–48.

Von Hippel, William, and Robert Trivers. 2011. "The Evolution and Psychology of Self-Deception." *Behavioral and Brain Sciences* 34(1): 1–56.

Journal of Economic Perspectives—Volume 30, Number 3—Summer 2016—Pages 165–188

The Preference for Belief Consonance

Russell Golman, George Loewenstein,
Karl Ove Moene, and Luca Zarri

> The great pleasure of conversation, and indeed of society, arises from a certain correspondence of sentiments and opinions, from a certain harmony of minds, which like so many musical instruments coincide and keep time with one another.
>
> —Adam Smith, *The Theory of Moral Sentiments*, 1759

Why are people who hold one set of beliefs so affronted by alternative sets of beliefs—and by the people who hold them? Why don't people take a live-and-let-live attitude toward beliefs that are, after all, invisibly encoded in other people's minds? In this paper, we present evidence that people care fundamentally about what other people believe, and we discuss explanations for why people are made so uncomfortable by the awareness that the beliefs of others differ from their own. This preference for belief consonance (or equivalently, distaste for belief dissonance) has far-ranging implications for economic behavior. It affects who people choose to interact with, what they choose to exchange information about, what media they expose themselves to, and where they choose to live

■ *Russell Golman is Assistant Professor of Social and Decision Sciences, Carnegie Mellon University, Pittsburgh, Pennsylvania. George Loewenstein is the Herbert A. Simon Professor of Economics and Psychology, Carnegie Mellon University, Pittsburgh, Pennsylvania. Karl Ove Moene is Professor of Economic Policy, University of Oslo, Oslo, Norway. Luca Zarri is Associate Professor of Economic Policy, University of Verona, Verona, Italy. Their email addresses are rgolman@andrew.cmu.edu, gl20@andrew.cmu.edu, k.o.moene@econ.uio.no, and luca.zarri@univr.it.*

† For supplementary materials such as appendices, datasets, and author disclosure statements, see the article page at
http://dx.doi.org/10.1257/jep.30.3.165 doi=10.1257/jep.30.3.165

and work. Moreover, when people are aware that their beliefs conflict with those of others, they often try to change other people's beliefs (proselytizing). If unsuccessful in doing so, they sometimes modify their own beliefs to bring them into conformity with those around them. A preference for belief consonance even plays an important role in interpersonal and intergroup conflict, including the deadliest varieties: Much of the conflict in the world is over beliefs—especially of the religious variety—rather than property (Svensson 2013).

Despite its importance for a wide range of economic and noneconomic outcomes, the preference for belief consonance has received relatively little attention from economists. Perhaps the most closely related research in economics examines the importance of identity (for example, Akerlof and Kranton 2000; Bénabou and Tirole 2011). Although it is typically taken for granted that groups will seek uniformity in the beliefs of their members, here we argue that the preference for belief consonance, and the motivational mechanisms that underlie it, provide a plausible explanation for why groups are so threatened by misalignment in the beliefs of their members.

We review the literatures in economics and allied disciplines dealing with the preference for belief consonance and related constructs. Perhaps most strikingly, we review evidence, and discuss possible explanations, for the curious fact that many of the most vicious disputes occur between individuals or groups who share a broad set of beliefs (consider Shiites and Sunnis or Catholics and Protestants) and revolve around differences in beliefs that can seem minor from the perspective of outsiders to the conflict.

Belief Consonance and Allied Concepts

In economics, the concept most closely related to the preference for belief consonance dates back to Adam Smith's (1759) discussion of "fellow-feeling" in the *Theory of Moral Sentiments.* As explicated by Robert Sugden (2002, 2005a), fellow-feeling is a positive sensation that arises when two people's emotional reactions to a common stimulus are aligned and there is common knowledge of the correspondence. As a striking example of fellow-feeling, Sugden (2005a) relates how soldiers who lived through the grim reality of trench warfare in World War I frequently wrote about the intensity of positive feelings of comradeship with their fellow-soldiers, which they believed would have been unlikely to arise in peacetime.

Although belief consonance is similar to fellow-feeling, it is not the same. At the most basic level, fellow-feeling has to do with feelings whereas belief consonance involves beliefs. Thus, if two people who are aware that each is from the opposite political party were to watch a debate together, and each reacted gleefully to the perceived triumph of their own candidate, fellow-feeling interpreted literally would predict that being together would *enhance* the experience of both (as they are both enjoying the debate). A preference for belief consonance would in contrast imply that the experience would be especially *unpleasant* to the extent that they were both aware that the confluence of feeling arose from divergent beliefs and interpretations of the event.

Much of the 20th century research most closely related to the preference for belief consonance was done by sociologists and psychologists. Sociologists coined the term "homophily"—literally, "love of the same"—to refer to people's propensity to associate with and form friendships with similar others. In their classic study of friendship in two urban neighborhoods, Lazarsfeld and Merton (1954) drew a distinction between *status homophily*, which captures the observed tendency of people to associate with other people possessing similar characteristics (such as race, gender, and religion), and *value homophily*, which reflects people's tendency to affiliate with those holding similar values.[1] Lazarsfeld and Merton discuss a range of possible interactions between status and value homophily. For example, people who associate with one another due to status homophily (say, because they belong to the same ethnic group), but who find that they hold different values, may either cease to associate, attempt to hide their differences when they interact, or change their values to bring them into closer conformity with one another. Value homophily is closely related to belief consonance in that differences in values are often closely related to differences in beliefs. Among economists, for example, differences in attitudes toward raising the minimum wage are closely related to differences in beliefs about the consequences of doing so, with causality almost certainly running in both directions.[2]

In psychology, Heider's (1946, 1958) pioneering Balance Theory posits that in human relationships there is a tendency towards balanced states in which the relations between individuals are harmonized. To illustrate, imbalance would arise if persons *A* and *B* liked one another, but *A* liked and *B* disliked person *C*. Heider discussed a range of behavioral reactions people might have in response to the perception of imbalance: for example, 1) avoiding discussion of imbalance-related topics, 2) distancing oneself from the other person either geographically or in terms of the closeness of the relationship, 3) attempting to change the other person's attitudes and/or beliefs, and 4) changing one's own attitudes and/or beliefs.

The idea that conflicting beliefs are important, albeit within an individual rather than across individuals, is also embedded in the once-influential theory of "cognitive dissonance" proposed by the social psychologist Leon Festinger (1957). Cognitive dissonance theory posits that individuals experience discomfort when they become aware that different beliefs they hold are in conflict. Akerlof and Dickens

[1] Psychologists drew a similar distinction in research on groups, distinguishing between diversity in *surface-level* characteristics like race, gender, and ethnicity and diversity in *deep-level* characteristics, such as experiences, preferences, and values (for example, Phillips and Loyd 2006).

[2] Although papers dating back more than a half-century featured value homophily prominently, more recent papers in sociology, as well as papers by economists who have picked up on the concept of homophily (for example, Currarini, Jackson, and Pin 2009, 2010), have mostly focused on status homophily, addressing the propensity, and consequences of, people's tendency to geographically sort and associate on the basis of objective characteristics like race/ethnicity or income. Indicative of this narrow focus, one influential review of the literature on homophily in the *Annual Review of Sociology* devoted only a single paragraph of its 30 pages to value homophily (McPherson, Smith-Lovin, and Cook 2001).

(1982) brought cognitive dissonance theory to economics, formalizing the theory as three propositions: 1) individuals have preferences not only over states of the world, but also over their beliefs about them; 2) individuals have some control over their beliefs; and 3) beliefs, once chosen, persist over time. Akerlof and Dickens apply their model to safety regulation, innovation, advertising, crime, and Social Security legislation. It is only a small extension of cognitive dissonance theory to assume that individuals attempt to maintain the same kind of balance between their own beliefs and the beliefs of those around them as they do between their own different beliefs.

Explaining the Preference for Belief Consonance

Why do people care about what others believe, and why do they prefer for others to believe what they themselves believe? A point of agreement among various explanations in the literature is that belief consonance strengthens a shared identity, whereas conflicting beliefs threaten one's identity, but different scholars have proposed different conceptions of identity leading to different reasons why protecting one's identity is so important.[3]

The first and most prevalent conception of identity is associated with *group membership*. People join, and identify with, groups because of the material, and possibly psychological, benefits that group membership confers. The preference for belief consonance then stems, according to the group membership perspective, from a desire to enhance one's connection to the group. Kahan and colleagues' "cultural cognition" project (for example, Kahan 2010) provides wide-ranging support for the idea that people bring their beliefs into conformity with those around them for (often rational) reasons connected to social identity. The theory of cultural cognition posits that individuals tend to conform their beliefs about disputed matters of fact to group values that define their members' cultural identities. A Republican, for example, might lose friends by expressing a belief that climate change exists or is caused by human activity, a personal cost that would dwarf the benefits they would personally obtain from articulating, and potentially acting on, opposing beliefs. According to this perspective, people want to hold beliefs similar to those of people with whom they want to associate, specifically for the purpose of strengthening their association to those people.

Beliefs formed through motivated reasoning will not necessarily be internally consistent, but the theory of cultural cognition further posits that individuals are motivated to develop some degree of internal consistency. For example, people tend to believe that behaviors they find moral are also socially beneficial (or at least benign) and that behaviors they find immoral are socially harmful. Kahan, Hoffman, and Braman (2009) illustrate this linkage by showing, for example, that conservatives not only condemn homosexuality, drug use, abortion, and, often,

[3] Akerlof and Kranton (2000, 2002, 2005, 2008) introduced the concept of identity to economic analysis, showing its usefulness across a broad spectrum of applications.

pre- or extramarital sex, but also tend to hold strong beliefs about the negative consequences of these behaviors. Whereas logical reasoning should ideally lead from evidence to conclusions (and perhaps to consensus), cultural cognition suggests that people first form their conclusions (in consensus with their in-group) and then interpret existing evidence in a way that bolsters these conclusions.

A second reason why people might want others to have similar views (or, equivalently, to have similar beliefs to others) is because they want to hold certain beliefs, and the presence of other people with different beliefs poses a threat to their own beliefs. In what follows, we will refer to this as the *protected beliefs* account.

Bénabou and Tirole (2011) propose by far the most developed perspective of this type. In their theory, individuals care about their own "deep values" such as moral standards, concern for others, strength of faith, and so on, but are also to some extent uncertain about, and hence are motivated to convince themselves of, their ability to live according to these ideals. Bénabou and Tirole assume that people have imperfect memory, but better memory for their own past behavior than for their own motives. Knowledge of this asymmetric retention leads people to engage in behaviors that are consistent with the self-identity that they want to maintain. In their model, people make investments based on their beliefs in order to remind themselves of what kind of person they are. Such investments might range from free expressions of belief to costly expenditures of time, money, and effort, which demonstrate commitment to a religious, national, cultural, or professional identity. Investments, including behaviors and beliefs, become "protected assets," much as individuals might protect property they own.[4] Encountering another person who behaves differently or who simply expresses discrepant beliefs diminishes the value of these investments and threatens one's view of oneself.

An interesting implication of Bénabou and Tirole's (2011) model is that identity-linked behaviors will be especially prominent when objective information about deep preferences is scarce, as illustrated by the often commented-upon zeal of new converts (for example, religious or political) whose loyalty to a cause has not yet been established, the exaggerated nationalism of the recent immigrant, and the notorious homophobia of people who have doubts about their own heterosexuality (Adams, Wright, and Lohr 1996).

Their model also predicts that when "deviant" behavior by peers (for example, violating norms and taboos) threatens a strong group identity, it may trigger a forceful reaction. Further, a norm violator's behavior has greater impact, the more similar to the group that person was previously thought to be—that is, the more correlated the violator's values had been to the group. Excommunication and apostasy are canonical examples of the harshest moral condemnations and punishments for insiders who threaten a group's valued beliefs.

Bénabou and Tirole (2011) assume that people desire a particular identity because it can be a source of willpower to sustain personal motivation or a source

[4] For a previous but less-developed account of beliefs as assets, see Abelson's (2007) paper aptly titled "Beliefs Are Like Possessions."

of positive anticipatory utility, but in the context of the preference for belief consonance it does not really matter why people want to protect their identity. A somewhat different "protected beliefs" account that we view as plausible in many circumstances is that people simply want to protect belief-related investments of time or money or other sacrifices that they have already made. If a Catholic had for many years been engaging in communion, praying to the Holy Trinity, and contributing money to the church, for example, it could be devastating to discover that a trusted priest had lost faith and converted to another faith. Although the invested resources cannot be recovered, and so by economic logic should be ignored, the preference for belief consonance reflects the fact that although costs are "sunk," it is natural that people would be reluctant to believe that their investments might have been a mistake (for example, Tykocinski, Pittman, and Tuttle 1995). Bénabou and Tirole's anticipatory account would predict that the preference for belief consonance is strongest early in life, when there is ample time to enjoy a desired identity or to make use of identity as a motivational tool. If, on the other hand, an individual is attached to her beliefs because she has made investments based on them, concern about belief consonance should be strongest later in life (when greater belief-based investments have accumulated).

Whether motivated by a desire to protect one's identity, or by the distaste for writing off a belief-based investment, the protected beliefs perspective leads to predictions about who will care about belief consonance, and in what specific situations. It predicts, for example, that people with intermediate levels of confidence in their beliefs, who are likely to be the most insecure about their identity, should have the strongest preference for belief consonance. People who are very confident or very unconfident in their beliefs should be less disturbed by discrepancies in beliefs; the former because their beliefs are unshakable (Babad, Ariav, Rosen, and Salomon 1987; Visser, Krosnick, and Simmons 2003) and the latter because they are already doubting their beliefs (and unlikely to have invested heavily based on such weak beliefs).

Another determinant of the preference for belief consonance is the *credibility* of the person holding the conflicting beliefs. Awareness that an expert has different beliefs from oneself should evoke stronger feelings of discomfort than the same beliefs held by a person whose opinion one doesn't respect. By the same token, people who held the same beliefs in the past but changed them are especially threatening, because the person's new views cannot be attributed to closed-mindedness.

The *reason* the other person has different beliefs from oneself should also matter. If other people believe something different because they do not have access to the same information, it is easier to assume that they would believe the same as oneself if they had access to one's own information. If they have come to different beliefs from the same information, this poses a much greater challenge to one's own interpretation of reality.

Sophistication about the preference for belief consonance can, perhaps paradoxically, actually make it easier to write off differences of opinion. Recognizing that other people do not like to consider opposing viewpoints allows a person to rationalize disagreement over beliefs as the result of the other person's stubborn denial of opposing arguments, without a need to reevaluate one's own view.

The frequency of encounters and the *visibility* of discrepant beliefs should matter, too. For someone with a preference for belief consonance, frequent contact with a person holding discrepant beliefs, or frequent exposure to the beliefs themselves (say, through the media), should tend to lower utility.

The protected beliefs perspective sheds light on a missing link in the group membership perspective. The group membership perspective does not explain *why* group membership imposes pressure to hold similar beliefs, except in situations in which group membership is *defined* on the basis of belief. If group membership is defined based on a social criterion, such as one's attendance at a particular school or one's ethnic group, then group identity provides no explanation for why people in the group would care about holding similar beliefs. The protected beliefs perspective, in contrast, provides an explanation for why it is so important for people in groups to hold similar beliefs: because the presence of other group members with discrepant beliefs forces a reevaluation of one's own core beliefs, and because other group members tend to have many of the properties just discussed—for example, they have access to similar information and one encounters them frequently.

Still, the group membership perspective can help to make sense of phenomena that the protected beliefs account fails to predict. It is, for example, hard to explain solely on the basis of an the protected beliefs perspective the common finding that people (and especially people with more extreme partisan attitudes) tend to overestimate the extremity of "out-group" views on issues such as affirmative action. Indeed, if anything, the protected beliefs perspective predicts the opposite, since considering the existence of large differences in beliefs should lead to questioning one's own beliefs. The group membership perspective, on the other hand, naturally implies that people will caricature out-group members and their beliefs as a way of defining the boundaries of one's own group and distinguishing in-group members from others.

Chambers, Baron, and Inman (2006) conducted two studies focusing on the contentious issues of abortion and politics. Both studies found that partisans tend to exaggerate differences of opinion with their adversaries, especially with regard to value issues they see as central to their own position. Van Boven, Judd, and Sherman (2012) observed similar effects in three studies, one with a nationally representative sample that evaluated candidates Obama and McCain before the 2008 Presidential election and two with samples of university students. Their results provide evidence of "polarization projection," by which they mean that individuals with more extreme partisan attitudes perceive greater polarization than do individuals with more moderate attitudes. Westfall, Van Boven, Chambers, and Judd (2015), drawing on over 30 years of national survey data from the American National Election Study, likewise found that individuals in the United States consistently overestimate polarization between the attitudes of Democrats and Republicans (see also Sherman, Nelson, and Ross 2003).

The protected beliefs and group membership perspectives offer complementary insights into the preference for belief consonance, and there is no real conflict between them. People want to achieve belief consonance both because it cements their connection to groups, because it protects core values and beliefs about the self, and likely because they don't want to write off investments that they made on the basis of their beliefs. Group membership confers independent benefits, and, because belief dissonance threatens cherished beliefs, groups tend to consist of like-minded individuals, so people often adapt their beliefs to fit into their social groups. At the same time, belief consonance reinforces people's cherished beliefs, which motivates people to associate with groups consisting of like-minded individuals and to ostracize those with discrepant beliefs.[5]

Consequences of the Preference for Belief Consonance

How do people respond to the threat or reality of belief dissonance? An economic perspective on the problem might argue that people should follow a cost–benefit approach. If the costs of any possible response exceed the expected benefit, then people should accept the discomfort of belief dissonance. If there are responses for which benefits exceed costs, however, the individual should be motivated to take the most advantageous approach to reducing the disutility of belief dissonance.

Motivated Belief Formation

When people are disturbed by others' discrepant beliefs, one option is to change their own beliefs to conform. This outcome should be especially likely when an individual regularly confronts multiple people who share common beliefs discrepant with his or her own beliefs, and especially when these individuals are relevant to the focal individual (for example, they are in the same social group). Such belief-conformity effects were demonstrated most famously in Asch's (1951) conformity experiments in which an experimental subject was embedded in a group of people who were all asked a basic question (specifically, which one of several lines shown on a screen was longest). Other than the "focal" subject, other group members were confederates who were instructed to give specific answers. When all of the confederates gave a patently wrong answer, many subjects conformed and gave wrong answers themselves. However, all it took was a single other dissenter for most subjects to provide the correct answer.

[5] Bénabou's (2008) analysis of ideology brings together both perspectives on identity. Ideology, according to Bénabou, "designates a system of beliefs that some group collectively upholds and maintains rigidly, even though it involves a substantial degree of reality denial or 'false consciousness'" (p. 322). Bénabou develops a model of ideologies as collectively sustained (and individually rational) distortions in beliefs, and shows how individuals' "subjective mental constructs" interact across agents and with institutions to generate biased perceptions of reality that persist over time and distort public policy.

Wood, Pool, Leck, and Purvis (1996) conducted studies with a group of students who were informed that a majority group of students at their university held a position on an attitude topic that differed from their own position (Study 1), or that a disliked minority group (such as the Ku Klux Klan) had expressed a position consistent with the participant's own positions (Study 2). In Study 1, participants who rated alignment with the majority group as more relevant to their personal identity were more likely to shift their attitudes to agree with the group's. In Study 2, participants who rated differentiation from the derogated minority group as more relevant to their personal identity were more likely to shift their attitudes to disagree with this group. In a follow-up study, Pool, Wood, and Leck (1998) found that participants who wanted to align themselves with a particular group reported lower self-esteem when they discovered that they disagreed with the group, and that individuals who wanted to differentiate themselves from a group reported lower self-esteem when they found that they agreed with the views of the group. Beyond showing these effects on self-esteem, they again found that whenever a source group was rated as highly self-relevant, participants changed their interpretations of questions either to align themselves with the majority group position or to distance themselves from the minority group position. Moreover, when participants were able to adjust their attitudes in the desired direction, they did not report this reduction in self-esteem.

Cohen (2003) conducted four experimental studies with groups of partisan college students aimed at testing the effects of group influence on attitude change. In the absence of information about the position of their own party on an issue, participants based their attitude on policy content and on their own ideological beliefs: Students characterized as liberals, for example, supported a generous welfare policy, while conservatives supported a stringent one. When information about the position of their party was available, however, participants supported the position endorsed by their party (regardless, in the case of welfare policy, of whether the policy was generous or stringent). Interestingly, participants denied having been influenced by the party positions to which they were exposed and claimed that their beliefs were driven purely by policy content. Participants figured out ways to interpret the policies, and their own values, so as to bring them into conformity, and were not aware that they were doing so.

A challenge in studying motivated belief formation is that beliefs are not directly observable, and it can be difficult to distinguish actual motivated belief formation from motivated *reporting* of agreement. That is, subjects could be motivated to say whatever is necessary to fit in, without actually believing it. A few cleverly designed studies, however, provide evidence that motivated belief formation is real and persistent (for a review, see Wood 2000). For example, Higgins and McCann (1984) had 159 subjects reveal their own beliefs to an audience whose beliefs had already been made public, and found that subjects' own impressions and attitudes were not only distorted toward conformity with those of the audience, but were still biased by this interaction two weeks later, when the audience was no longer present.

An important (and perhaps inadequately appreciated) feature of motivated belief formation, including that which is motivated by the desire for belief consonance, is that people do not, in general, simply arrive at the beliefs they are motivated to hold. Rather, they shift toward beliefs they want to hold through a process of sifting through evidence in a selective fashion (for an early investigation see Darley and Gross 1983; for further evidence, see Babcock, Loewenstein, Issacharoff, and Camerer 1995; for a theoretical model of motivated belief formation based on biased interpretation of evidence, see Rabin and Shrag 1999). As a consequence of such biased information processing, groups with opposing values which are presented with identical evidence often end up becoming more polarized in their beliefs: Lord, Ross, and Lepper (1979) provide an experimental demonstration of this phenomenon. Moreover, although one might expect people with greater scientific expertise to process information in an unbiased fashion, research by Kahan (2015) finds, quite to the contrary, that those who measure higher in scientific knowledge/expertise are most likely to hold polarized beliefs which reflect their political and cultural affinities, as if they use their expertise not to reach reasoned judgments, but rather to rationalize their biased processing of evidence.

Proselytizing

Instead of conforming their own beliefs to those around them, individuals might choose the alternative strategy of attempting to change the beliefs of others to conform to their own. People will be more likely to take this course when they believe the prospects for doing so, relative to the investment required, are favorable. This has the natural implication that individuals holding a minority viewpoint in a large group should be less likely to proselytize and more likely to change their own views than those confronting a smaller number of individuals with discrepant beliefs, since changing the views of large numbers of people is likely to be challenging.

Proselytizing can be a risky strategy, however. If one's attempts to persuade others are unsuccessful, a natural inference is that one's own position is inherently unpersuasive and possibly false. A sophisticated individual should take this risk into account before embarking on an attempt to do so. On the other hand, successful proselytizing can provide a powerful "shot in the arm" for those who care deeply about particular beliefs but feel that their beliefs are threatened.

Based on these considerations, we should expect that proselytizing will be especially common both for those who feel confident about their views (and as a result, presumably confident about their prospects of converting others), and for people who care deeply about their beliefs but perceive that they are threatened, to the point where they are willing to embark on a high-risk strategy to bolster them. The empirical literature provides support for both of these predictions.

Supporting the first prediction, Visser, Krosnick, and Simmons (2003) found that individuals who were especially confident of their attitude towards global warming or air pollution (and attached high importance to the issue) were very likely to attend discussions related to that issue and to exert active efforts to persuade others to adopt their views. Supporting the second, three experiments conducted by Gal and Rucker (2010) showed that *shaken* confidence in beliefs tends to increase people's propensity to persuade others. Across three studies, subjects who were made to feel less confident exerted more effort in advocating their beliefs, and were more likely to attempt to persuade others of their beliefs. As the authors note, proselytizing seemed to function as a means for helping less-confident individuals to bolster their views and to resolve their own doubts. In one experiment, they also found that the effect of shaken confidence on advocacy was affected by other people's receptivity to the advocated message: advocacy was more likely when individuals believed that there was a possibility of changing the opinion of another person.

Selective Information-Seeking and Conversational Minefields

While some disagreements are inevitable, people do have some ability to influence the set of views to which they are exposed (Akerlof and Dickens 1982). For example, although one might think that people would want to expose themselves to news sources that would expand their knowledge and insight, research on media bias finds that people prefer to receive information from media sources that are unlikely to challenge their existing beliefs (Gentzkow and Shapiro 2008, 2010). A Pew Research Center (2014) report on political polarization in the American public reveals, perhaps not surprisingly, that there is a strong correlation between the outlets that people name as their main sources of information about news and politics, and their own political views. Forty-seven percent of "consistent conservatives" named Fox News as their main news source about government and politics, and 88 percent reported that they trust Fox News, whereas 50 percent of "consistent liberals" named either NPR, the *New York Times*, CNN, or MSNBC as their main news source (Mitchell, Gottfried, Kiley, and Matsa 2014). People's distaste for having their beliefs challenged creates powerful incentives for media sources not to "rock the boat"—that is provide belief-challenging perspectives—for fear of losing faithful customers who might bail if exposed to unwelcome viewpoints. Indeed, research on ideological slant in news coverage finds that US daily newspapers tend to slant their coverage of stories in a fashion that retains and consolidates their audiences (Gentzkow and Shapiro 2010).

In his insightful treatise *Republic.com 2.0*, Sunstein (2007) hypothesizes that, although the greater diversity of information available online makes it possible in theory to expose oneself to a wide range of diverse perspectives, the actual result is to enable people to expose themselves more selectively to perspectives that accord with, and rarely challenge, their existing views. Sunstein warns against "the risks posed by any situation in which thousands or perhaps millions or even tens of millions of people are mainly listening to louder echoes of their own voices." Consistent with Sunstein's argument, Gentzkow and Shapiro (2011) find that online news consumption is more ideologically segregated than offline news consumption.

Bénabou (2008) argues that the tendency of citizens to engage in ideological denial provides a new rationale for why societies set up (and should set up) commitment devices such as constitutional rights to free speech and independence of the press, which make it more likely that bad news will surface sooner or later, thus decreasing the expected return of investing in denial. Bénabou and Tirole's (2011) identity-based account of the preference for belief consonance predicts that people will be more willing to expose themselves to belief-contradicting media in the short run if they believe that such exposure is inevitable in the long-run.

As discussed earlier, the protected beliefs account of the preference for belief consonance predicts that people should be especially averse to hearing dissonant beliefs espoused by people or news sources that they might otherwise respect. A natural coping mechanism is to lose respect for news outlets or people with whom one disagrees. Thus, it is common to hear conservatives disparage the sources of news that are popular among liberals, like the *New York Times* or NPR, and it is common to hear liberals disparage Fox News.

People don't only get information from the media, however. They also exchange information about their beliefs with friends, acquaintances, and coworkers. In such interactions, the preference for belief consonance creates a dynamic of interpersonal interaction in which people avoid topics they might disagree about, as described by Sugden (2005b, p. 67):

> Different topics are gradually introduced into the conversation, exploiting connections with what has already been said, with the general aim of finding a topic on which the two partners have common opinions or beliefs. If a topic begins to provoke disagreement, it is dropped. Issues on which people are liable to have strong and opposed private feelings are avoided as *conversational minefields*: recall the familiar saying that religion, sex and politics (some people say religion, sex and money) should never be introduced into a conversation [italics ours].

In *Hearing the Other Side*, Mutz (2006) provides evidence that Americans are generally reluctant to discuss political issues, but especially with people who disagree with them. Her research relies on three data sources: the 1992 and 2000 National Election Survey components of the Comparative National Election Project and a 1996 survey funded by the Spencer Foundation. It shows that people appear mainly reluctant to be exposed to oppositional viewpoints in intimate social networks (as compared with loosely connected social networks), as well as when they hold extreme positions (compared to moderates and independents), which would follow naturally from people's distaste for discussing politics with others who disagree with them. Mutz also finds that there is more exposure to disagreement in networks that are nonwhite, low in socioeconomic status, and populated by people low in knowledgeability about politics.

The reluctance to share discrepant beliefs with others can lead to a phenomenon discussed by psychologists and sociologists termed *pluralistic ignorance*, which arises when everyone believes *X*, but everyone believes that everyone besides themselves believes not-*X*. For example, research on campus alcohol consumption finds that college students often mistakenly believe that they are more uncomfortable with campus alcohol practices than the average student (Prentice and Miller 1993). Similarly, Van Boven (2000) finds that many university students publicly espoused what they view as "politically correct" attitudes—for example, supporting affirmative action—that they questioned in private. We would expect pluralistic ignorance to be most likely to occur in cohesive, homogeneous groups, the members of which should be reluctant to "stick their necks out" and share views that they assume are discrepant with those of the majority.

Belief-Driven Clustering

One straightforward implication of the preference for belief consonance is that people should choose to associate with—that is, become friends with, work with, and even have romantic relationships with—others who share their beliefs. In their original paper on homophily, Lazarsfeld, and Merton (1954) provided evidence for such clustering based on an investigation of two small towns, in which liberals disproportionately selected other liberals as close friends, and conservatives did the same. The Pew Research Center (2014) report mentioned earlier found that online clustering in social media space follows a similar pattern: 52 percent of consistent liberals and 66 percent of "consistent conservatives" on Facebook declared that most of their close friends share their own political views. Forty-four percent of "consistent liberals" say they have blocked or defriended someone due to disagreement about politics (Mitchell et al. 2014).

There is considerable evidence that the desire for belief consonance affects who people choose to date and marry. Alford et al. (2011), for example, offer evidence from almost 8,000 US spouses that, while physical and personality attributes fail to show a significant positive correlation across spouses, political attitudes display extremely strong interspousal correlations. The authors examine 28 individual items and find particularly high correlations regarding school prayer, abortion, gay rights, and party affiliation. Liberal wives are much more likely to have liberal husbands, and conservative wives are much more likely to have conservative husbands. The researchers find, further, that the political similarity of spouses derives to a large extent from assortative mating rather than from spousal assimilation or social homogamy (marriage based on characteristics such as socioeconomic status, class, or religion). If the above correlations were the result of assimilation of beliefs, we should expect to observe that similarity increases over the life of the relationship. Instead, the correlations seem to be more or less constant over time: specifically, adding five years to the length of the marriage raises the correlations by .01—a very modest increase compared to the typical levels of correlations in the data (around .60). Huber and Malhotra (2013), using a novel dataset from a national online dating community, conducted an experiment to investigate the influence of (pre-match) political predispositions on people's initial formation of romantic relationships. The two studies show that when choosing from among potential relationship partners, individuals prefer those who have similar political views and levels of political engagement. Their experimental results show that it is political orientations specifically, rather than correlated attributes, that underlies the apparent preference for politically similar dating partners.

The preference for belief consonance can also affect where people choose to locate geographically. In *The Big Sort*, Bishop (2008) provides diverse evidence to document, since the 1970s, a general trend for Americans to sort geographically based on (mainly political) beliefs. For both cultural and policy-related reasons, the United States is unusual among developed countries in terms of the ease with which people relocate geographically (Molloy, Smith, and Wozniak 2011). Combining that with the increasing polarization of politics, the United States has features that contribute to making it a prime location for belief-driven segregation.

The preference for belief consonance also affects the economic associations that workers enter into, and the consequences of these associations. Complete worker ownership is an interesting case. Bhuyan (2007) finds it is often inspired by commonly held values like equality, self-responsibility, and democracy at the workplace. Sharing the same beliefs in such enterprises can be a great advantage, making other collective goals easier to achieve. "When workers share similar values," Craig, Pencavel, Farber, and Krueger (1995, p. 160) conclude, based on their empirical studies of cooperatives in the US plywood industry, "disputes within the producing unit are less likely to occur, monitoring costs tend to be lower, and social sanctions are probably more effective in deterring malfeasance." Other research shows that workers are willing to pay a substantial premium (in the sense of working for lower effective wages) to work in cooperative enterprises (Craig and Pencavel 1992).

Belief-Driven Favoritism and Conflict

A substantial body of research, much from psychology but some from economics, documents the prevalence of intra-group favoritism and outgroup hostility, and, most importantly from the perspective of this paper, the important role played by beliefs in these phenomena. The general pattern of these studies is to divide the subjects into groups by some criteria, which in different studies can be gender, race, or field of study, or just about anything from sports-team loyalty and music preferences to political affiliations. The different groups then perform an exercise designed to measure levels of cooperation or trust both within and between groups. Taken together, these experimental studies support the idea that, with respect to intra-group and inter-group relationships, people care about shared beliefs, and especially beliefs about politics and religion, and that they generally care more about these beliefs than about other potential dimensions of identity.

In one such study, Kranton, Pease, Sanders, and Huettel (2013) divided undergraduate student subjects into groups which (in two conditions) were based either on preferences for poetry and art or on political affiliation. Subjects then allocated resources between themselves and others who were either part or not part of their own group. Subjects were more likely to behave selfishly, and even to destroy resources to deprive others of money, when dealing with a different group, and group membership based on political affiliation produced stronger effects than that based on artistic/poetry preferences.

Iyengar and Westwood (2015) investigate "partisan affective polarization," by which they refer to the tendency of people identifying as Republicans or Democrats to view opposing partisans negatively, and copartisans positively. They conducted a study in which the beliefs of 2,000 adults were measured with an "implicit association test" designed to measure attitudes that people have but are not consciously aware of holding. Positive views of in-group members and negative views of out-group members were evident not only in explicit, but also implicit measures of attitudes. Further, using classic experimental (trust and dictator) games, they found that players acted more pro-socially towards members of their own political party than toward members of the opposing political party. In contrast, they did not observe such a discrepancy in behavior for those in the same or different ethnic groups.

In an experiment that compared the impact of a broad range of group differences, Ben-Ner, McCall, Stephane, and Wang (2009) assigned undergraduate participants to groups based on different criteria. In the first study, they found that, all of the belief-based membership categories (political views, sports-team loyalty, religion, and music preferences) led to greater cooperation than any nonbelief-based categories (such as birth order, dress type, body type, socioeconomic status, and gender) with the sole exception of family ties, which led to greatest cooperation. A follow-up study found that generosity in a dictator game was greatest between those who shared political views, followed by those who shared religious affiliation, nationality, or body type.

Although the most relevant research on the inter- and intra-group consequences of belief consonance and dissonance focuses on relatively mild outcomes, such as allocation of small amounts of money, belief dissonance between groups can have more momentous consequences. When members of groups with conflicting beliefs interact with outgroup members, neither changing one's own beliefs nor proselytizing are likely to be viable strategies for an individual, because the former would produce belief dissonance with their own group, and the latter would fail because those in the other group are, by the same token, unlikely to be persuadable. If the groups cannot move away from one-another, and the constant reminder of the conflicting beliefs is sufficiently threatening, groups may resort to violent conflict to try to limit exposure to the threatening beliefs by seeking to silence the other group, or in some cases even by eliminating their members.

Indeed, much of the conflict in the world is over beliefs, rather than land or property, and especially over religious beliefs. Of all recorded armed conflicts in the world in the period 1975–2010, according to statistics assembled by Svensson (2013), 28 percent had a "religious dimension in the incompatibility." In regions that are more conflict-prone than average, the percent of conflicts revolving around religious incompatibilities is especially high. For instance, in the Middle East and North Africa region, there were 430 conflicts during the 1975–2010 period, and 38 percent of these appeared, at least on the surface, to involve religious incompatibilities (Svensson 2013, table 1).

The Surprising Potency of Small Differences

Some of the most vociferous disagreements occur between people who—at least from an outsider's perspective—would seem to have very similar beliefs. In the studies just cited examining the source of armed conflicts in the world, for example, almost half of these conflicts were between different sects of groups within the same broad religious tradition.

Drawing attention to the nastiness of disputes between people holding nearly identical views, Sigmund Freud referred in *The Taboo of Virginity* (1917 [1991]) to the "narcissism of small differences," commenting that "it is precisely the differences in people who are otherwise alike that form the basis of hostility between them." The sociologist Pierre Bourdieu made a similar point in his treatise *La Distinction* (1979, English translation in 1984, p. 479), observing that "social identity lies in difference, and difference is asserted against what is closest, which represents the greatest threat."

Empirical research from social psychology and anthropology has documented the surprising potency of small differences. In a 1982 overview article in social psychology, Tajfel summarizes the results of three experimental studies that all find evidence for the importance of small differences for intergroup hostility (Turner 1978; Turner, Brown, and Tajfel 1979; Brown, as reported in Brown and Turner 1981). The studies find that groups with similar values display more intergroup discrimination in competitive situations than groups with dissimilar values. They also show that group members are more ready to sacrifice self-interest for the collective benefit of the in-group when they are dealing with outgroups that are more similar to the in-group.

Further evidence of the potency of small differences comes from research by psychologists on "horizontal hostility." In a series of surveys, White and Langer (1999) and White, Schmitt, and Langer (2006) find that members of minority groups express more unfavorable attitudes about members of other minority groups than about members of majority groups. In particular, people express more hostility toward other minority groups when the other minority groups are more mainstream than their own group. The pattern of horizontal hostility is also evident from a study of members of political parties in Greece by White, Schmitt, and Langer (2006). The authors asked eight party members from each of the four main parties to give a 10-point rating for the social traits of honesty, intelligence, fiscal responsibility, and attractiveness of hypothetical candidates from different parties. Again they find strongly negative evaluations of potential members of similar, but more-mainstream, parties.

In real conflicts, the most comprehensive and systematic investigation of the importance of small differences was undertaken by the Dutch anthropologist Anton Blok (1998, 2001), who drew on existing datasets and empirical findings on the basis of which he concluded that "the fiercest battles often take place between people who have a lot in common" (Blok 1998). In the civil wars in the former Yugoslavia, for example, the most severe fighting took place in the regions that had the smallest differences in ethnic and religious composition between groups and the highest incidences of mixed groups and intermarriages (Blok 2001; Hayden 1996). The differences that divide the fighting parties in many other conflicts are also minor: for example, between the Uzbek minority and the Kyrgyz majority in the conflict in Kyrgyzstan; between Indians and Pakistanis in the conflict in Punjab; between the Greeks and the Turks in the conflict in Cyprus; and between Tutsis and Hutus in Rwanda. The historian Gerard Punier (1995) argues, in his book *The Rwanda Crisis*, that the genocide in 1994 happened after a period in which economic and social differences between Hutus and Tutsis had narrowed. He discusses how the two groups had long lived side by side, had been involved in intermarriages, and how they neither have had separate homelands, languages, or religions. In all these conflicts, subtle differences in beliefs are often the major distinguishing feature, and in some cases the only difference, between the fighting parties. Hatred and suspicion based on these belief differences seem to increase in intensity the more similar the groups are on other dimensions.

The protected beliefs perspective helps to explain the surprising potency of small differences. Bénabou and Tirole (2011) conclude on the basis of their model that "discordant actions are threatening to a person's self-concept when the individuals involved are similar to him." The reason is that people recognize the alignment of another person's beliefs with their own as an informative signal about the other person's credibility. If I am confident about my own beliefs, then the observation that another person holds similar beliefs should lead me to perceive the other person as generally credible. This credibility caused by the general confluence of our beliefs is what renders especially threatening any remaining differences in our beliefs.

Conclusion

In this paper, we have sought to accomplish three goals. First, we have drawn attention to the importance of the preference for belief consonance, as well as connections to related topics discussed in economics and allied disciplines. Despite extensive discussion of homophily in economics and other social and behavioral science disciplines, there has been a striking neglect, in these literatures, of phenomena specifically related to belief consonance.

Second, we have reviewed alternative accounts of why people value belief consonance. One account (for example, Kahan 2015) views the preference for belief consonance as derivative of the desire to conform to the beliefs of a group one is, or would like to be, a member of. An alternative protected beliefs account, articulated in greatest detail by Bénabou and Tirole, views the preference for belief consonance as derivative of the desire to protect core values and beliefs about oneself. Although the protected beliefs account generates more, and more-nuanced, predictions about what types of people and situations will result in stronger or weaker preferences for belief consonance, we have noted that each account helps to explain different stylized facts, and we argue that the two explanations should be viewed as complementary rather than as competing.

Third, we have identified and discussed evidence for a wide range of social and economic consequences of the preference for belief consonance, including motivated belief formation, proselytizing, selective information exposure, belief-driven clustering, and belief-driven favoritism and conflict. With the additional assumption that people judge the validity of others' beliefs based on how many other beliefs they share, moreover, it is possible to explain a phenomenon that, to the best of our knowledge, has not been previously explained: why small differences in beliefs cause such great discomfort and so often lead to violent conflict. Although the evidence we review runs the gamut, from laboratory to field and from observational to experimental, none of the experimental evidence comes from field experiments. Given the importance of the phenomenon of the preference for belief consonance, we believe that this should be an important priority for future research.

Although our focus in this paper has been on the emotional and behavioral consequences of differences in *beliefs* between individuals and groups, very similar analysis could apply to differences in values or attitudes. In practice, beliefs, values, and attitudes tend to be very closely aligned. Just as people like to maintain consistency between the different beliefs they hold, people also seek to maintain consonance between their attitudes and beliefs—to hold beliefs that reinforce their attitudes, and attitudes that reinforce their beliefs (Kahan, Hoffman, and Braman 2009).

The economics profession is, of course, not immune from the polarizing effects of the preference for belief consonance. In his famous essay on the "Methodology of Positive Economics," Milton Friedman (1953) optimistically argued that most disputes that seem to be over values are actually over beliefs, which implied to him that "differences in principle can be eliminated by the progress of positive economics." Friedman illustrated his point with the example of minimum-wage legislation, arguing:

> [U]nderneath the welter of arguments offered for and against such legislation there is an underlying consensus on the objective of achieving a 'living wage' for all, to use the ambiguous phrase so common in such discussions. The difference of opinion is largely grounded on an implicit or explicit difference in predictions about the efficacy of this particular means in furthering the agreed-on end.

However, more than 20 years after the influential paper by Card and Krueger (1994) which found that raising the minimum wage in New Jersey increased rather than decreased youth employment in the fast food industry, there has been little convergence in scientific perspective between the sides of the minimum wage debate despite decades of follow-up research. Instead, both sides seem able to rationalize the existing evidence so that it supports their pre-existing beliefs, often in a way that keeps their beliefs consonant with their political allegiances.

Many philosophers and political scientists have commented on the value of openness to a wide range of viewpoints. Himelboim, McCreery, and Smith (2013) point out three examples: "Habermas (1989) assumed that exposure to dissimilar views will benefit the inhabitants of a public sphere by encouraging greater interpersonal deliberation and intrapersonal reflection." Arendt (1968) claimed that exposure to conflicting political views plays an integral role in encouraging "enlarged mentality." In the 19th century, John Stuart Mill (1859 [1956], p. 21) wrote about the lack of contact with opposing viewpoints in this way:

> If the opinion is right, they are deprived of the opportunity of exchanging error for truth; if wrong, they lose what is almost as great a benefit, the clearer perception and livelier impression of truth produced by its collision with error.

The preference for belief consonance undermines these desirable properties of free intellectual exchange. It leads people to interact with other people, and media, who share, and hence tend to reinforce, their existing views.

The political climate in the United States at the time this paper is going to press underlines the importance, and hence the value of studying and understanding the causes and consequences of the preference for belief consonance. Analyzing trends with ten questions designed to measure partisanship,[6] the already cited study by the Pew Research Center (2014) found that the share of Americans with consistently conservative or consistently liberal views increased from 10 percent in 1994 to 21 percent in 2014. In 1994, 40 percent of Republicans were more liberal than the median Democrat and 30 percent of Democrats were more conservative than the median Republican. By 2014, these numbers had shrunk dramatically, to 8 percent and 6 percent.

The Pew report documents not only increasing polarization of views, segregation by views, and selective exposure to media, but also increasing animosity between people holding differing views. In each party, the share of highly negative views of those in the opposing party more than doubled from 1994 to 2014. The consequences of polarization go beyond friendship and politics, and reach areas like labor market discrimination. In one study reported earlier, Iyengar and Westwood (2015) asked 1,021 individuals drawn from the Survey Sampling International Panel to select one of two graduating high school seniors for a scholarship. They were told that an anonymous donor had contributed $30,000 to a scholarship fund, that the selection committee had deadlocked over two finalists, and that they had commissioned a survey to decide the winner. The two candidates differed in academic achievement, and also, depending on experimental condition, one of two characteristics: political affiliation (cued through membership in a partisan extracurricular group) or a racial identity (cued through a stereotypical African American/European American name and membership in an extracurricular group). Approximately 80 percent of Republican and Democrat respondents proposed to award the scholarship to the student who shared their own politics. This difference was much larger than the tendency for European

[6] All ten questions asked respondents to report which of two statements—for example, "Blacks who can't get ahead in this country are mostly responsible for their own condition" versus "Racial discrimination is the main reason why many black people can't get ahead these days"—came closer to their own views.

Americans and African Americans to award the scholarship to members of their own ethnic group.

Belief consonance is not without upsides. For example, it can lead to greater trust and solidarity within organizations and groups, which can be good for solving collective action problems. When people share beliefs and values, there are many things they can do together that would otherwise be impossible. Sharing the same beliefs can enhance collective rationality and democracy together with social and economic equality at the relevant local level. Both participation in governance and equal sharing of the benefits seem to enhance productivity.

At the national level, all this was evident when the small open economies in Scandinavia initiated their process of wage compression and welfare expansion in the 1930s with a shared belief that economic openness was important. With the perception that the entire economy was dependent on foreign demand, it was easier to accept that wages throughout the economy needed to be set at a level that exporting industries could tolerate, and that social insurance was needed to mitigate the consequences of fluctuations in the world market (Barth, Moene, and Willumsen 2014). Sharing the belief that economic openness was decisive, the Scandinavian countries could implement protection without protectionism, which resulted in half of the US wage inequality and twice the US welfare state generosity. These beliefs, and their consequences, still remain. The share of the population that wants protective measures against foreign competition is only 29 percent in Sweden, 35 percent in Denmark and Norway, in contrast to 61 percent in the United States (Melgar, Milgram-Baleix, and Rossi 2013).

At the firm level, in most countries, there are differences across companies in beliefs and values, in part because workers select companies and companies select workers with similar values to their own (Lazear 1995; Van den Steen 2010; Besley and Ghatak 2005). These shared beliefs constitute the culture of the enterprise. The resulting homogeneity within firms reduces differences in objectives, mitigating agency problems and extending the scope for delegation. Thus, there are clear gains of homogeneity of beliefs, although the literature also warns against a possible overinvestment in homogeneity (Van den Steen 2010).

While the preference for belief consonance may make perfect sense for a utility-maximizing individual, and may confer benefits in limited situations, we believe that the larger literature on belief consonance suggests that it is a largely negative force for society as a whole, through its contribution to diverse social ills including intolerance, political polarization and deadlock, and intergroup conflict. Greater tolerance of disagreement might make the world a more productive and hospitable place in which to coexist.

■ *The authors thank Roland Bénabou for comments on an earlier draft of this paper, seminar participants at the Toulouse School of Economics and Institute for Advanced Study, Cass Business School, and the Harvard University Economics and Psychology Departments. We are also grateful for financial support from the Einaudi Institute for Economics and Finance (EIEF).*

References

Abelson, Robert P. 2007. "Beliefs Are Like Possessions." *Journal for the Theory of Social Behavior* 16(3): 223–50.

Adams, Henry E., Lester Wright Jr., and Bethany A. Lohr. 1996. "Is Homophobia Associated with Homosexual Arousal?" *Journal of Abnormal Psychology* 105(3): 440–45.

Akerlof, George A., and William T. Dickens. 1982. "The Economic Consequences of Cognitive Dissonance." *American Economic Review* 72(3): 307–19.

Akerlof, George A., and Rachel E. Kranton. 2000. "Economics and Identity." *Quarterly Journal of Economics* 115(3): 715–53.

Akerlof, George A., and Rachel E. Kranton. 2002. "Identity and Schooling: Some Lessons for the Economics of Education." *Journal of Economic Literature* 40(4): 1167–1201.

Akerlof, George A., and Rachel E. Kranton. 2005. "Identity and the Economics of Organizations." *Journal of Economic Perspectives* 19(1): 9–32.

Akerlof, George A., and Rachel E. Kranton. 2008. "Identity, Supervision, and Work Groups." *American Economic Review* 98(2): 212–17.

Alford, John R., Peter K. Hatemi, John R. Hibbing, Nicholas G. Martin, and Lindon J. Eaves. 2011. "The Politics of Mate Choice." *Journal of Politics* 73(2): 362–79.

Arendt, Hannah. 1968. "Truth and Politics." In *Between Past and Future: Eight Exercises in Political Thought*, by Hannah Arendt. New York: Viking Press.

Asch, Solomon E. 1951. "Effects of Group Pressure on the Modification and Distortion of Judgments." In *Groups, Leadership and Men*, edited by Harold Guetzkow, pp. 177–90. Pittsburgh, PA: Carnegie Press.

Babad, Elisha Y., Ayala Ariav, Ilana Rosen, and Gavriel Salomon. 1987. "Perseverance of Bias as a Function of Debriefing Conditions and Subjects' Confidence." *Social Behaviour* 2(3): 185–93.

Babcock, Linda, George Loewenstein, Samuel Issacharoff, and Colin Camerer. 1995. "Biased Judgments of Fairness in Bargaining." *American Economic Review* 85(5): 1337–43.

Barth, Erling, Karl O. Moene, and Fredrik Willumsen. 2014. "The Scandinavian Model—An Interpretation." *Journal of Public Economics* 117: 60–2.

Bénabou, Roland. 2008. "Ideology." *Journal of the European Economic Association* 6(2–3): 321–52.

Bénabou, Roland, and Jean Tirole. 2011. "Identity, Morals and Taboos: Beliefs as Assets." *Quarterly Journal of Economics* 126(2): 805–55.

Ben-Ner, Avner, Brian P. McCall, Massoud Stephane, and Hua Wang. 2009. "Identity and In-group/Out-group Differentiation in Work and Giving Behaviors: Experimental Evidence." *Journal of Economic Behavior and Organization* 72(1): 153–70.

Besley, Timothy, and Maitreesh Ghatak. 2005. "Competition and Incentives with Motivated Agents." *American Economic Review* 95(3): 616–36.

Bhuyan, Sanjib. 2007. "'People' Factor in Cooperatives: An Analysis of Members' Attitudes and Behavior." *Canadian Journal of Agricultural Economics* 55(3): 275–98.

Bishop, Bill. 2008. *The Big Sort: Why the Clustering of Like-Minded America is Tearing Us Apart.* New York: Mariner Books.

Blok, Anton. 1998. "The Narcissism of Minor Differences." *European Journal of Social Theory* 1(1): 33–56.

Blok, Anton. 2001. *Honour and Violence.* Cambridge: Polity Press.

Bourdieu, Pierre. 1979 [1984]. *Distinction: A Social Critique of the Judgment of Taste.* Translation published 1984. (First Published 1979 in French as *La Distinction, Critique sociale du judgement* by Les Editions de Minuit, Paris.) London: Routledge.

Brown, Rupert J., and John C. Turner. 1981. "Interpersonal and Intergroup Behavior." In *Intergroup Behavior*, edited by John C. Turner and Howard Giles, 33–65. Oxford: Blackwell.

Card, David, and Alan B. Krueger. 1994. "Minimum Wages and Employment: A Case Study of the Fast-Food Industry in New Jersey and Pennsylvania." *American Economic Review* 84(4): 772–93.

Chambers, John R., Robert S. Baron, and Mary L. Inman. 2006. "Misperceptions in Intergroup Conflict: Disagreeing About What We Disagree About." *Psychological Science* 17(1): 38–45.

Cohen, Geoffrey L. 2003. "Party over Policy: The Dominating Impact of Group Influence on Political Beliefs." *Journal of Personality and Social Psychology* 85(5): 808–22.

Craig, Ben, and John Pencavel. 1992. "The Behavior of Worker Cooperatives: The Plywood Companies of the Pacific Northwest." *American Economic Review* 82(5): 1083–1105.

Craig, Ben, John Pencavel, Henry Farber, and Alan Krueger. 1995. "Participation and Productivity: A Comparison of Worker Cooperatives and Conventional Firms in the Plywood Industry." *Brookings Papers on Economic Activity: Microeconomics*, 121–74.

Currarini, Sergio, Matthew O. Jackson, and Paolo Pin. 2009. "An Economic Model of Friendship: Homophily, Minorities, and Segregation." *Econometrica* 77(4): 1003–45.

Currarini, Sergio, Matthew O. Jackson, and Paolo Pin. 2010. "Identifying the Roles of Race-based Choice and Chance in High School Friendship Network Formation." *PNAS* 107: 4857–61.

Darley, John M., and Paget H. Gross. 1983. "A Hypothesis-Confirming Bias in Labeling Effects." *Journal of Personality and Social Psychology* 44(1): 20–33.

Festinger, Leon. 1957. *A Theory of Cognitive Dissonance.* Evanston, Illinois: Row Peterson.

Freud, Sigmund. 1917 [1991]. "The Taboo of Virginity." In *On Sexuality: Three Essays on the Theory of Sexuality and Other Works.* Harmondsworth: Penguin Books.

Friedman, Milton. 1953. "The Methodology of Positive Economics." Part I of *Essays in Positive Economics*, p. 3–43. University of Chicago Press.

Gal, David, and Derek D. Rucker. 2010. "When in Doubt, Shout! Paradoxical Influences of Doubt on Proselytizing." *Psychological Science* 20(11): 1701–07.

Gentzkow, Matthew, and Jesse M. Shapiro. 2008. "Competition and Truth in the Market for News." *Journal of Economic Perspectives* 22(2): 133–54.

Gentzkow, Matthew, and Jesse M. Shapiro. 2010. "What Drives Media Slant? Evidence from U.S. Daily Newspapers." *Econometrica* 78(1): 35–71.

Gentzkow, Matthew, and Jesse M. Shapiro. 2011. "Ideological Segregation Online and Offline." *Quarterly Journal of Economics* 126(4): 1799–1839.

Habermas, Jürgen. 1989. *The Structural Transformation of the Public Sphere: An Inquiry into a Category of Bourgeois Society.* Cambridge MA: MIT Press.

Hayden, Robert M. 1996. "Imagined Communities and Real Victims: Self Determination and Ethnic Cleansing in Yugoslavia." *American Ethnologist* 23(4): 783–801.

Heider, Fritz. 1946. "Attitudes and Cognitive Organization." *Journal of Psychology* 21(1): 107–112.

Heider, Fritz. 1958. *The Psychology of Interpersonal Relations.* New York: J. Wiley and Sons.

Higgins, E. Tory, and C. Douglas McCann. 1984. "Social Encoding and Subsequent Attitudes, Impressions, and Memory: 'Context-Driven' and Motivational Aspects of Processing." *Journal of Personality and Social Psychology* 47(1): 26–39.

Himelboim, Itai, Stephen McCreery, and Marc Smith. 2013. "Birds of a Feather Tweet Together: Integrating Network and Content Analyses to Examine Cross-Ideology Exposure on Twitter." *Journal of Computer Mediated Communication* 18(2): 40–60.

Huber, Gregory, and Neil Malhotra. 2013. "Dimensions of Political Homophily: Isolating Choice Homophily along Political Characteristics." Working Paper 3108, Stanford University.

Iyengar, Shanto, and Sean J. Westwood. 2015. "Fear and Loathing across Party Lines: New Evidence on Group Polarization." *American Journal of Political Science* 59(3): 690–707.

Kahan, Dan. 2010. "Fixing the Communications Failure." *Nature*, January 21, 463(7279): 296–97.

Kahan, Dan M. 2015. "Climate Science Communication and the Measurement Problem." *Political Psychology* 36(S1): 1–43.

Kahan, Dan M., David A. Hoffman, and Donald Braman. 2009. "Whose Eyes Are You Going to Believe? Scott v. Harris and the Perils of Cognitive Illiberalism." *Harvard Law Review* 122(3): 838–906.

Kranton, Rachel, Matthew Pease, Seth Sanders, and Scott Huettel. 2013. "Identity, Group Conflict, and Social Preferences." Unpublished paper.

Lazarsfeld, Paul, and Robert K. Merton. 1954. "Friendship as a Social Process: A Substantive and Methodological Analysis." In *Freedom and Control in Modern Society*, edited by M. Berger, T. Abel, and C. H. Page, pp. 18–66. New York: Van Nostrand.

Lazear, Edward P. 1995. "Corporate Culture and the Diffusion of Values." In *Trends in Business Organization: Do Participation and Cooperation Increase Competitiveness?* edited by Horst Siebert. Tubingen: Mohr.

Lord, Charles G., Lee Ross, and Mark R. Lepper. 1979. "Biased Assimilation and Attitude Polarization: The Effects of Prior Theories on Subsequently Considered Evidence." *Journal of Personality and Social Psychology* 37(11): 2098–2109.

McPherson, Miller, Lynn Smith-Lovin, and James M. Cook. 2001. "Birds of a Feather: Homophily in Social Networks." *Annual Review of Sociology* 27: 415–44.

Melgar, Natalia, Juliette Milgram-Baleix, and Máximo Rossi. 2013. "Explaining Protectionism Support: The Role of Economic Factors." *ISRN Economics,* International Scholarly Research Notice, vol. 2013; article ID 954071. http://dx.doi.org/10.1155/2013/954071.

Mill, John Stuart. 1859[1956]. *On Liberty.* Liberal Arts Press.

Mitchell Amy, Jeffrey Gottfried, Jocelyn Kiley, and Katerina Eva Matsa. 2014. *Political Polarization and Media Habits,* October 21. Pew Research Center.

Molloy, Raven, Cristopher L., Smith, and Abigail Wozniak. 2011. "Internal Migration in the United States." *Journal of Economic Perspectives* 25(3): 173–96.

Mutz, Diana C. 2006. *Hearing the Other Side: Deliberative versus Participatory Democracy.* Cambridge University Press.

Pew Research Center. 2014. *Political Polarization in the American Public: How Increasingly Ideological Uniformity and Partisan Antipathy Affect Politics, Compromise and Everyday Life.* June. http://www.

people-press.org/files/2014/06/6-12-2014-Political-Polarization-Release.pdf.

Phillips, Katherine W., and Denise L. Loyd. 2006. "When Surface and Deep-level Diversity Collide: The Effects on Dissenting Group Members." *Organizational Behavior and Human Decision Processes* 99(2): 143–60.

Pool, Gregory J., Wendy Wood, and Kira Leck. 1998. "The Self-Esteem Motive in Social Influence: Agreement with Valued Majorities and Disagreement with Derogated Minorities." *Journal of Personality and Social Psychology* 75(4): 967–75.

Prentice, Deborah A., and Dale T. Miller. 1993. "Pluralistic Ignorance and Alcohol Use on Campus: Some Consequences of Misperceiving the Social Norm." *Journal of Personality and Social Psychology* 64(2): 243–56.

Punier, Gérard. 1995. *The Rwanda Crisis 1959–1994: History of a Genocide.* London: Hurst.

Rabin, Matthew, and Joel L. Schrag. 1999. "First Impressions Matter: A Model of Confirmatory Bias." *Quarterly Journal of Economics* 114(1): 37–82.

Sherman, David K., Leif D. Nelson, and Lee D. Ross. 2003. "Naïve Realism and Affirmative Action: Adversaries Are More Similar Than They Think." *Basic and Applied Social Psychology* 25(4): 275–89.

Smith, Adam. 1759[1976] *The Theory of Moral Sentiments.* Oxford University Press.

Sugden, Robert. 2002. "Beyond Sympathy and Empathy: Adam Smith's Concept of Fellow-feeling." *Economics and Philosophy* 18(1): 63–87.

Sugden, Robert. 2005a. "Correspondence of Sentiments: An Explanation of the Pleasure of Mutual Interaction." In *Economics and Happiness: Framing the Analysis*, edited by Luigino Bruni and Pier Luigi Porta. Oxford University Press.

Sugden, Robert. 2005b. "Fellow Feeling." In *Economics and Social Interaction: Accounting for Interpersonal Relations*, edited by Benedetto Gui and Robert Sugden. Cambridge University Press.

Sunstein, Cass R. 2007. *Republic.com 2.0.* Princeton University Press.

Svensson, Isak. 2013. "One God, Many Wars: Religious Dimensions of Armed Conflict in the Middle East and North Africa." *Civil Wars* 15(4): 411–30.

Tajfel, Henri. 1982. "Social Psychology of Intergroup Relations." *Annual Review of Psychology* 33: 1–39.

Turner, John C. 1978. "Social Comparison, Similarity, and Ingroup Favoritism." In *Differentiation Between Social Groups: Studies in the Social Psychology of Intergroup Relations*, edited by Henri Tajfel, 235–50. London: Academic.

Turner, John C., Brown, Rupert J., and Henri Tajfel. 1979. "Social Comparison and Group Interest in Ingroup Favoritism." *European Journal of Social Psychology* 9(2): 187-204.

Tykocinski, Orit E., Thane S. Pittman, and Erin E. Tuttle. 1995. "Inaction Inertia: Foregoing Future Benefits as a Result of an Initial Failure to Act." *Journal of Personality and Social Psychology* 68(5): 793–803.

Van Boven, Leaf. 2000. "Pluralistic Ignorance and Political Correctness: The Case of Affirmative Action." *Political Psychology* 21(2): 267–76.

Van Boven, Leaf, Charles M. Judd, and David K. Sherman. 2012. "Political Polarization Projection: Social Projection of Partisan Attitude Extremity and Attitudinal Processes." *Journal of Personality and Social Psychology* 103(1): 84–100.

Van den Steen, Eric. 2010. "On the Origin of Shared Beliefs (and Corporate Culture)." *Rand Journal of Economics* 41(4): 617–48.

Visser, Penny S., Jon A. Krosnick, and Joseph P. Simmons. 2003. "Distinguishing the Cognitive and Behavioral Consequences of Attitude Importance and Certainty: A New Approach to Testing the Common-Factor Hypothesis." *Journal of Experimental Social Psychology* 39: 118–41.

Westfall, Jacob, Leaf Van Boven, John R. Chambers, and Charles M. Judd. 2015. "Perceiving Political Polarization in the United States: Party Identity Strength and Attitude Extremity Exacerbate the Perceived Partisan Divide." *Perspectives on Psychological Science* 10(2): 145–58.

White, Judith B., and Ellen J. Langer. 1999. "Horizontal Hostility: Relations between Similar Minority Groups." *Journal of Social Issues* 55(3): 537–59.

White, Judith B., Michael T. Schmitt, and Ellen J. Langer. 2006. "Horizontal Hostility: Multiple Minority Groups and Differentiation from the Mainstream." *Group Processes and Intergroup Relations* 9(3): 339–58.

Wood, Wendy. 2000. "Attitude Change: Persuasion and Social Influence." *Annual Review of Psychology* 51: 539–70.

Wood, Wendy, Gregory J. Pool, Kira Leck, and Daniel Purvis. 1996. "Self-Definition, Defensive Processing, and Influence: The Normative Impact of Majority and Minority Groups." *Journal of Personality and Social Psychology* 71(6): 1181–93.

Journal of Economic Perspectives—Volume 30, Number 3—Summer 2016—Pages 189–212

Motivated Bayesians: Feeling Moral While Acting Egoistically

Francesca Gino, Michael I. Norton, and Roberto A. Weber

A growing body of research yields ample evidence that individuals' behavior often reflects an apparent concern for moral considerations. Using a broad definition of morality—to include varied non-egoistic motivations such as fairness, honesty, and efficiency as possible notions of "right" and "good"—economic research indicates that people's behavior often reflects such motives (Fehr and Schmidt 2006; Abeler, Becker, and Falk 2014). Perhaps this should not come as a surprise to economists, given that Adam Smith prominently highlighted such motivations in *The Theory of Moral Sentiments* in 1759—17 years before *The Wealth of Nations.*

A natural way to interpret evidence of such motives using an economic framework is to add an argument to the utility function such that agents obtain utility both from outcomes that yield only personal benefits and from acting kindly, honestly, or according to some other notion of "right" (Andreoni 1990; Fehr and Schmidt 1999; Gibson, Tanner, and Wagner 2013). Indeed, such interpretations can account for much of the existing empirical evidence. However, a growing body of research at the intersection of psychology and economics produces findings inconsistent with such straightforward, preference-based interpretations for moral behavior. In particular, while people are often willing to take a moral act that imposes personal

■ *Francesca Gino is the Tandon Family Professor of Business Administration, Harvard Business School, Boston, Massachusetts. Michael I. Norton is the Harold M. Brierley Professor of Business Administration, Harvard Business School, Boston, Massachusetts. Roberto A. Weber is Professor of Economics, University of Zurich, Zurich, Switzerland. Their email addresses are fgino@hbs.edu, mnorton@hbs.edu, and roberto.weber@econ.uzh.ch.*

† For supplementary materials such as appendices, datasets, and author disclosure statements, see the article page at
http://dx.doi.org/10.1257/jep.30.3.189 doi=10.1257/jep.30.3.189

material costs when confronted with a clear-cut choice between "right" and "wrong," such decisions often seem to be dramatically influenced by the specific contexts in which they occur. In particular, when the context provides sufficient flexibility to allow plausible justification that one can both act egoistically while remaining moral, people seize on such opportunities to prioritize self-interest at the expense of morality. In other words, people who appear to exhibit a preference for *being* moral may in fact be placing a value on *feeling* moral, often accomplishing this goal by manipulating the manner in which they process information to justify taking egoistic actions while maintaining this feeling of morality.

As an example of how such motivated beliefs help people easily reinterpret their egoistic behavior, consider Fritz Sander, a German engineer employed by Topf & Sons during World War II, whose work included designing—and attempting to patent—more efficient incineration devices for use in Nazi concentration camps.[1] Following the war, he justified his actions as morally consistent with his professional obligations: "I was a German engineer and key member of the Topf works, and I saw it as my duty to apply my specialist knowledge in this way to help Germany win the war, just as an aircraft construction engineer builds airplanes in wartime, which are also connected with the destruction of human beings" (as quoted by Fleming 1993). Such justifications abound in less-extreme cases as well. Enron chief executive officer Jeffrey Skilling (2002 [2011]), following the firm's bankruptcy and his own convictions for conspiracy and fraud, testified: "I was immensely proud of what we accomplished. We believed that we were changing an industry, creating jobs, helping resuscitate a stagnant energy sector, and … trying to save consumers and small businesses billions of dollars each year. We believed fiercely in what we were doing." Skilling proceeded to testify that he was "not aware of any inappropriate financing arrangements" and that he had left the company "solvent and highly profitable." Of course, one could ask whether Sanders and Skilling were simply lying—that is, knowingly attempting to exculpate their misdeeds by arguing that they were the product of nobler motives. Research on self-serving approaches to morality, however, suggests that they may have processed information and facts in a biased way, allowing them to feel that their questionable behavior was morally justifiable.

In this paper, we will argue that there is a widespread tendency for individuals to exploit justifications and uncertainties present in decision-making environments in order to act egoistically—and, possibly, dishonestly or unethically—without feeling that what they are doing is "bad." That is, people often appear concerned less with the morality of their actions or the outcomes they produce, and more with what the actions they take reveal about them as moral beings. People want to believe they are moral, and prefer actions that support this belief—sometimes independently of whether those actions are themselves actually moral. To facilitate this belief, people often acquire and process information about what is "moral" or "immoral"

[1] See "Topf & Songs as Partners of the SS—The Patent Application" (http://www.topfundsoehne.de/cms-www/index.php?id=120&l=1).

in self-serving ways—and these biased beliefs, rather than a preference for morality itself, may drive much human behavior in contexts involving morality. In decisions involving morality, we argue that people often act as "motivated Bayesians"—while they gather and process information before and during the decision-making process, they tend to do so in a way that is predictably biased toward helping them to feel that their behavior is moral, honest, or fair, while still pursuing their self-interest. Hence, while classical Bayesians will both seek out the most informative evidence and process it in an unbiased way, motivated Bayesians will also be influenced by the evidence that they encounter but will be biased both in choosing which information to acquire and in their interpretation of such information in order to facilitate beliefs in their own morality.

We begin by describing psychological research on motivated reasoning, the domain-general process by which people's goals and emotions influence the manner in which they collect and evaluate information during decision making. We then discuss two ways in which people act as motivated Bayesians when faced with moral decisions, each having the property that people interpret evidence self-servingly to facilitate egoistic behavior at the expense of some moral concern: self-serving judgments of morality and self-serving interpretations of reality. First, we argue that people often form self-serving judgments of what, exactly, constitutes fair or moral behavior or outcomes. When there is some flexibility in interpreting what is "right" and "wrong" or "moral" or "immoral," people's judgments of the morality of an act are often biased in the direction of what best suits their interests. Second, we argue that a similar but distinct phenomenon occurs when people actually alter their judgments of objective qualities—such as their own abilities or the quality of competing options—as a way of making egoistic behavior appear more moral. Finally, we argue that motivated Bayesian reasoning in moral decision making has important implications for many behaviors relevant for economics and policy. In domains including charitable giving, corruption and bribery, and discrimination in labor markets, the ability of people to pursue egoistic objectives while maintaining a belief in their own morality has important consequences for their behavior.

Motivated Reasoning and Motivated Bayesians

Decades of research in psychology shows that people care about their self-concept and expend a great deal of effort maintaining a positive image of the self, often by engaging in motivated reasoning (Steele 1988; Kunda 1990). Kunda (1987), for example, shows that people's explanations for the successes of others tend to reflect favorably on themselves: when asked to indicate the extent to which several factors had contributed to the success of a target person's marriage, participants rated attributes that they personally possessed (such as being the youngest child, or having an employed mother) as more important than characteristics they did not possess. In other words, participants' templates of success in marriage were self-serving. Similarly, people are quick to attribute their successes to their own qualities

("I got an A because I am smart") but their failures to situational factors ("I got an F because the professor is an idiot") (Weiner 1985).

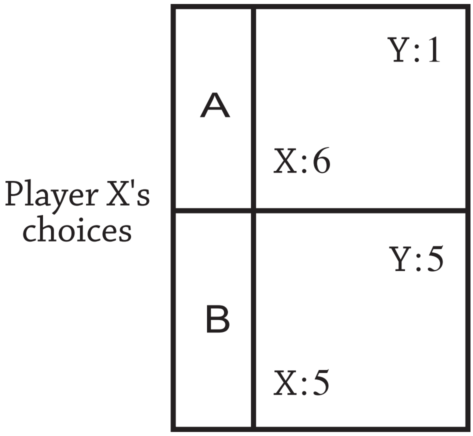
However, a crucial aspect of such motivated reasoning is that this ability to manipulate is not without limit. As noted by Kunda (1990, p. 480), people reach the conclusions they want to reach, "but their ability to do so is constrained by their ability to construct seemingly reasonable justifications for these conclusions." In short, people cannot simply believe anything they want to believe, but are instead—at least in part—constrained by the evidence they encounter and the conclusions that might plausibly be supported by such evidence.

We use the term "motivated Bayesian" to describe this general type of biased information processing. In textbook Bayesian reasoning encountered in introductory statistics courses, people have probability distributions of prior beliefs and then update these beliefs with an unbiased evaluation of any new evidence they encounter. Motivated Bayesians bias this process, for example, by ignoring or underweighting unfavorable evidence or by manipulating the inferences that they draw from the evidence.[2] For example, in a choice context involving morality, a motivated Bayesian has prior beliefs about her own moral qualities. Making, say, an egoistic choice at the expense of some moral objective or obligation should lead an unbiased Bayesian to update (in this case, by downgrading) her beliefs about her moral qualities. However, a motivated Bayesian confronting such a choice will manipulate the information she acquires and how she processes that information in order to reach the conclusion that her egoistic behavior is, in fact, not reflective of immorality. That is, people can be quite creative at manipulating their perceptions of a situation in order to make egoism appear "not that bad" from a moral perspective.

As a specific example, consider the situation analyzed by Batson, Kobrynowicz, Dinnerstein, Kampf, and Wilson's (1997) study, in which participants in a laboratory experiment distribute two tasks between themselves and another participant: a positive task (where correct responses to a task earn tickets to a raffle) and a negative task (not incentivized and described as "rather dull and boring"). Participants were informed: "Most participants feel that giving both people an equal chance—by, for example, flipping a coin—is the fairest way to assign themselves and the other participant to the tasks (we have provided a coin for you to flip if you wish). But the decision is entirely up to you." Half of participants simply assigned the tasks without flipping the coin; among these participants, 90 percent assigned themselves to the positive task. However, the more interesting finding is that among the half of participants who chose to flip the coin, 90 percent "somehow" ended up with the positive task—despite the distribution of probabilities that one would expect from a two-sided coin. Moreover, participants who flipped the coin rated their actions as

[2] This description falls within the more general perspective of treating people as "quasi-Bayesians" in behavioral economic theory (Camerer and Thaler 2003). Under this modeling approach, people make a few systematic mistakes in how they process information, but otherwise employ Bayesian inference procedures. An example of motivated quasi-Bayesian information processing that shares features with the processes we describe is provided by Rabin and Schrag's (1999) model of "confirmatory bias."

Figure 1
Baseline Game from Dana, Weber, and Kuang (2007)



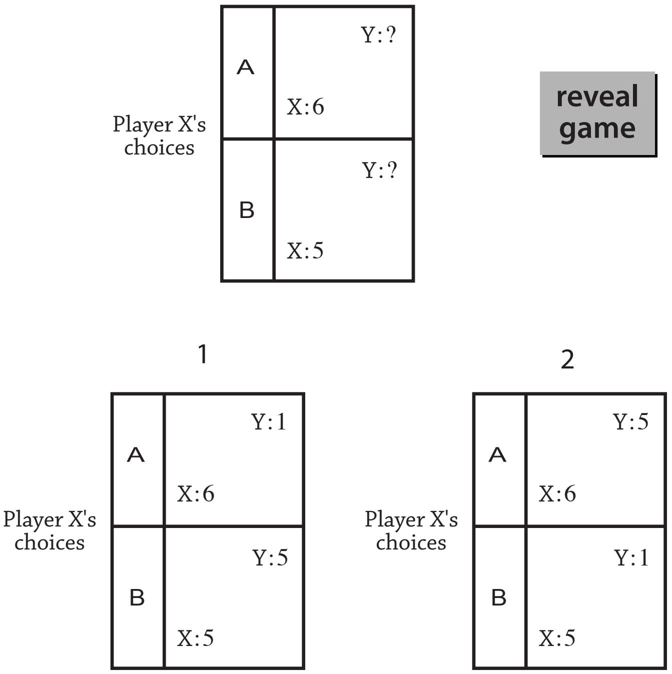
Source: Dana, Weber, and Kuang (2007, Figure 1: Interface for baseline treatment).

more moral than those who did not—even though they had ultimately acted just as egoistically as those who did not flip in assigning themselves the positive task. These results suggest that people can view their actions as moral by providing evidence to themselves that they are fair (through the deployment of a theoretically unbiased coin flip), even when they then ignore the outcome of that coin flip to benefit themselves. Follow-up research on children aged 6 to 11 suggests that this pattern of behavior has a developmental trend (Shaw et al. 2014). As children get older, they remain just as likely to assign themselves to the positive task: what changes with age is their likelihood of flipping the coin—that is, of attempting to gather evidence of their morality rather than actually behaving morally. Hence, consistent with motivated Bayesian reasoning, the act of flipping the coin—perhaps enough times to produce a favorable outcome—seems to provide sufficient evidence to decision makers that their egoistic behavior is in fact consistent with moral behavior.

As another example, consider the decision illustrated in Figure 1, drawn from Dana, Weber, and Kuang (2007). Choice *A* can be viewed as the egoistic act, because it gives the decision maker (*X*) $6 instead of $5 but the other person (*Y*) $1 instead of $5. Choice *B* arguably incorporates other considerations—such as equality and total welfare—that can be interpreted as moral. When confronted with this decision, 74 percent of participants in an experiment using monetary incentives selected the latter option, essentially giving up $1 in order to act in concordance with some apparent moral consideration.

But consider the seemingly similar decision in Figure 2, also from Dana, Weber, and Kuang (2007). In this case, the decision maker again faces a choice between options that offer more (*A*) or less (*B*) money. But, now, the consequences for the other party are unknown, as reflected by the "?" symbol representing unknown payoffs. There are two possible—and equally likely—states of the world,

Figure 2
Hidden Information Game from Dana, Weber and Kuang (2007)



Source: Dana, Weber, and Kuang (2007, Figure 2: Interface for hidden information treatment).

depicted at the bottom of the figure, each with a different set of payoffs that might result from the decision maker's actions. In the first case, the payoffs are identical to those in the baseline case in Figure 1—choosing option *A* rewards the decision maker but harms the other party. But in the other case, acting egoistically also benefits the other person and yields the highest total earnings. That is, in the second case being egoistic is also being moral. Importantly, all decision makers had to do was to click a button ("reveal game") in order to find out the true payoffs. In roughly 50 percent of cases, actual payoffs were identical to those in the baseline, and morally motivated individuals in these cases could have easily acquired the information necessary to sacrifice self-interest for the sake of the greater good.

Despite the preponderance of individuals being willing to choose a more equal distribution with larger social payoffs in the baseline experiment, only 37 percent of participants did so in the hidden-information case in which the payoffs were identical to those in the baseline. That is, even though the resulting outcomes and the ability of individuals to implement those outcomes were identical, simply starting people off in a state of ignorance about the consequences of their actions diminished, by half, the frequency with which people sacrificed personal wealth in pursuit of a moral objective. Moreover, roughly half of the decision makers did not bother to click the button and acquire the payoff information. In this context, decision makers appear to treat ignorance—even if it is a self-imposed absence of evidence that could easily be eliminated—as an excuse for acting egoistically. As motivated Bayesians, participants treat an action taken under willful ignorance as less indicative of an underlying egoistic motivation. This general result was subsequently replicated in other studies (Larson and Capra 2009; Matthey and Regner 2011; Grossman and van der Weele 2013; Feiler 2014).

Importantly in the Dana, Weber, and Kuang (2007) example, people cannot simply convince themselves that choosing ($6, $1) over ($5, $5) is the "moral" thing to do—otherwise, most people would do it in the baseline situation. Instead, they require some "wiggle-room" to reach the desired conclusion, provided by an informational default that allows the perception that choosing ($6, $?) over ($5, $?)—even when in a state of self-imposed ignorance—is not that bad. This is possible when they have the ability to interpret their own behavior favorably, in the manner of a motivated Bayesian.[3]

The research we review next provides further insights into the processes by which people manipulate their perceptions—of what is fair, of the likely outcomes of random processes, of perceived quality, and even of their own abilities—when doing so allows them to maintain a positive moral image while also garnering more personally desirable outcomes.

Self-Serving Judgments of Morality

One way in which people engage in motivated evaluations of the morality of their own behavior is through a flexible construction of beliefs regarding what is "moral." For example, as we describe below, people may be more psychologically comfortable stating something that is not factually true when it is more likely that it *could* have been true (Schweitzer and Hsee 2002; Shalvi, Dana, Handgraaf, and De Drue 2011), or when a lie also benefits someone else (Wiltermuth 2011; Gino, Ayal, and Ariely 2013). A motivated Bayesian can interpret the moral implications of a lie self-servingly, thereby making it easier to act dishonestly. Motivated judgments can also influence perceptions of what is "fair" or "just." In many contexts, it is not straightforward to conclude how much one person deserves relative to another. In such cases, people often interpret the evidence regarding fairness and justice in self-serving ways—by evaluating what is moral through the lens of what also happens to be most personally rewarding. Our goal in this section is to show that judgments of what, precisely, is moral often possess some flexibility and that a motivated Bayesian

[3] Such "self-signaling" can be captured by models in which individuals do not have complete access to their underlying moral motivations when making moral judgments about their own behavior. Instead they draw inferences about their own motivations through actions they observe themselves taking (Bénabou and Tirole 2011; Grossman and van der Weele forthcoming).

can rely on such flexibility to pursue egoistic objectives while maintaining the feeling of adherence to moral standards.

The notion that people are self-serving in how they form judgments of justice is nicely illustrated in experiments on pre-trial bargaining (Babcock, Loewenstein, Issacharoff, and Camerer 1995; Babcock and Loewenstein 1997). In one study, law students were given a civil tort case and assigned to litigate one side of the case. After reviewing the case information, they provided estimates of the award actually granted by a judge in the case, with monetary incentives for accuracy. They also provided assessments of what would constitute a fair settlement. The judgments differed dramatically—by about 50 percent of the average settlement amount—between those "lawyers" assigned to argue the plaintiff's case versus those assigned to represent the defense. Importantly for our argument that people are motivated Bayesians, the difference was much smaller when people reviewed the case material *before* finding out which side of the case they would represent. That is, being forced to process the evidence and develop initial judgments of what constitutes a fair and unbiased settlement *before* having an incentive to view certain outcomes as more or less fair, subsequently prevented participants from having the flexibility to interpret the evidence as supporting a personally favorable notion of justice.

Other studies show that when people can choose among different standards of fairness, very little information is needed for them to favor the standards that are personally beneficial. For example, consider a situation where two people are working on a joint project, but one person's work has produced \$20 and another has produced \$10. Now suppose that one person decides unilaterally how to divide the total \$30 earned by the pair. One could divide the total \$30 either with an equitable division rule (\$15, \$15) or with a meritocratic one that allocates rewards according to account inputs (\$20, \$10). Either has some justification as a "moral" or "just" way to divide jointly produced rewards. In several studies, many people appear to identify the fair distribution of rewards in such cases as the one that best suits their financial interests (Frohlich, Oppenheimer, and Kurki 2004; Messick and Sentis 1979; Konow 2000).

As a concrete example of how motivated Bayesians construct self-serving judgments of what is just or fair, Rodriguez-Lara and Moreno-Garrido (2012) had pairs of participants answer quiz questions, which yielded a shared reward based on the number of correct answers provided by the pair. Importantly, the productivity of individuals' answers, in terms of how much they contributed to the reward, varied across participants. For example, in one variant of the experiment, one person generated 150 pesetas for each correct answer, while the other generated 200 pesetas. One participant in each pair was then randomly given discretion over how to allocate the combined "earnings" produced by the pair. Rodriguez-Lara and Moreno-Garrido identified three possible allocation rules that such an allocator might employ based on different judgments of what is "fair." Under an "egalitarian" rule, the proposer and the allocator receive the same amount of money, independent of their individual productivity. Under an "accountability" rule, participants are accountable for what they can actually control, which in this case is the number of correct answers, but not accountable for the randomly determined productivity per answer. Hence, under accountability, participants receive money in proportion to the number of their correct answers. Finally, under a "libertarian" rule, participants receive an allocation equal to the money that they generated on the quiz based on their correct answers and their random productivity. The results provide clear evidence of motivated Bayesian reasoning. When the allocator's productivity was lower than that of the recipient, allocators relied more on the accountability rule and less on the libertarian rule—that is, they were more likely to allocate according to a rule that rewarded correct answers but not the random productivity shocks. However, when allocators were randomly assigned to be the ones whose output generated more revenue, the importance of accountablity and libertarianism was reversed—allocators were more likely to incorporate these random shocks as part of the entitlements in a just reward. Importantly, only 10 percent of the participants in Rodriguez-Lara and Moreno-Garrido's study kept everything for themselves—doing so feels clearly unjust and immoral. But while perhaps acting somewhat more "morally," the remaining 90 percent tended to form self-serving judgments of fairness consistent with motivated Bayesian reasoning. As Konow (2000) shows, such self-serving judgments of fairness can even constrain one's judgments of what is fair when subsequently dividing money among others as a disinterested third party.

Motivated Bayesians can similarly convince themselves that their actions are more moral than purely egotistical behavior when the situation gives them license to do so, even when the resulting outcomes are the same as those obtained through egotistical acts. This is the case in the study, discussed above, by Dana, Weber, and Kuang (2007): people seem to be more comfortable implementing unequal and inefficient outcomes when they can do so under a veil of self-imposed ignorance.

Another way motivated Bayesians can perceive the same egoistic act as more moral is by acting through an intermediary, which seems to diminish perceptions of moral responsibility. In a study by Hamman, Lowenstein, and Weber (2010), participants could either act egoistically, at the expense of another, by making decisions themselves or by selecting someone to make such decisions on their behalf. In one experimental treatment, participants decided unilaterally how to divide $10 with an anonymous and passive recipient—in a repeated version of the well-known "dictator game." In another treatment, participants hired "agents" to make the allocation decisions on their behalf. Importantly, the subject doing the hiring had all the market power, so agents had to compete for employment by trying to implement the level of sharing that those participants desired. When participants made allocation decisions themselves, a slight majority (51 percent) shared a positive sum with the recipient. However, when acting through intermediaries, this proportion declined to 13 percent—driven by the fact that participants sought out those agents willing to share the least on their behalf. Moreover, when asked to evaluate their behavior, decision makers who acted through agents felt less responsible for the unfair outcomes they had produced and perceived them as fairer. Hence, simply being able to hand off their "dirty work" to someone else can make people evaluate their pursuit of egoistic motives as less wrong. Once again, slightly different paths to the same egoistic outcome can seem more moral when accompanied by a superficial justification. Other studies involving intermediaries that reveal conceptually similar results include Drugov, Hamman, and Serra (2013) and Erat (2013).

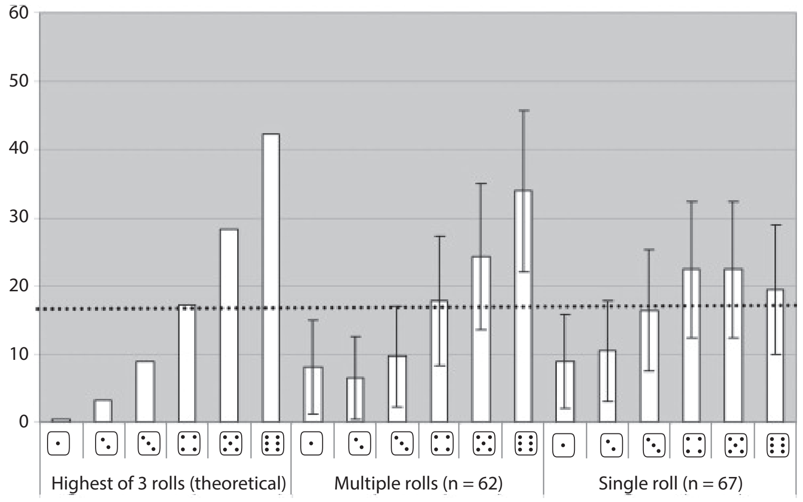
This ability to interpret evidence in a manner favorable to both one's egoism and perceptions of one's morality can be found in contexts beyond those involving sharing and distributing wealth. Many investigations of dishonesty, often led by psychologists interested in morality and behavioral ethics, provide evidence that slightly different paths of behavior toward the same egoistic end can provide individuals with flexibility to favorably interpret the morality of their behavior and the actions that they ultimately take (for examples, see Mazar, Amir, and Ariely 2008; Gino, Norton, and Ariely 2010; Shalvi, Gino, Barkan, and Ayal 2015).

As one example, a study by Shalvi et al. (2011) gave participants the opportunity to lie—by misreporting the outcome of a die roll—in order to obtain more money: higher numbers meant higher payoffs. Hence, an individual could report an outcome of six to obtain the highest possible earnings, and not even the experimenter could identify whether that individual had actually rolled a six. Shalvi et al. either had people roll the die once and report that outcome or, in a "multiple rolls" condition, roll the die three times with the instruction to report only the first roll. Panel 1 of Figure 3 shows the theoretical distribution of reporting the highest of best-of-three rolls. Panel 2 shows the distribution of reported outcomes when participants rolled the die multiple times, while Panel 3 shows the distribution for those who only rolled the die once. People appear to lie more when they roll the die multiple times with instructions to report only the first roll than when they only roll it once. Critically, the distribution of reported die rolls in the multiple rolls case is similar to a best-of-three distribution, suggesting that having observed a favorable die-roll outcome among one of the rolls *that did not count* allowed people to feel more morally justified in reporting that roll as their outcome. That is, if an outcome "could have been true"—in that the individual observed it actually happen—then lying about it seems to provide less-clear evidence of immorality than simply concocting an outcome that was never observed. Rather than treating the counterfactuals as irrelevant, these participants, like motivated Bayesians, incorporate all die roll outcomes as relevant evidence if doing so allows them to win more money by reporting a higher score.

The above examples share a common feature of motivated Bayesian reasoning. The decision maker presumably starts with a belief about his or her own concerns for egoism and morality, and then decides whether to take an action that provides evidence of the strength of these two motives.[4] However, rather than processing this evidence in an unbiased manner, a motivated Bayesian uses the context surrounding the choice to bias the inference drawn from one's own actions. Whether because a motivated Bayesian "did not know" the consequences of actions through willful

[4] For examples of models in which decision makers' actions provide signals of underlying motivations, see Bénabou and Tirole (2006, 2011), Ariely and Norton (2008), and Grossman and van der Weele (2013).

Figure 3
Distributions of Reported Die Rolls



Source: Shalvi, Dana, Handgraaf, and De Dreu (2011).
Note: In an experiment by Shalvi et al. (2011), people were either asked to roll the die once and report that outcome or, in a "multiple rolls" condition, roll the die three times with the instruction to report only the first roll. Higher reported numbers meant higher payoffs. Panel 1 of Figure 3 shows the theoretical distribution of reporting the highest of best-of-three rolls. Panel 2 shows the distribution of reported outcomes when participants rolled the die multiple times, while Panel 3 shows the distribution for those who only rolled the die once

ignorance or because the person was reporting outcomes that "could have" been true, this person, despite acting egoistically, reaches self-serving conclusions that such acts do not reflect a lack of morality.

Self-Serving Interpretations of Reality

A separate and distinct type of motivated Bayesian reasoning involves not changing one's interpretation of the evidence regarding what is fair/unfair or moral/immoral, but instead changing one's perception of the evidence itself in order to arrive at a more positive moral impression of one's behavior. Such self-serving information processing is common in people's evaluations of their own characteristics and abilities, even in contexts that do not involve tradeoffs between egoism and morality. Several studies document that people seek out and attend to information that reinforces the belief that they are better than others in domains such as intelligence and attractiveness, overweighting positive information and underweighting negative information (Mobius, Niederle, Niehaus, and Rosenblat 2011; Eil and Rao 2011). For instance, Quattrone and Tversky (1984) found that people who were told that greater tolerance to immersing one's body in cold water indicated longer (or shorter) longevity subsequently increased (or decreased) the amount of time for which they endured such a task. Hence, a motivated Bayesian who wants to believe in personal longevity may manipulate the evidence that is used in this judgment.

In the domain of moral behavior, people also seem to manipulate beliefs about their own abilities, particularly when doing so makes cheating seem less bad. Consider this statement by disgraced cyclist Lance Armstrong, stripped of his Tour de France victories after a doping scandal: "When you win, you don't examine it very much, except to congratulate yourself. You easily, and wrongly, assume it has something to do with your rare qualities as a person" (Armstrong and Jenkins 2003). Evidence that people misconstrue information about their morally questionable actions to instead provide evidence of their competence is provided by Chance, Norton, Gino, and Ariely (2011). They conducted a series of studies using a paradigm in which participants earned money for answering questions on an IQ test. Some participants took a standard IQ test, while others took the same test but with the answers printed at the bottom—allowing them to "check their work." Not surprisingly, those with the answers at the bottom scored higher on the test and made more money. But the key finding for our purpose occurs when both groups were then shown a second test, which had no answers at the bottom, and were incentivized to predict their performance on that test. An unbiased individual who had used the visible answers on the first test to obtain a higher score would presumably recognize this fact and anticipate lower performance on the second test. However, a motivated Bayesian might instead ignore the presence of the answers—or any effect they may have had on performance on the first test—and instead attribute good performance to personal intelligence, or assume it is driven by what Armstrong called "rare qualities as a person." Consistent with the latter account, people's predictions showed that they disregarded the presence of the answers and instead predicted that they would continue to perform well on the second test, attributing success to their innate "genius" rather than to cheating. Moreover, because payment for performance on the second test was based in part on the accuracy of predictions, these overestimations of performance resulted in motivated Bayesians making less money than people who never had the answers and were not tempted to cheat. When forced to take multiple tests without answers—a process that provided a stream of accurate feedback about their true ability—people were slow to correct their inflated beliefs; but when given another opportunity to cheat and perform well, they were quick to regain their faith in their enhanced abilities (Chance, Gino, Norton, and Ariely 2015).

People also manipulate their beliefs about the likely outcomes of random processes when doing so facilitates egoistic behavior. For example, Haisley and Weber (2010) presented participants with two options. An "other-regarding" option yielded payoffs for the decision maker and for a passive recipient that were relatively equal, for example, \$2.00 and \$1.75, respectively. The "self-interested" option gave the decision maker more money (for example, \$3.00) and gave the recipient a lower payoff involving risk—for example, a lottery paying \$0.50 with $p = 0.5$ and \$0 with $p = 0.5$.

Hence, the self-interested choice was guaranteed to make the recipient worse off, but by how much depended on the outcome of the lottery. The key manipulation in the study was the nature of the lotteries. In a simple-risk condition, the lottery was an objective $p = 0.5$ lottery, where ten red and ten blue chips were placed in a bag and one was drawn at random, with participants free to choose the winning color for the recipient. In an "ambiguous" lottery condition, the composition of the bag was unknown—participants were told that some random combination of red and blue chips had been determined prior to the experiment. Hence, the ambiguous lottery was objectively identical to the lottery involving known simple risk—in both cases there is a 0.5 probability of a ball of each color being selected—but its description created uncertainty about the precise color composition of the bag that would determine outcomes.[5]

The main hypothesis tested by Haisley and Weber (2010) was whether the vague nature of the ambiguous lottery would provide participants the flexibility to manipulate their beliefs about the likely outcome. That is, if participants can convince themselves that the ambiguous lottery is likely to yield a positive payoff with greater probability—since the probability could be anywhere between 0 and 1—then the self-regarding option appears less harmful for the recipient. Indeed, self-interested choices were selected in 73 percent of cases in the ambiguity condition, but only 59 percent of cases under simple risk. Here, the presence of ambiguous consequences for another seems to facilitate egoistic behavior.

Two pieces of evidence from the Haisley and Weber (2010) study particularly suggest a role for motivated Bayesian information processing. First, Haisley and Weber included another treatment dimension to examine whether first inducing participants to express their natural attitudes toward ambiguity, which are typically negative, would subsequently limit their flexibility to interpret ambiguity favorably. In the "constrained" treatment condition, participants started the experiment by choosing which type of lottery they preferred for themselves: one involving simple risk or one involving ambiguity. Consistent with classic evidence of "ambiguity aversion," a large majority of participants preferred the lottery involving simple risk. Importantly, only *after* expressing these attitudes toward ambiguity, did these subjects perform the main choice task, in which they chose whether to take more money for themselves and give the recipient a lottery, which involved either simple risk or ambiguity. Unlike the "unconstrained" participants discussed above, "constrained" participants did not exhibit more frequent self-interested behavior under ambiguity (59 percent) than under simple risk (63 percent). These results show that people who have just expressed an unfavorable view of ambiguity then find it difficult to switch to a favorable view when it becomes convenient to do so.

A second piece of evidence comes from asking participants to estimate the expected value of the payoff to the recipient produced by their choices, with

[5] Having less information about the actual composition of the bag typically induces "ambiguity aversion," whereby the ambiguous lottery is perceived as less desirable (Fox and Tversky 1995; Sarin and Weber 1993).

incentives for accuracy. Participants in the experiment make four choices that potentially affected the payoffs for a recipient. This part of the experiment also included a group of participants who made hypothetical choices, which they knew had no real consequences, so there was no incentive to engage in belief manipulation. Each participant played the game four times, resulting in four choices. Haisley and Weber (2010) calculated the degree to which the different types of participants over- or underestimated the expected value for the recipient resulting from their choices. Figure 4 shows the average estimate bias, cumulative across four choices, for the different groups of participants. The greatest degree of overestimation (by $0.89 across four choices) was demonstrated by "unconstrained" participants making choices under ambiguity; in no other case does ambiguity produce significantly greater overestimation of the value of lotteries, relative to simple risk. Thus, the only group that seems to adopt a strongly favorable view of the likely consequences of their choices is the group that has both an incentive to do so and the flexibility to manipulate their beliefs (having not been recently constrained by stating which kind of lottery they would choose for themselves).

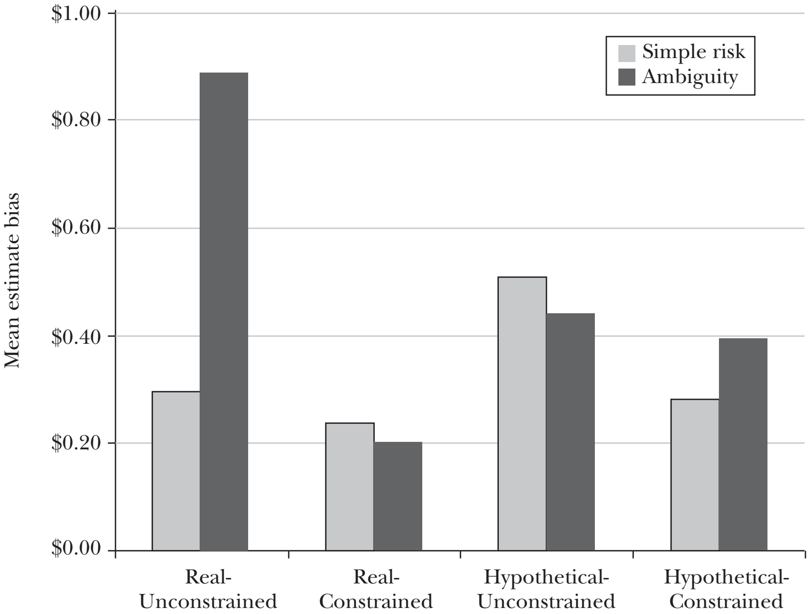
In the study by Haisley and Weber (2010), the choice confronting participants is one in which acting egoistically gives the other participant an unfavorable lottery. Hence, an individual sufficiently concerned with not prioritizing egoism over fairness may find it difficult to take such an action from a moral perspective. However, a convenient opportunity to satisfy both objectives arises if one can reinterpret the evidence to suggest that the unattractive lottery for the other party is, in fact, more attractive than it actually is.

Recent work provides additional evidence of motivated Bayesian reasoning in which people change their beliefs or preferences in order to facilitate egoistic acts. For example, one participant who may benefit by taking money from another may feel better about doing so when the first participant has some reason to feel convinced that the other intends to act unkindly as well (Di Tella, Perez-Truglia, Babino, and Sigman 2015). In the next section, we discuss some additional examples that are particularly relevant for policy questions of interest to economists.

Why the Psychology of Self-Serving Moral Judgments Matters for Economists

As we have shown, self-serving judgments of morality and self-serving interpretations of reality are two common ways in which people act as motivated Bayesians. Much of the pioneering evidence of this phenomenon—and a large part of the existing knowledge—comes from laboratory experiments in psychology, where the idea that people are self-serving in information processing has long been of central interest (Hastorf and Cantril 1954; Festinger 1957). An important question is the extent to which motivated Bayesian reasoning is relevant for the domains that typically interest economists. Below, we discuss several economic contexts in which the kind of motivated reasoning we describe above likely plays an important role.

Figure 4
Overestimation of Consequences for Another



Source: Haisley and Weber (2010).
Note: This experiment involved choosing between two options: one yielding relatively egalitarian payoffs and another yielding more money for the decision maker, less for the other, and making the others' payoff the result of a lottery. The experiment varied whether the lottery involved simple risk (a known 0.5 probability) or ambiguity (a probability anywhere from 0 to 1). In the "constrained" treatment condition, participants started the experiment by choosing which type of lottery they preferred for themselves. The "unconstrained" treatment did not have this component. Some participants made hypothetical choices, which they knew had no real consequences, while others made real choices, which they knew would affect another person. The participants played the game four times, making four choices. The participants were asked to estimate the expected value for the recipient resulting from their choices, with incentives for accuracy. The figure shows the mean estimate bias, cumulative across four choices. See text for details.

Charitable Giving

A natural application of the insights on how motivated Bayesians confront tradeoffs between egoism and sharing wealth is to the domain of charitable giving, which constitutes both a sizeable portion of economic activity and an active area of economic research. Part of the interest among economists lies in understanding why people voluntarily donate to help others—a behavior potentially consistent with a moral motivation such as valuing the well-being of aid recipients or feeling pleasure from the act of giving (Andreoni 1990; Dunn, Aknin, and Norton 2008). However, if people prefer to act selfishly while at the same time believing that they are concerned with fairness and morality—and can employ motivated reasoning to satisfy both objectives—then we might observe them relying on excuses and justifications to avoid making costly charitable donations. Indeed, research suggests that avoiding charitable donation requests is easier for participants than declining the requests once they are made and that, therefore, participants may go out of their way to avoid the request altogether (Flynn and Lake 2008; Lazear, Malmendier, and Weber 2012; DellaVigna, List, and Malmendier 2012; Andreoni, Rao, and Trachtman forthcoming). As with the research reviewed above on willful ignorance, such behavior is consistent with people having some flexibility in how they judge the morality of their actions—and choosing a course of action, when it is available, that yields less giving without a direct challenge to their moral standing.

People may also manipulate their beliefs about the attractiveness of a charitable donation—similarly to the phenomenon observed by Haisley and Weber (2010)—when doing so gives them justifications for acting egoistically. For instance, Exley (2016) examines people given the option to make a donation to a charity, but with some risk that the charity may not receive the donation—as when there is potential waste or corruption. Specifically, she compares situations involving a "self–charity tradeoff," in which people choose between a monetary allocation to be received personally or a monetary allocation to a charity where one of the two allocations involves risk, with other situations involving "no self–charity tradeoff," in which people choose between either a certain amount of money or a risky lottery for themselves, or a certain amount of money or a risky lottery for a charity. By varying the certain amount against which a risky lottery is compared, Exley can observe how much subjects appear to value risky lotteries for themselves or for a charity, and how this is influenced by the presence or absence of a self–charity tradeoff.

Figure 5 shows that when there is no tradeoff between egoism and helping the charity, in the left panel, people treat risk equivalently whether it affects their earnings or those of the charity—that is, they discount the "value" of a given amount of risky money similarly based on the probability that the money might not be received. However, when it comes to decisions involving a tradeoff between egoism and helping the charity, in the right panel, attitudes toward risk diverge considerably. In cases that involve, for example, a choice between keeping money for oneself or giving a risky lottery for the charity, choices reflect a much greater devaluation of lotteries involving risk for the charity than for oneself. In fact, in the right panel, for choices in which one can either give riskless money to the charity or allocate money to a risky lottery for oneself ("self lottery"), people appear to become risk-loving—overvaluing lotteries relative to their expected value—presumably because doing so creates the justification for keeping more money.

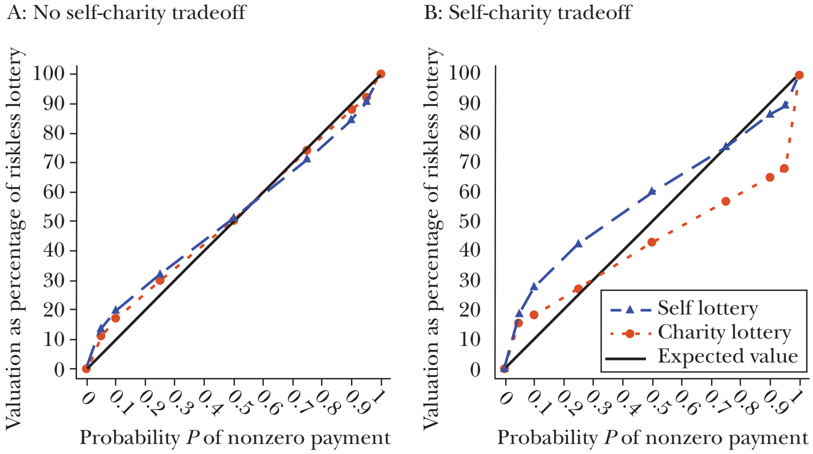
Hence, although participants' donation decisions reflect concern for the charity, when they can justify giving less by altering their attitudes toward risk to make donation relatively less attractive, they do so. Statements such as, "I would donate, but it would just go to waste" or "the charity's overhead is too high," may reflect motivated Bayesian information processing in action, coming up with justifications for not giving.

Discrimination

Another domain in which motivated Bayesians may find creative ways around doing the "right" thing is discrimination. If people are adept at altering the values that they subjectively place on seemingly objective criteria in order to justify ethically

Figure 5

Valuation for Money to Oneself or to a Charity, Based on Risk and on Whether the Decision Involves a Tradeoff between Egoism and Helping the Charity



Source: Exley (2016).

Note: Four situations are compared: 1) a certain monetary allocation or one involving risk, both for oneself (A: Self lottery); 2) a certain monetary allocation or one involving risk, both for a charity (A: Charity lottery); 3) a certain monetary allocation for a charity or one involving risk for oneself (B: Self lottery); and 4) a certain monetary allocation for oneself or one involving risk for a charity (B: Charity lottery). The experiment varies the certain amount against which a risky lottery is compared.

questionable preferences, this may allow them to reach the conclusion that a minority applicant for a position is worse on such "objective" criteria without believing that they themselves are actively discriminating or doing anything morally wrong.

Norton, Vandello, and Darley (2004) capture this alteration of decision criteria directly: men were asked to choose between male and female candidates for a stereotypically male job. Some participants read that the man was better educated but had less experience; others, that he had more experience but less education. Across both conditions, the majority of men selected the male applicant. Most relevant for our account, males claimed that gender played no role in their decision, instead citing education (but only when the male had more education) or experience (but only when the male had more experience) as the basis for their decision. Similar apparent manipulation of preferences is observed in a study by Snyder, Kleck, Strenta, and Mentzer (1979) in which participants chose which of two rooms to sit in to watch a movie. In one room, a person in a wheelchair was also waiting to watch the film; the other room was empty. There were two conditions: the film was either the same in both rooms (offering no excuse to avoid the disabled person) or different (offering a justification for choosing to watch the movie alone). Participants were more likely to choose to watch the movie alone when the two movies were different, presumably because this difference allowed them to claim that the movie in the "solo" room was objectively better—rather than admit to bias against sitting with the handicapped person.

By allowing motivated information processing to influence their perceptions of what constitutes an “attractive” option or candidate, individuals may find it easy to discriminate without believing they are doing so. Therefore, apparent and striking inconsistency between employers’ claims that they do not engage in racial discrimination and their clearly race-based hiring decisions (Pager and Quillian 2005) may seem perfectly justifiable to the motivated Bayesian engaged in such discrimination.[6]

Such motivated Bayesian information processing may also provide an explanation for the finding that the returns to qualifications are lower for employment applicants from minority groups against which there is discrimination and that this can be partially explained by how prospective employers search for information on applicants (Bertrand and Mullainathan 2004; Bartoš, Bauer, Chytilová, and Matejka forthcoming). A motivated Bayesian employer who wants to discriminate, but feels wrong doing so blatantly, may search for reasons to favor a nonminority candidate over one from a minority group. Indeed, in interviews with 55 hiring managers, Pager and Karafin (2009) show that although managers held strong beliefs about the relative performance of black and white employees, they were often unable to generate any instances in their experience to support those impressions, suggesting that rather than updating beliefs with an unbiased evaluation of new evidence, as a classic Bayesian would, these managers were selectively weighting and interpreting information that supported their biased views.

Corruption and Bribery

Situations in which individuals are tempted to accept a bribe or favor a family member for a lucrative appointment also create the ideal conditions for motivated information processing (Hsee 1996). A motivated Bayesian may be quite adept at reaching a conclusion that the familiar candidate is the best qualified or that the vendor offering the highest bribe also offers the best use of public funds. Hence, an official awarding a prestigious sports tournament to a country that has also offered a lucrative personal payment may be able to convince himself that the country is really the most deserving based on “objective” criteria.

The application of this kind of reasoning to corruption is demonstrated by Gneezy, Saccardo, Serra-Garcia, and van Veldhuizen (2016). They use a task in which two participants compete over who can write the best joke (about economists), with the winner receiving a $10 prize. The prize is awarded according to the judgment of a third participant “referee” who picks the winner. The two competing participants can attempt to bribe the referee by sending part of the show-up fee that they receive in cash at the beginning of the experiment to the referee. In a “Before” condition, the referee receives any bribe in the same envelope as the written joke. Therefore,

[6] Such motivated Bayesian “nondiscrimination” can also occur in charitable donations. Fong and Luttmer (2011) find that varying the perceived race of the recipient of a charitable donation does not affect giving directly. However, nonblack donors who are led to believe that recipients are more likely to be black evaluate those recipients as less worthy of aid—for example, by choosing to believe the recipients are more likely responsible for their poverty—and, in turn, give less.

the referee observes the bribe at the same time as opening the envelope to read the jokes. In an "After" condition, the bribes and the jokes are in separate envelopes and the referee sees the bribes only after first reading the jokes. Note that these two versions change very little in terms of the tradeoff between morality and egoism. Someone who wishes to ignore the bribe and simply go with the best joke can do so in either case, which is also true for someone who wishes to simply select the egoistic option and ignore the quality of the jokes. However, a motivated Bayesian's judgments of the quality of the jokes may be swayed by which one is accompanied by the greatest personal benefit. At the same time, a motivated Bayesian who has already read the jokes and formed beliefs about their quality, before learning of the bribes, should find it harder to retroactively convince herself that the joke with the higher bribe is "better."

Consistent with motivated Bayesian reasoning, the timing of knowledge of the bribe appears to affect participants' willingness to be swayed by it. Eighty-four percent of participants in the "Before" condition selected the joke accompanied by the larger bribe, even though only 56 percent of these jokes were rated better by evaluators with no incentive. However, learning of the bribes only after reading the jokes constrains referees' judgments of joke quality: In the "After" condition, a lower proportion (73 percent) selected the joke accompanied by the larger bribe, and a much higher proportion selected the joke rated objectively better (81 percent). Hence, people are unsurprisingly swayed by bribes—but more so when they have the ability to interpret joke quality in a self-serving way.

Attitudes Toward Market Outcomes

Wealthier people often hold less-favorable attitudes toward redistribution (Alesina and Giuliano 2011). For example, Di Tella, Galiani, and Schargrodsky (2007) found that squatters in settlements in Argentina who were exogenously assigned property rights subsequently changed their perceptions of the inherent justice of a free-market system. In particular, these "lucky" individuals were more willing to support statements endorsing the belief that success results from hard work and that money is valuable for happiness. The correlation between personal circumstances and beliefs about the morality of the free-market system and potential resulting inequality might simply reflect self-interest: people express support for those policies that they believe to be most personally rewarding. However, motivated reasoning offers an alternative interpretation. Specifically, if motivated Bayesians can process information in a manner that allows them to reach the conclusion that what is personally rewarding is also that which is moral, then the above relationship may arise without people believing that they are compromising their morality. Instead, they may convince themselves—based on the information to which they attend and that they deem important—that the appropriate notions of fairness and justice are those that also happen to correspond to their own self-interest.

Relatedly, notions of what constitutes fair market wages may reflect self-serving biases and motivated information processing (Babcock, Wang, and Loewenstein

1996). For example, in a study by Paharia, Kassam, Greene, and Bazerman (2009) participants reported being relatively unwilling to hire a domestic worker to clean their house at a below-poverty level wage even when the worker was willing to accept this wage. When the decision was framed as hiring the worker through a placement agency ("Super Cleaners"), however, participants were far more likely to hire the worker. As in the study by Hamman, Loewenstein, and Weber (2010) that we reviewed earlier, inserting a third-party intermediary offers a degree of moral cover for what constitutes a "fair" wage.

Similar self-serving justifications may influence *consumers'* desire for products that raise ethical questions, such as those produced with sweatshop labor or those that may harm the environment. When presented with undesirable products produced with sweatshop labor, participants reported being uninterested in purchasing such unethical products; when products were desirable, on the other hand, purchase interest increased hand in hand with justifications for that increased interest, evidenced by greater agreement with sentiments such as "sweatshops are the only realistic source of income for workers in poorer countries" (Paharia, Vohs, and Deshpandé 2013). Moreover, Ehrich and Irwin (2005) show that people who care about a particular issue—such as the environment—are often *less* likely to seek out product information on that attribute. Because learning about negative environmental impact would constrain purchase, motivated Bayesian consumers avoid the chance of learning in order to allow them to feel good about purchasing behavior. These experiments again show that people motivated by egoistic concerns demonstrate remarkable celerity in using and misusing information to meet self-serving goals while continuing to feel moral.

Conclusion

Economists have developed extensive literatures on topics related to the tradeoffs people make between self-interest and moral considerations such as equality, social welfare, and honesty (Hoffman, McCabe, and Smith 1996; Charness and Rabin 2002; Frey and Meier 2004; Gneezy 2005; Fischbacher and Föllmi-Heusi 2013; Abeler, Becker, and Falk 2014), and have devoted considerable attention to corruption and its potential influence on economic development (Shleifer 2004; Bertrand, Djankov, Hanna, and Mullainathan 2007; Olken 2007). These streams of research have advanced our understanding of both the characteristics of individuals likely to lead them to compromise morality in pursuit of personal gain and the conditions under which such behavior is most likely.

We argue that an underexplored element in much of this research is the frequent tendency of decision makers to engage in motivated information processing—acting as motivated Bayesians—thereby resolving the apparent tension between acting egoistically and acting morally. Individuals' flexibility and creativity in how they acquire, attend to, and process information may allow them to reach the desirable conclusion that they can be both moral and egoistic at the same time. The extensive literature in psychology and growing literature in economics reviewed above provide compelling evidence that behavior in many domains with a moral component is often driven by such self-serving information processing, suggesting that incorporating the underlying psychology into economic models is a worthwhile endeavor for future investigation.

References

Abeler, Johannes, Anke Becker, and Armin Falk. 2014. "Representative Evidence on Lying Costs." *Journal of Public Economics* 113(May): 96–104.

Alesina, Alberto, and Paola Giuliano. 2011. "Preferences for Redistribution." Chap. 4 in *Handbook of Social Economics*, vol. 1, edited by Jess Benhabib, Alberto Bisin, and Matthew O. Jackson. North Holland.

Andreoni, James. 1990. "Impure Altruism and Donations to Public Goods: A Theory of Warm-Glow Giving." *Economic Journal* 100(401): 464–77.

Andreoni, James, Justin M. Rao, and Hannah Trachtman. Forthcoming. "Avoiding The Ask: A Field Experiment on Altruism, Empathy, and Charitable Giving." *Journal of Political Economy.*

Ariely, Dan, and Michael I. Norton. 2008. "How Actions Create—Not Just Reveal—Preferences." *Trends in Cognitive Sciences* 12(1): 13–16.

Armstrong, Lance, and Sally Jenkins. 2003. "Every Second Counts." New York: Doubleday.

Babcock, Linda, and George Loewenstein. 1997. "Explaining Bargaining Impasse: The Role of Self-Serving Biases." *Journal of Economic Perspectives* 11(1): 109–26.

Babcock, Linda, George Loewenstein, Samuel Issacharoff, and Colin Camerer. 1995. "Biased Judgments of Fairness in Bargaining." *American Economic Review* 85(5): 1337–43.

Babcock, Linda, Xianghong Wang, and George Loewenstein. 1996. "Choosing the Wrong Pond: Social Comparisons in Negotiations that Reflect a Self-Serving Bias." *Quarterly Journal of Economics* 111(1): 1–19.

Bartoš, Vojtěch, Michal Bauer, Julie Chytilová, and Filip Matejka. Forthcoming. "Attention Discrimination: Theory and Field Experiments with Monitoring Information Acquisition." *American Economic Review.*

Batson, Daniel C., Diane Kobrynowicz, Jessica L. Dinnerstein, Hannah C. Kampf, and Angela D. Wilson. 1997. "In a Very Different Voice: Unmasking Moral Hypocrisy." *Journal of Personality and Social Psychology* 72(6): 1335–48.

Bénabou, Roland, and Jean Tirole. 2006. "Incentives and Prosocial Behavior." *American Economic Review* 96(5): 1652–78.

Bénabou, Roland, and Jean Tirole. 2011. "Identity, Morals, and Taboos: Beliefs as Assets." *Quarterly Journal of Economics* 126(2): 805–855.

Bertrand, Marianne, Simeon Djankov, Rema Hanna, and Sendhil Mullainathan. 2007. "Obtaining a Driver's License in India: An Experimental Approach to Studying Corruption." *Quarterly Journal of Economics* 122(4): 1639–76.

Bertrand, Marianne, and Sendhil Mullainathan. 2004. "Are Emily and Greg More Employable than Lakisha and Jamal? A Field Experiment on Labor Market Discrimination." *American Economic Review* 94(4): 991–1013.

Camerer, Colin, and Richard H. Thaler. 2003. "In Honor of Matthew Rabin: Winner of the John Bates Clark Medal." *Journal of Economic Perspectives* 17(3): 159–76.

Chance, Zoë, Michael I. Norton, Francesca Gino, and Dan Ariely. 2011. "Temporal View of the Costs and Benefits of Self-Deception." *PNAS* 108(S3): 15655–59.

Chance, Zoë, Francesca Gino, Michael I. Norton, and Dan Ariely. 2015. "The Slow Decay and Quick Revival of Self-Deception." *Frontiers in Psychology*, vol. 6, Article 1075.

Charness, Gary, and Matthew Rabin. 2002. "Understanding Social Preferences with Simple Tests." *Quarterly Journal of Economics* 117(3): 817–69.

Dana, Jason, Roberto A. Weber, and Jason Xi Kuang. 2007. "Exploiting 'Moral Wiggle Room': Experiments Demonstrating an Illusory Preference for Fairness." *Economic Theory* 33(1): 67–80.

DellaVigna, Stefano, John A. List, and Ulrike

Malmendier. 2012. "Testing for Altruism and Social Pressure in Charitable Giving." *Quarterly Journal of Economics* 127(1): 1–56.

Di Tella, Rafael, Sebastian Galiani, and Ernesto Schargrodsky. 2007. "The Formation of Beliefs: Evidence from the Allocation of Land Titles to Squatters." *Quarterly Journal of Economics* 122(1): 209–41.

Di Tella, Rafael, Ricardo Perez-Truglia, Andres Babino, and Mariano Sigman. 2015. "Conveniently Upset: Avoiding Altruism by Distorting Beliefs about Others' Altruism." *American Economic Review* 105(11): 3416–42.

Dunn, Elizabeth W., Lara B. Aknin, and Michael I. Norton. 2008. "Spending Money on Others Promotes Happiness." *Science* 319(5870): 1687–88.

Drugov, Mikhail, John Hamman, and Danila Serra. 2013. "Intermediaries in Corruption: An Experiment." *Experimental Economics* 17(1): 78–99.

Ehrich, Kristine R., and Julie R. Irwin. 2005. "Willful Ignorance in the Request for Product Attribute Information." *Journal of Marketing Research* 42(3): 266–77.

Eil, David, and Justin M. Rao. 2011. "The Good News–Bad News Effect: Asymmetric Processing of Objective Information about Yourself." *American Economic Journal: Microeconomics* 3(2): 114–38.

Erat, Sanjiv. 2013. "Avoiding Lying: The Case of Delegated Deception." *Journal of Economic Behavior & Organization* 93(September): 273–78.

Exley, Christine L. 2016. "Excusing Selfishness in Charitable Giving: The Role of Risk." *Review of Economic Studies* 83(2): 587–628.

Fehr, Ernst, and Klaus M. Schmidt. 1999. "A Theory of Fairness, Competition, and Cooperation." *Quarterly Journal of Economics* 114(3): 817–68.

Fehr, Ernst, and Klaus M. Schmidt. 2006. "The Economics of Fairness, Reciprocity and Altruism—Experimental Evidence and New Theories." Chap. 8 in *Handbook of the Economics of Giving, Altruism and Reciprocity*, vol. 1, edited by Serge-Christophe Kolm and Jean Mercier Ythier, 615–91. Elsevier.

Feiler, Lauren. 2014. "Testing Models of Information Avoidance with Binary Choice Dictator Games." *Journal of Economic Psychology* 45 (December): 253–67.

Festinger, Leon. 1957. *A Theory of Cognitive Dissonance.* Evanston, IL: Row & Peterson.

Fischbacher, Urs, and Franziska Föllmi-Heusi. 2013. "Lies in Disguise—An Experimental Study on Cheating." *Journal of the European Economic Association* 11(3): 525–47.

Fleming, Gerald. 1993. "Engineers of Death." *New York Times*, July 18, sec. E.

Flynn, Francis J., and Vanessa K. B. Lake. 2008. "'If You Need Help, Just Ask': Underestimating Compliance with Direct Requests for Help." *Journal of Personality and Social Psychology* 95(1): 128–43.

Fong, Christina M., and Erzo F. P. Luttmer. 2011. "Do Fairness and Race Matter in Generosity? Evidence from a Nationally Representative Charity Experiment." *Journal of Public Economics*, Charitable Giving and Fundraising Special Issue 95(5–6): 372–94.

Fox, Craig R., and Amos Tversky. 1995. "Ambiguity Aversion and Comparative Ignorance." *Quarterly Journal of Economics* 110(3): 585–603.

Frey, Bruno S., and Stephan Meier. 2004. "Social Comparisons and Pro-Social Behavior: Testing 'Conditional Cooperation' in a Field Experiment." *American Economic Review* 94(5): 1717–22.

Frohlich, Norman, Joe Oppenheimer, and Anja Kurki. 2004. "Modeling Other-Regarding Preferences and an Experimental Test." *Public Choice* 119(1–2): 91–117.

Gibson, Rajna, Carmen Tanner, and Alexander F. Wagner. 2013. "Preferences for Truthfulness: Heterogeneity among and within Individuals." *American Economic Review* 103(1): 532–48.

Gino, Francesca, Shahar Ayal, and Dan Ariely. 2013. "Self-Serving Altruism? The Lure of Unethical Actions that Benefit Others." *Journal of Economic Behavior & Organization* 93(September): 285–92.

Gino, Francesca, Michael I. Norton, and Dan Ariely. 2010. "The Counterfeit Self: The Deceptive Costs of Faking It." *Psychological Science* 21(5): 712–20.

Gneezy, Uri. 2005. "Deception: The Role of Consequences." *American Economic Review* 95(1): 384–94.

Gneezy, Uri, Silvia Saccardo, Marta Serra-Garcia, and Roel van Veldhuizen. 2016. "Motivated Self-Deception, Identity and Unethical Behavior." Working paper.

Grossman, Zachary, and Joël van der Weele. 2013. "Self-Image and Strategic Ignorance in Moral Dilemmas." University of California at Santa Barbara, Economics Working Paper Series qt0bp6z29t. Department of Economics, UC Santa Barbara. https://ideas.repec.org/p/cdl/ucsbec/qt0bp6z29t.html.

Haisley, Emily C., and Roberto A. Weber. 2010. "Self-Serving Interpretations of Ambiguity in Other-Regarding Behavior." G*ames and Economic Behavior* 68(2): 614–25.

Hamman, John, George Loewenstein, and Roberto A. Weber. 2010. "Self-interest through Delegation: An Additional Rationale for the Principal–Agent Relationship." *American Economic Review* 100(4): 1826–46.

Hastorf, Albert H., and Hadley Cantril. 1954. "They Saw a Game; A Case Study." *Journal of Abnormal and Social Psychology* 49(1): 129–34.

Hoffman, Elizabeth, Kevin McCabe, and Vernon L. Smith. 1996. "Social Distance and Other-Regarding Behavior in Dictator Games." *American Economic Review* 86(3): 653–60.

Hsee, Christopher K. 1996. "Elastic Justification: How Unjustifiable Factors Influence Judgments." *Organizational Behavior and Human Decision Processes* 66(1): 122–29.

Konow, James. 2000. "Fair Shares: Accountability and Cognitive Dissonance in Allocation Decisions." *American Economic Review* 90(4): 1072–91.

Krupka, Erin L., and Roberto A. Weber. 2013. "Identifying Social Norms Using Coordination Games: Why Does Dictator Game Sharing Vary?" *Journal of the European Economic Association* 11(3): 495–524.

Kunda, Ziva. 1987. "Motivated Inference: Self-Serving Generation and Evaluation of Causal Theories." *Journal of Personality and Social Psychology* 53(4): 636–47.

Kunda, Ziva. 1990. "The Case for Motivated Reasoning." *Psychological Bulletin* 108(3): 480–98.

Larson, Tara, and C. Monica Capra. 2009. "Exploiting Moral Wiggle Room: Illusory Preference for Fairness? A Comment." *Judgment and Decision Making* 4(6): 467–74.

Lazear, Edward P., Ulrike Malmendier, and Roberto A. Weber. 2012. "Sorting in Experiments with Application to Social Preferences." *American Economic Journal: Applied Economics* 4(1): 136–63.

Matthey, Astrid, and Tobias Regner. 2011. "Do I Really Want to Know? A Cognitive Dissonance-Based Explanation of Other-Regarding Behavior." *Games* 2(1): 114–35.

Mazar, Nina, On Amir, and Dan Ariely. 2008. "The Dishonesty of Honest People: A Theory of Self-Concept Maintenance." *Journal of Marketing Research* 45(6): 633–44.

Messick, David M., and Keith P. Sentis. 1979. "Fairness and Preference." *Journal of Experimental Social Psychology* 15(4): 418–34.

Mobius, Markus M., Muriel Niederle, Paul Niehaus, and Tanya S. Rosenblat. 2011. "Managing Self-Confidence: Theory and Experimental Evidence." NBER Working Paper 17014.

Norton, Michael I., Joseph A. Vandello, and John M. Darley. 2004. "Casuistry and Social Category Bias." *Journal of Personality and Social Psychology* 87(6): 817–31.

Olken, Benjamin A. 2007. "Monitoring Corruption: Evidence from a Field Experiment in Indonesia." *Journal of Political Economy* 115(2): 200–249.

Pager, Devah, and Diana Karafin. 2009. "Bayesian Bigot? Statistical Discrimination, Stereotypes, and Employer Decision Making." *Annals of the American Academy of Political and Social Sciences* 621(January): 70–93.

Pager, Devah, and Lincoln Quillian. 2005. "Walking the Talk? What Employers Say versus What They Do." *American Sociological Review* 70(3): 335–80.

Paharia, Neeru, Karim S. Kassam, Joshua D. Greene, and Max H. Bazerman. 2009. "Dirty Work, Clean Hands: The Moral Psychology of Indirect Agency." *Organizational Behavior and Human Decision Processes* 109(2): 134–41.

Paharia, Neeru, Kathleen D. Vohs, and Rohit Deshpandé. 2013. "Sweatshop Labor Is Wrong Unless the Shoes are Cute: Cognition Can Both Help and Hurt Moral Motivated Reasoning." *Organizational Behavior and Human Decision Processes* 121(1): 81–88.

Quattrone, George A., and Amos Tversky. 1984. "Causal versus Diagnostic Contingencies: On Self-Deception and on the Voter's Illusion." *Journal of Personality and Social Psychology* 46(2): 237–48.

Rabin, Matthew, and Joel L. Schrag. 1999. "First Impressions Matter: A Model of Confirmatory Bias." *Quarterly Journal of Economics* 114(1): 37–82.

Rodriguez-Lara, Ismael, and Luis Moreno-Garrido. 2012. "Self-interest and Fairness: Self-serving Choices of Justice Principles." *Experimental Economics* 15(1): 158–75.

Sarin, Rakesh K., and Martin Weber. 1993. "Effects of Ambiguity in Market Experiments." *Management Science* 39(5): 602–15.

Schweitzer, Maurice E., and Christopher K. Hsee. 2002. "Stretching the Truth: Elastic Justification and Motivated Communication of Uncertain Information." *Journal of Risk and Uncertainty* 25(2): 185–201.

Shalvi, Shaul, Jason Dana, Michel J. J. Handgraaf, and Carsten K. W. De Dreu. 2011. "Justified Ethicality: Observing Desired Counterfactuals Modifies Ethical Perceptions and Behavior." *Organizational Behavior and Human Decision Processes* 115(2): 181–90.

Shalvi, Shaul, Francesca Gino, Rachel Barkan, and Shahar Ayal. 2015. "Self-serving Justifications: Doing Wrong and Feeling Moral." *Current Directions in Psychological Science* 24(2): 125–30.

Shaw, Alex, Natalia Montinari, Marco Piovesan, Kristina R. Olson, Francesca Gino, and Michael I. Norton. 2014. "Children Develop a Veil of Fairness." *Journal of Experimental Psychology: General* 143(1): 363–75.

Shleifer, Andrei. 2004. "Does Competition Destroy Ethical Behavior?" *American Economic Review* 94(2): 414–18.

Skilling, Jeffrey K. 2002 [2011]. "Jeff Skilling's Congressional Testimony." Testimony given February 7, 2002 to the Subcommittee on

Oversight and Investigations. Posted April 24, 2011 at *Enron Online: The Enron Blog*. http://enron-online.com/2011/04/24/jeff-skillings-congressional-testimony/.

Snyder, Melvin L., Robert E. Kleck, Angelo Strenta, and Steven J. Mentzer. 1979. "Avoidance of the Handicapped: An Attributional Ambiguity Analysis." *Journal of Personality and Social Psychology* 37(12): 2297–2306.

Steele, Claude M. 1988. "The Psychology of Self-Affirmation: Sustaining the Integrity of the Self." In *Advances in Experimental Social Psychology*, Vol. 21: *Social Psychological Studies of the Self: Perspectives and Programs*, 261–302. San Diego, CA, US: Academic Press.

Weiner, Bernard. 1985. "An Attributional Theory of Achievement Motivation and Emotion." *Psychological Review* 92(4): 548–73.

Wiltermuth, Scott S. 2011. "Cheating More When the Spoils Are Split." *Organizational Behavior and Human Decision Processes* 115(2): 157–68.

Journal of Economic Perspectives—Volume 30, Number 3—Summer 2016—Pages 213–234

In Defense of the NSF Economics Program

Robert A. Moffitt

In 1975, Senator William Proxmire of Wisconsin began a series of highly publicized "Golden Fleece" awards for purportedly frivolous federally funded research projects. His first award was to the National Science Foundation for a social science grant of $84,000 to investigate the reasons that people fall in love. Not coincidentally, in 1981, the federal Office of Management and Budget proposed to Congress a 75 percent reduction in the NSF social science budget, with some of that funding being shifted to the natural sciences, because social science research "was of lesser importance to the economy than the support of the natural sciences." While that specific reduction did not pass Congress, the NSF Economics Program budget did fall by 40 percent from about $10 million in fiscal year 1980 to approximately $6 million three years later. The budget did not re-attain its 1980 real value until 1996 but, since then, it has gone through periods of expansion and contraction. By fiscal year 2013, the real Economics Program budget had again fallen below its 1980 value.

The stagnation of the NSF Economics Program budget has not been the result of general fiscal challenges facing the federal government, but rather of a targeted effort to promote other areas of research at NSF. Indeed, the overall NSF budget has grown sevenfold since 1980 while the Economics Program budget has risen only two-and-a-half times. As a result, the Economics Program budget currently constitutes one-half of 1 percent of the total NSF budget, down from a little over 1 percent in 1980.

■ *Robert A. Moffitt is the Krieger-Eisenhower Professor of Economics, Johns Hopkins University, Baltimore, Maryland. His email address is moffitt@jhu.edu.*

† For supplementary materials such as appendices, datasets, and author disclosure statements, see the article page at
http://dx.doi.org/10.1257/jep.30.3.213 doi=10.1257/jep.30.3.213

In the last several years, reductions in economics and social science research funding have again been proposed. A Congressional study of the National Science Foundation in 2011 argued that there was widespread waste in its funding (Coburn 2011). It cited the Panel Study of Income Dynamics (partly funded by the NSF Economics Program) as one example of waste, and also listed a study of dynamic pricing in a perishable goods market as a second example. The latter study, later published in the *Journal of Political Economy*, involved dynamic pricing of how ticket prices to sporting events change as the time of the event approaches (Sweeting 2012).

In February 2013, Eric Cantor of Virginia, then the Majority Leader of the House of Representatives, stated: "Funds currently spent by the government on social science … will be better spent helping find cures to diseases." He later added that he was strongly in favor of increasing biomedical funding at NSF and "reducing funding for lower-priority programs like social and political science research" (Cantor 2013a, b). In May 2015, the House of Representatives passed authorizing legislation that, breaking with precedent, required specific appropriations for different NSF directorates—rather than letting NSF itself set scientific priorities—and specified a 45 percent reduction in the NSF directorate for social and behavioral sciences. In June 2015, the House of Representatives passed appropriations legislation that would have redirected a substantial amount of the NSF budget for the social and behavioral sciences to the natural sciences and engineering.[1]

Economics research has also been challenged in agencies other than NSF. In July 2012, the House Subcommittee on Labor, Health and Human Services, and Education recommended to the full House Appropriations Committee a prohibition on any funds given to the National Institutes of Health (NIH) to be used "for any economic research programs, projects, or activities" (COSSA 2012). While this legislation did not pass Congress, the NIH announced administrative regulations in November 2015 that singled out economics research for special treatment, explicitly prohibiting several specific areas of health economics from NIH support.

Economists are rarely of a single opinion on any subject, and the value of government support for economic research is no exception. Milton Friedman (1981) argued that the cuts proposed to the Economics Program at NSF should be extended to the natural sciences; indeed, he proposed that the National Science Foundation be abolished. He argued that the peer review process in academic journals does not reward innovative research and that academic institutions, foundations, and donations from private individuals would be better funders of economics research than the government. This led to a lively debate in the pages of this journal, with then

[1] This legislation is still pending in Congress. Social science funding at NSF other than the Economics Program has also been targeted. In March 2013, Congress passed legislation, later signed by President Obama, that directed that NSF could no longer fund any projects in its political science program unless they were certified by the NSF Director as "promoting national security or the economic interests of the United States" (Consolidated and Further Continuing Appropriations Act of 2013, P.L. 113-6, Division B, Title III, Sec. 543, enacted March 26), another interference in the process of peer review by which the merit of scientific investigations is ordinarily determined. This restriction was lifted in legislation passed by Congress in January 2014.

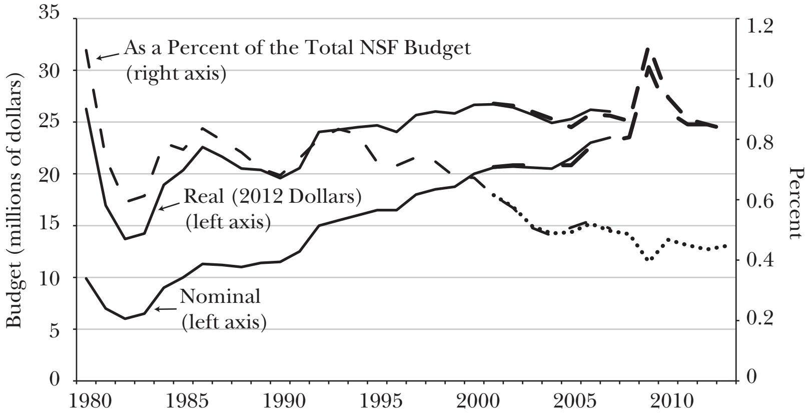
AEA President-Elect Zvi Griliches (1992) arguing in support of NSF funding for economics—citing the importance of peer review in picking the most scientifically worthy studies to fund and bemoaning the public perception that economics funding "is just another barrel of pork"—and Friedman and others (1994), including Merz et al. and Laband et al., taking the opposing side several issues later.

In this essay, I reconsider the case for government funding of economic research, with the NSF Economics Program as the leading example. I first set the stage with some background and statistics concerning the NSF Economics Program, and I present some examples of how NSF-supported economic research has made major contributions to a more informed discussion of policy tradeoffs and alternatives. I then tackle the more difficult question of whether the type of economics research funded by NSF would be supported by other agents and institutions in the absence of NSF with a detailed discussion of the issue. It is of course impossible to know what would happen in the counterfactual world where the NSF did not exist or was much smaller than it is now. I use the traditional public goods argument for governmental support for research that is in the general societal interest to argue that neither private firms, universities (at least those without large endowments), foundations, nor private individuals would likely provide comprehensive support for basic economics research across all areas in the discipline in the way that the NSF Economics Program does, although there are several specific areas of economics research where nongovernment funders have been important. Finally, I briefly review the small empirical literature examining the impact of research on outcomes related to scientific productivity, a literature which is weak and inconclusive and, in fact, virtually nonexistent for government economics funding specifically. Despite the weakness of this empirical evidence, I argue that the a priori case for government funding of comprehensive, general-purpose basic economics research is very strong. Further, the NSF Economics Program has an excellent record of funding research in areas that have made major contributions to the discipline and to public policy.

Overview of the NSF Economics Program

The NSF Economics Program was created in 1960 and funds basic research in economics across all subfields in the discipline (the definition of "basic research" is a little fuzzy and will be discussed further below). The Program currently receives proposals twice a year, proposals which are first evaluated by a set of outside reviewers drawn from the profession and then by a committee of NSF-appointed economists with rotating terms. The committee members read the outside reviews, add their own critical evaluations, and then rank the proposals. The NSF program officers then fund proposals based on the rankings and on available funds. The average size of an award currently is approximately $75,000 per year (although there is considerable variation around that average) including indirect costs and with a typical duration of three years. In recent years, the Program receives about 200–300 proposals per year and approximately 20–30 percent of proposals are funded, hence

Figure 1
NSF Economics Program Budget, 1980–2013



Source: The series from 1980–2007 are from Plott (2010, figure 3). The series from 2001–2013 are from personal correspondence with Dr. Nancy Lutz. The increase in 2009 is from the American Recovery and Reinvestment Act.

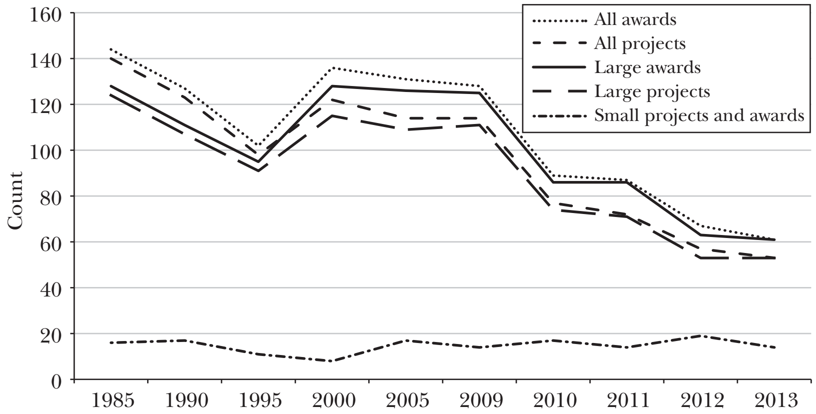
making about 60 grants per year. Approximately one-quarter of funded proposals go to young investigators rather than senior and established researchers.

Figure 1 shows the size of the NSF Economics Program budget in real and nominal terms from 1980 to 2013 as well as its value relative to the total NSF research budget. In the last few years, total spending of the Economics Program has been about $25 million. As noted, its real value was lower in 2013 than in 1980 (the one-shot infusion of additional funds a few years ago was part of the American Recovery and Reinvestment Act of 2009). Figure 2 shows trends in the number of Projects and Awards since 1985. Separate awards are often made to two or more researchers who are located at different institutions but their research constitutes only one project, so there are always more awards than projects. But both have generally declined over time, especially since 2009, and the decline in projects and awards has mainly occurred in the funding of large grants (note that, for small grants, there is no distinction between awards and projects since they are always made to only one institution). Figure 3 shows trends in the median size of awards and projects, both for all grants as well as for small and large ones separately. Award and project sizes in aggregate grew significantly in the early 1990s, then grew more modestly after that.[2]

NSF Economics makes awards with budgets that cover faculty salary, research assistant expenses, lab experiment and capital expenses, field experiment expenses,

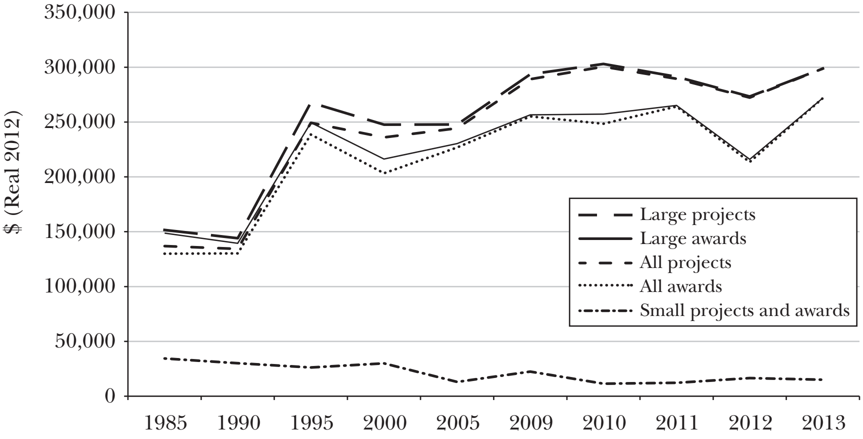
[2] The drop in 2012 was for idiosyncratic reasons, as applications with multiple Principal Investigators declined, NSF program staff made an effort to reduce the size of many awards, and some funds in that year were used to pay off awards made in earlier years.

Figure 2
Number of NSF Economics Awards and Projects, 1985–2013



Source: Personal communication with Dr. Nancy Lutz. Small awards are those of $50,000 or less (2012 dollars) and large awards are those of more than $50,000.

Figure 3
Median Size of Awards and Projects, 1985–2013
(in real 2012 dollars)



Source: Personal communication with Dr. Nancy Lutz. Small awards are those of $50,000 or less (2012 dollars) and large awards are those of more than $50,000. Figures are in real 2012 CPI-U-RS (Consumer Price Index for Urban Consumers, Research Series) dollars.

the cost of data acquisition, travel, and other items. The current average breakdown of budgets across these categories is approximately 30 percent for the salary of the Principal Investigator(s), 25 percent for graduate student support, 32 percent for indirect costs, and 13 percent for everything else.[3] In addition, while the official NSF position on salary support is that it will pay up to two months of summer salary, at a rate of two-ninths of the faculty member's nine-month salary, it reserves the right to negotiate the amount and, in practice, it has generally limited the total amount to $25,000 per year since 2009. That amount is currently about 15 percent of the mean salary of full professors in economics at PhD-granting institutions (Scott and Siegfried 2014) and hence less than two-ninths. NSF funding cannot be used to pay for a lighter teaching load (although much NIH funding does so).

NSF reviews and ranks proposals on two criteria: Intellectual Merit and Broader Impacts. The former basically covers the scientific merit of the project, including the importance of the question and the validity and quality of the research design. It is important to note that NSF tries not to take a position on the relative importance of different subfields within economics, different methodological approaches, or theory versus empirical work. It accepts proposals, and appoints review committee members, who represent all the major fields in economics, from pure theory to applied microeconomics, and from econometric theory to macroeconomics. It is fair to say that it makes an attempt to fund the highest-quality proposals in each major subfield of the discipline.

The Broader Impacts criterion, at least in Economics, refers in many cases to the promise of the project to bear on some issue of public policy or guidance for future policy decisions. A substantive emphasis on the criteria used in the Broader Impacts has been present for most of the NSF's history (Rothenberg 2010), but the term was made official in its current form only in 1996. Over time, reviewers have been encouraged to give it more weight in their rankings. For economics proposals, however, almost any topic can be argued to affect knowledge of the real world, however tangentially, and hence to have some bearing on policy at least in some indirect way. In addition, NSF sees its role as supporting basic research, which includes research that is relevant to policy only indirectly. So-called line agencies of the federal government, which administer specific government programs and fund research on those programs, have a different goal than basic research.

The Broader Impacts criterion also includes whether the project benefits teaching, training, and learning of students; whether underrepresented minorities are likely to benefit; and whether research infrastructure is enhanced by the project. These criteria presumably reflect social goals that NSF, either by itself or from the

[3] These percentages are from the budgets as presented in the proposals at the time of initial awards, but differ from the actual allocations because NSF allows considerable freedom to principal investigators to shift funds across categories during the later periods of the award. Like all government agencies, NSF pays average cost, not marginal cost, thereby subsidizing some of the activities of the organizations receiving the funding.

direction of the legislative or executive branches, have judged to be in the society's best interest.

Examples of NSF-Supported Research on Policy Issues

The NSF Economics program provides support to basic research, although that term differs from its common usage in economics. Economists distinguish between "pure theory" and "applied theory," between "pure econometrics" and "applied econometrics," and between "microeconomic theory" and "applied microeconomics," for example. But all these fields are basic in the sense used in government research funding, for even applied research in economics often does not concern specific programs (think of the vast literature on estimating the rate of return to education, for example, or the estimation of wage elasticities of labor supply). Nevertheless, much of the "basic" research funded by NSF has indeed concerned policy issues, which is not surprising since so much of the research in the discipline in general is policy-oriented and has become more so over time. Although most of that research has been empirical, there have been significant theoretical developments in policy areas like optimal taxation, market structure and antitrust, and school choice designs, to name only three.

The argument here will be that the economic research on policy issues that NSF has funded represents a major intellectual achievement and has greatly informed public discussions of policy issues in a large number of areas. However, whether economic research on policy issues has had a significant impact on policy itself is a separate question and one on which the views of economists differ. Plott (2010), for example, offers a lengthy discussion of the fundamental contributions of economic research in general to policy, ranging across a vast number of areas, and his discussion overlaps heavily with the types of projects funded by NSF since NSF tries to support general economic research. Indeed, the volume in which Plott's essay appears is devoted to the demonstration of the impact of economic research on policy in a variety of areas. On the other hand, many economists are quick to point out the failure of economic principles to affect policy, citing examples like the failure to enact a carbon tax, the existence of a seriously inefficient tax system, and the failure of economic research on welfare programs to have any effect on welfare reform policy.

Table 1 lists a (decidedly nonexhaustive) sampling of topics that NSF has supported over the years that have had a major impact on policy discussions, and sometimes on policy itself, together with just one or two illustrative citations for each. For example, the Economics Program has supported many studies in the area of environmental economics, cap-and-trade systems, carbon taxation, and related issues, including supporting the work of Nordhaus, Oates, and other economists working in the area.[4] The early promotion of emissions trading systems by

[4] All the topics noted in Table 1, and all economists who are named in this section, can be found in the NSF list of past awards at http://www.nsf.gov/awardsearch/.

Table 1
Areas of Policy Contribution by NSF-Supported Research

Topic	*Description*	*Illustrative Citations*
Emissions trading	Extensive research supported by NSF Economics on emissions trading and cap-and-trade systems was followed by federal legislation for sulfur dioxide trading and by several state-level trading systems	Baumol and Oates (1971)
Monetary policy	Research establishing the importance of expectations, predictable policy rules, and aggressive responses to inflation led to a long period of stability of national GDP	Taylor(1980)
Measurement of inflation	NSF-sponsored research contributed to the discovery of multiple biases in the construction of the Current Price Index that had resulted in an overstatement of inflation by over 1 percentage point per year, affecting COLAs for Social Security	Boskin et al. (1996)
Inflation-indexed bonds	Much research supported by the Program has concerned asset markets and their prices and how inflation-indexed bonds offer investors a more secure asset, contributing to the understanding of such bonds offered by the US Treasury since 1997	Campbell et al. (2009)
Trade liberalization	The NSF program supported numerous studies of the complex relationship between trade liberalization and growth, contributing to the public policy discussions that led to major liberalizations in the 1990s	Grossman and Helpman (1994)
Deregulation	NSF supported a large number of studies of deregulation of the airline industry, electricity markets, financial markets, and the hospital industry, studies which influenced public policy on deregulation over several decades	Bailey and Williams (1988); Rassenti and Smith (1998)
Government auctions	NSF-supported research on auction design led to improvements in government auctions of the radio spectrum yielding a revenue gain of $100 billion. Economic research also informed auction designs for airport slots, offshore oil leases, forests, toxic assets, and mineral rights	Milgrom and Weber (1982); Wilson (1992)
Antitrust	Among many other topics, economic research on cross-price elasticities of demand in differentiated product markets have influenced DoJ and FTC Horizontal Merger Guidelines	Berry, Levinsohn, and Pakes (1995)
Kidney transplantation	Research on methods of overcoming obstacles to kidney exchange among incompatible donors led to new organizations of exchanges that saved thousands of lives	Roth et al. (2005)
Private pensions	Behavioral economics research on undersaving for retirement and passive behavior in the face of default rules led to major 2006 federal legislation changing default rules and other pension characteristics	Choi et al. (2002); Thaler and Benartzi (2004)

economists contributed to the turnaround of opinion on the usefulness of the approach and to both national and state-level policy enactments utilizing variants of the idea. However, once again, it can be argued that economists' ideas on this topic have not influenced policy to nearly the extent they should.

NSF-funded economics research supported the development of macroeconomic theories with model-consistent expectations and associated studies of monetary theories and macroeconomic dynamics, including the work of Kydland, Lucas, Phelps, Prescott, Sargent, and Wallace in the 1970s and 1980s. This work has had fundamental repercussions both on academic theories of the business cycle and economic growth, as well as on thinking about government policy. Some might argue that this work has had more methodological influence on subsequent policy-oriented work in macroeconomics than a direct influence on policy (for example, on monetary neutrality), but Taylor (2010) argues otherwise. One occasionally hears the argument that NSF funding decisions are biased toward research that casts a favorable light on government programs, but much of this research provides a definitive counterexample because much of it suggests the weakness of certain government macroeconomic policies. NSF also supported research by Taylor leading to the well-known Taylor Rule, which has also had a major impact on monetary policy.

Improved price indices and the measurement of inflation provide an example of a topic commonly regarded by the public as an obscure technical problem but which has implications for virtually every area of the discipline and for the application of economics to the real economy. NSF Economics has supported the research of Boskin, Diewert, Gordon, Griliches, Hausman, Jorgensen, Pollak, and Rosen, among many others, on this topic. Much of that NSF-funded research fed into the analysis and recommendations of the well-known Boskin Commission Report (Boskin et al. 1996; discussed in a six-paper "Symposium on Measuring the CPI" in the Winter 1998 issue of this journal). This research had a major impact on price index development in the federal government, especially the development of the Consumer Price Index for Urban Consumers, Research Series (CPI-U-RS), which is now widely preferred to the plain vanilla CPI. Another important policy-related topic in inflation concerns the development and impact of inflation-indexed bonds, which are now being offered by the US Treasury and are a key component of the portfolio of many investors (in this journal, see Wilcox 2008). Economic researchers who work on pricing in asset markets and portfolio decisions have been supported by NSF Economics, and their research has contributed to the understanding of those Treasury bonds, as illustrated by the work of Campbell, Shiller, and Viceira.

Yet another area of NSF Economics support has been in the controversial area of trade liberalization and growth. Particularly in the 1990s, NSF supported the research of Edwards, Grossman, Helpman, Paul Romer, and many others working in the area. The research by economists made major contributions to the complex issues involved in trade liberalizations. More recently, the Program has supported numerous studies of international trade between developed countries as well, both general models of trade as well as of specific trade agreements such as the North

American Free Trade Agreement (see Kehoe and Ruhl 2013, for one of many possible examples). Again, however, the politics of trade barriers has proved to be an obstacle to fully incorporating the lessons of economic research in the area.

NSF Economics has supported a large number of studies in the area of industrial organization, including research on estimation of cross-price elasticities of demand in differentiated product markets, which has had a major influence, both theoretically and computationally, on the development of Horizontal Merger Guidelines jointly authored by the US Department of Justice and the Federal Trade Commission (2010). Going farther back, the extensive theoretical and empirical literature on resale price maintenance, which dates at least back to influential economic research in the 1960s (for example, Telser 1960), has been supported by the NSF Economics Program. Another major area of NSF support in the area of industrial organization has been support of economic research on deregulation, applied to many different industries including airlines, electricity, banking, and hospitals (in addition to the illustrative citations in the table, see Kahn 1971; Bailey, Graham, and Kaplan 1985). Legislative actions and decisions in both the executive branch and the courts have used this research to make major changes in public policy toward regulation.

The Economics Program has also given major support to research on auction methods. This is an area where there was rapid development of theory starting in the late 1980s, accelerating in the 1990s, and continuing to this day. Research of Ledyard, McMillan, Milgrom, Plott, Roth, Smith, Wilson, and others has been supported. This research had a direct impact on the auctioning of the radio spectrum in 1995 with a revenue gain of $100 billion to the federal government (McAfee, McMillan, and Wilke 2010). In another area of mechanism design, NSF has supported research on the development of kidney exchanges, which has led to important reworking of those exchanges in actual practice. Somewhat similar methods have been used to generate deferred-acceptance algorithms for public schools in New York and Boston that solved serious problems with methods that had been used previously (for a review, see Roth 2010).

Economic research on behavioral economics has also been supported by the NSF Economics program. One example of NSF-support research with a policy impact concerns default rules for saving. Years of research by economists on this issue led to federal legislation in 2006 that changed pension default provisions. More recently, an Executive Order by the White House (2015) directed all agencies to use behavioral economics insights in the design of their programs. The Executive Order also mentioned another recent governmental reform based on NSF-supported research, a reform to simplify and streamline college financial aid application forms as of the 2017–2018 application year. Those forms have been a long-standing source of barriers to application because of their complexity and length. NSF-supported research in this area includes Bettinger, Long, Oreopolous, and Sanbonmatsu (2012), which showed that simplifications in, and streamlining of, the Free Application for Federal Student Aid (FAFSA) application form increased application rates and, ultimately, college attendance rates.

The NSF Economics Program has also supported many individual economists who are widely regarded as having made major contributions to the discipline. The program has supported every Nobel Prize winner in Economics since 1998 and almost every John Bates Clark medal winner since 1961. It has supported the research of ten out of the last eleven chairs of the Council of Economic Advisors, including those serving under both Republican and Democratic administrations, constituting further testimony to the support of economists who engage with real-world policy problems. The Economics Program has provided partial support to the Carnegie-Rochester Public Policy conference and the Brookings Panel on Economic Activity. Datasets like the Panel Study of Income Dynamics (PSID), whose core funding is provided by NSF, and the Health and Retirement Study (HRS), whose core funding is provided by the National Institutes of Health, have generated thousands of published articles in economics journals: for example, a Google Scholar search on the Panel Study of Income Dynamics yields 24,000 hits as well as 12,900 hits for the Health and Retirement Survey.

None of this evidence proves that NSF funds have been solely responsible for the research or that the research would not have been done otherwise. (The next section takes up this question.) In addition, the topics supported by the NSF have also been supported by many other funders and institutions as well. However, the evidence does demonstrate that NSF has successfully identified some of the best research in the discipline and has supported projects in areas of economics that have had a major impact on informed discussions of public policy, and often on policy itself.

Would Economics Research Be Underfunded in the Absence of NSF?

The key question for assessing the value of NSF Economics Program funding is: what is the marginal social benefit of another dollar of funding or, alternatively, the marginal social benefit of the total budget of the Program? A subquestion concerns the marginal social benefit of funding different types of research (theory versus empirical, for example). In principle, these should be empirical questions to be answered with data but, perhaps unsurprisingly, determining the answers is difficult and there is essentially no credible existing evidence addressing these questions. In the absence of determining evidence, circumstantial and indirect reasoning must be brought to bear.

First, I should make one general point: it is important to recognize the small size of the NSF Economics Program. It funds only 60 new grants per year, spread out over all subfields in economics, and each grant lasts about three years, which means that about 180 are in place in any given year. There are about 12,000 AEA members at academic institutions in about 800 departments of economics, and if NSF were to fund only the 180 best research projects from those members, a very large number of meritorious proposals would obviously go unfunded. In addition, given its tiny size, it is not surprising that the aggregate impact of NSF on research in the entire discipline of economics is small and that most papers published in economics journals are not NSF-supported. But this is irrelevant to the question of whether the marginal benefit of another dollar of NSF spending is large or small: indeed, it may only demonstrate that the NSF Economics budget should be dramatically increased.

Also, the annual NSF Economics Program budget of $25 million is miniscule by almost any comparison. The NIH spends approximately $194 million per year on a wide range of economics projects, which is about eight times the NSF budget, even though it is focused only on health economics (Schuttinga 2011). Moreover, most NIH economics funding is basic research in the governmental meaning of the word. To believe that the current allocation is an optimal allocation of government expenditure, one also must believe that NSF dollars spent on economics topics like those in Table 1 have a very low marginal social benefit relative to dollars spent on health economics. It is likely that spending within the Federal Reserve System on basic research in macroeconomics (again, just one subset of economics) is also larger than at the National Science Foundation.

Other major funders of economics research also provide support in excess of $25 million or slightly lesser amounts but for much more specialized uses. The National Bureau of Economic Research spends $32 million per year on its research programs and administrative expenses[5] yet concentrates its activities on only a select number of areas, mostly empirical, in the discipline. The Russell Sage Foundation, a small foundation focusing entirely on social policy funding, spends $13 million per year on its activities, more than a half of what NSF Economics spends, despite the relative narrowness of its agenda.[6]

Despite the small scale of the NSF Economics Program, there are those who believe that its marginal social benefit is small, or even negligible. The primary argument in favor of this position is that economists who are funded by NSF would have done the same research without the funding. Cochrane (2012) represents the views of many economists in arguing that academic economists use government funding to conduct the research they would have done otherwise. Milton Friedman, as noted, was of the view that universities, foundations, and private philanthropists should be the funders of economic research instead of the government. A particularly stark way to address this issue is to ask the hypothetical question of whether, if NSF were abolished as Friedman proposed, other institutions would pick up the slack and provide the same funding as NSF currently does or, more generally, whether the same amount of research would take place. If so, then all the NSF research supporting the topics in Table 1 would have been conducted anyway and the marginal social benefit of NSF is effectively zero. While such a counterfactual is inherently speculative, it is worth conducting this thought experiment in some detail because it puts the potential marginal social benefit of NSF into useful

[5] NBER Summary Financial Statement for the fiscal year ended June 30th, 2015; available at http://www.nber.org/info.html (accessed June 28, 2016).

[6] Audited financial statement for fiscal year 2015, http://www.russellsage.org/about/financial-statements.

perspective. The rest of this section discusses, therefore, whether private firms, universities, foundations, or private philanthropists would fund economic research to the same degree as NSF.

Private firms fit the classic public goods model, where the free rider problem would lead to underprovision of a public good by a private firm. Research that produces public knowledge fits the two conditions for a pure public good: public knowledge is nonrival, because one person's consumption of it does not diminish another person's consumption, and it is nonexcludable because, once published, no one can be denied access to it (Nelson 1959; Arrow 1962).[7] Private actors acting in their own self-interest would not support the optimal level of the good or service because they would not be able to capture all of the societal benefits.

The failure of private firms to conduct basic research is particularly likely since such research has major impacts only after cumulative years of collective effort and research. In economics, the linkage from research to public policy can be long, diffuse, and uncertain. The scholarly achievements on the topics delineated in Table 1 took years of research from a large number of economists, building on each other's work, and the outcome of the full body of work could not have been anticipated in advance nor could the marginal contributions of any single study or even small group of studies be identified sufficiently in advance to warrant private investment.

Having said this, some large firms do hire staff economists to conduct research that is academic in nature or hire consultants to conduct research on topics that have value as basic research. The field of auctions is a prominent example, and many of the most fundamental contributions to the field were the result of private consulting contracts. More recently, firms like Microsoft, Amazon, Google, and other web-based firms have sponsored economic research that has led to publications on basic research topics published in the leading journals. In addition, research in many other areas of economics published in the journals is a result of consulting agreements, which can be verified by the recent enactment of disclosure statements for papers published in the *American Economic Review.*

Nevertheless, these examples are the exception rather than the rule. Most economic research sponsored by private firms must ultimately be seen as benefitting the bottom line of the firm, and the research that is sponsored has to obey boundaries where at least some such benefit can be established. Private firms rarely sponsor research on most types of pure theory, theoretical econometrics, or fundamental foundations of the macroeconomy. Even when firms occasionally support basic research, such efforts often are eventually closed down, as the example of the celebrated Bell Labs research shop in the 1970s and early 1980s demonstrates.

Research universities are more likely candidates for picking up the slack since they are nonprofit institutions for whom a primary goal is basic research. But it is

[7] Here I ignore the fact that most prominent journals charge prices for their product and that many individuals do not have access to journals through institutional subscriptions and would have to pay to gain access.

worth parsing this possibility in detail by considering the research university environment in the United States: that is, what the goals of universities are, along with their business model and level of financial resources. For example, public universities are funded by state legislatures whose goals are only partly to establish nationwide research excellence and whose main interest is in providing education to the residents of the state. Recent trends in reduced funding of state universities reveal that university research is a declining priority. State universities typically also do not have sufficient funds to hire graduate research assistants for their faculty, to pay for data acquisition above a nominal level, to establish research centers entirely supported by state funds, or to pay for the expenses associated with building an experimental lab and running lab experiments. NSF funding does support all of those expenses.

Private universities have a greater financial potential for research support, but it is only the top 20 or so universities with large endowments that have the funds to support research expenses other than faculty time. Most private research universities in the United States have modest endowments at best and are unable to support hiring of research assistants or expenses for data acquisition, experimental labs, or research centers.

Salary payments for faculty time is a separate issue, for the business model of most research universities is to pay salary for only nine months of the year, so that summer research must be funded by outside sources solicited by the researcher. This is widely regarded as an accounting fiction, as most academic economists continue to do research out of intellectual interest or for career incentives year-round whether they find extra funding or not; they simply make their "9-month" salary last 12 months. If they do the same summer research with NSF funding that they would do without it, NSF funding is simply a transfer from the taxpayer to the researcher.[8] However, the elasticity of substitution probably differs by the type of research—for example, whether the research has empirical content. Further, it should be recalled that salary support constitutes only 30 percent of the average research grant, which therefore constitutes an upper bound on the transfer. But ultimately the question of whether a faculty member would work at a faster pace or more intensively in the summer with a grant than without it is again an empirical question that is not possible to determine with current data and on the basis of current research. Further, even if it could be established that the average elasticity of substitution is nonzero, or it could be established that certain fields have a higher elasticity than others, it is difficult to imagine how NSF could incorporate this information into its review and award process. Asking the NSF staff or NSF review committees or reviewers to make a judgement on how much of a proposed project would be conducted without NSF funding would be an impossible task and lead to

[8] If the funds were granted for the researcher's paid research time during the nine months of the year, the university would reduce its salary payment by the amount of the time spent on the project and then the university, not the researcher, would be the recipient of the transfer, but still with no change in aggregate research output. But if the funds were granted for the researcher's teaching time, and the university used the released salary funds to hire another teacher, teaching output would remain unchanged but aggregate research output would rise (there is of course an opportunity cost to the other teacher's time).

discretionary judgments that would be unpopular and fraught with error. It is also unlikely that NSF should simply rule out entire fields of research on the grounds that the average elasticity in those fields is high. This is why NSF instead ignores this issue and just uses the criteria discussed above—Intellectual Merit and Broader Impacts—to make its awards.

Large-scale data collection is an area where substitution is least likely to occur. No university would be willing to support a dataset like the Panel Study of Income Dynamics—a dataset that is used by macro- and well as micoeconomists—out of its own private funds, nor would foundations. Other government agencies would not do so because the PSID is insufficiently focused on the programmatic concerns of line agencies (indeed, the PSID was created by the Department of Health and Human Services in the 1960s, but eventually they chose to drop it given its general focus, and NSF picked it up). Nor would other government statistical agencies pick up the PSID; it would be the view of the Census Bureau and the Bureau of Labor Statistics, for example, that the PSID is mainly used by academics to address narrow research questions, whereas they see their mission as to more directly provide descriptive statistics to Congress, and the federal government in general, charting the state of American society and its economy.[9]

General-purpose data collection is one area where a good case can be made for direct salary support for economists. Even with government funding for all the nonpersonnel expenses necessary to collect data, most researchers would prefer not to spend their time being involved in a large-scale data collection exercise for a public-use dataset that will lead to publications mostly by other economists; they would rather work on another research paper of their own. Without salary compensation, most economists would simply not engage in that activity. It is difficult to imagine that universities, even those private universities with the largest endowments, would initiate data collection projects like the Panel Study of Income Dynamics, the Health and Retirement Survey, or many financial datasets purely out of their own private funds.

Finally, private foundations and private philanthropy could pick up much of the research that would be funded by a government research agency like NSF. On the face of it, this outcome seems very unlikely, for foundations typically have specific missions, defined by their founders or donors or board members, which focus on certain public issues of interest. Foundations also typically do not make awards on the basis of peer review or merit review, but instead generally make awards through a private solicitation process. Similarly, individual private philanthropic donors typically have specific research agendas in mind and usually are interested in only very applied research on a particular topics of personal interest. Indeed, major increases

[9] This point is reinforced by the experience of the Survey of Income and Program Participation, which is funded by the Census Bureau, and the National Longitudinal Survey, which is funded by the Bureau of Labor Statistics. Both are most heavily used for research and not for the main missions of their agencies and, as a consequence, these datasets are continually at risk of deep funding cuts within their agencies or complete elimination when agency budgets are tight.

in private support of research have already occurred in the natural sciences, where large donors have picked up some of the slack from reductions in government funding. They have only provided funds for narrowly defined topics of interest, and the consequences for the advance of knowledge in the natural sciences have been problematic (Broad 2014).

If government funding of research did not exist, it is possible that foundation boards and private donors might see their responsibilities differently and might fund some portion of the basic research that would, in the alternative universe, have been government funded. But there would surely be limits to that support, and it almost surely would not support the broad, comprehensive agenda of research in all areas of economics that an agency like NSF supports.

As with private firms, there are exceptions in some areas. A number of private foundations sponsor billions of dollars of research, a slice of which goes to economists: the Gates Foundation, the Rockefeller Foundation, the Ford Foundation, the Hewlett Foundation, and many others. A recent report stated that $52 billion was given by US foundations in 2014 (Foundation Center 2014), and if only a tiny sliver of that went to economists, it would still be vastly greater than the NSF Economics Program budget. Much of the economic research in developing countries is supported by foundations, including many of the randomized controlled trials currently being conducted there. However, this is the exception rather than the rule. Most subfields in economics do not have foundation support of this kind, and it is unlikely that new foundations would spring up to serve other subfields in economics were NSF to be abolished. Further, from a societal point of view, it would again seem to be a distortion of resource allocation to devote disproportionate support to selected fields in this way. In addition, once again, it is difficult to imagine how NSF could deal with this issue, except, for example, by deciding they would no longer support randomized controlled trials in developing countries because these can be funded by other institutions.

The conclusion to be drawn from this discussion is that a significant body of economic research would almost surely be lost if NSF were not to exist. The loss would differ by field and by whether other funds are available, and the loss would be concentrated on research at public universities and less-endowed private universities that do not have the funds to support nonsalary expenses for projects that would require them. The collection of large-scale datasets would be significantly reduced with subsequent damage to the state of scientific knowledge.

Empirical Evidence on the Impact of Government Funding of Research

Very little systematic empirical work has been done to estimate the impact of the NSF Economics Program. In correspondence published in this journal, Laband, Piette, Ralston, and Tollison (1994) regressed citations to papers published in the leading economics journals on whether the author(s) had previously received an NSF award, finding a positive effect of an additional 5.6 percent citations. But as the authors recognized, those who receive award funding may be those who would have done better research than those who do not receive funding, and the positive correlation they found could have occurred even if the funding had no effect on whether either group did their research. This is a form of the well-known "selection" problem, and in this case, the authors argued that the best interpretation of their findings was that NSF was picking the better proposals. Arora and Gambardella (2005) conducted a similar but more extensive investigation, but again not controlling for selection, and found NSF grants in the late 1980s to have considerably smaller effects, which were present mostly for younger researchers. One possible difference between the two analyses is that Arora and Gambardella did not use total citations but used a quality-weighted index based on journal impact factors. But neither of these analyses can be given much weight given their inability to control for selection. Jaffe (2002) has argued that funding agencies need to build more evaluation structures into their award procedures, possibly by randomization, and this would seem to apply to the NSF Economics Program as well.

Although pertaining to a different federal government agency and not restricted to the subject of economics, two recent studies of funding from the National Institutes of Health (NIH) may be relevant. Jacob and Lefgren (2011) used a regression discontinuity design to estimate the effect of winning an NIH grant on subsequent research productivity measured by citations, leveraging the fact that proposals are awarded at NIH on the basis of a well-defined score and looking at scores above or below a particular cutoff. Jacob and Lefgren found very small effects on future productivity, with a winning proposal leading to only one additional publication over the next five years, about a 7 percent increase. The positive effect was also mainly concentrated among younger researchers, consistent with the notion that older researchers are more likely to have alternative sources of funding. As is always the case with regression discontinuity designs, and as acknowledged by the authors, the estimated effects only apply to marginal applicants, and it could be that the effects on inframarginal applicants are larger (or smaller). The authors also discussed at length whether the small effect arose because most marginal proposals that were not funded by NIH obtained funding from some other source. The authors' data were not definitive on this question, but they suggested that other funding for rejected proposals could often be obtained from coauthors who obtained grants from other NIH or from non-NIH sources, from NSF (although these grants could be small in magnitude), or from other sources such as foundations and universities. While this may seem to contravene the arguments in the last section, this result is likely to be heavily influenced by the predominance of biomedical funding at NIH, which may have more alternative funding sources.[10]

Freeman and Van Reenen (2009) studied whether the doubling of the NIH research budget from 1998 to 2003 had positive effects on research activity in the

[10] Li and Agha (2015) also examine the correlation between NIH priority scores and later publication outcomes, but the authors explicitly disavow any attempt to separate selection of better proposals into higher priority scores from a true impact of the funding itself.

biomedical sciences. They stressed that the doubling and a subsequent rapid deceleration created severe problems for biomedical researchers because of adjustment costs incurred in ramping up research facilities and then adjusting downwards. This "stop and go" cycle is a familiar problem in science funding, especially in the case of biomedical researchers. The deceleration has been argued to disadvantage young researchers in particular (Stephan 2012). Freeman and Van Reenen argued that the impact of large increases and decreases in science funding depend on how funding agencies make decisions about how many awards to make, the value of each, and whether they are made to younger rather than older researchers, and they argued that NIH had not paid sufficient attention to this issue. For present purposes, the "stop and go" cycle creates problems of inference for any study that uses the number of awards or amount of funding as a determining variable because it implies there are significant lags in any output response and that short-run and long-run effects are likely to be quite different.

Another, older, literature of possible relevance, albeit even more tangential to NSF Economics funding, takes a "knowledge production function" approach to the effect of research and development on economic growth, productivity, or some other measure of output (as in Griliches 1979). While early papers applied the approach to investigate the effects of industry research and development, other papers studied the effect of academic research on growth and productivity, often using a traditional growth accounting framework. Although there are endogeneity and identification issues in these studies, most find a positive effect (Jaffee 1989; Adams 1990; Mansfield 1991; Adams and Griliches 1996). A more relevant literature for present purposes is that examining whether public funding of research and development in the sciences has positive or negative effects on private research and development—or put differently, whether public and private research and development are complements or substitutes. The relatively large literature on this topic was reviewed by David, Hall, and Toole (2000), who found results all over the map, and very sensitive to specification, level of aggregation, and other issues (see also Diamond 2008). The authors concluded that the evidence, taken as a whole, is ambivalent on the main question of interest. A side note is that much of the research cited in this paragraph was funded by the NSF Economics Program.

Conclusion

The NSF Economics Program is under challenge in Washington. However, the evidential basis for supporting a reduction in its budget is essentially nonexistent. On the contrary, the circumstantial and indirect evidence for the opposite position—that the Program is dramatically underfunded—is strong. The number of grants made is tiny and has been declining over time, and the real budget is no larger than it was in 1980, despite the tremendous growth and productivity of the discipline over the last 35 years. The size of the budget is miniscule compared to that of other federal research funding agencies and as compared to that spent by some other institutions that fund economics research in specialized areas. Yet the NSF program has supported research in a number of areas related to public policy where that research has had a major impact on the discipline, on thinking about policy problems, and often on policy itself, demonstrating the high value of the program. Finally, while some fraction of the funds expended to support faculty summer research time may merely substitute for time they would have spent on the research even without that funding, this fraction is likely to be small as a percent of the average NSF grant and likely to apply only to certain types of research. Overall, the elimination of NSF support for research expenses, especially those of a nonsalary nature, on a broad range of basic research topics would almost surely lead to the disappearance of much research at universities other than those with large endowments and would not be replaced by funding from other institutions.

The critical missing element in existing discussions of this issue is a strong empirical basis demonstrating the marginal social benefit of NSF spending, either for marginal increases or decreases in its budget or for its spending as a whole. It would be valuable to know that marginal benefit not only for the current mix of NSF spending but also for specific projects such as general purpose data collection and data purchase, expenses for research assistants, the lab, and randomized controlled trials, and for empirical work versus theory, either microeconomic theory or econometric theory, for example. It would be interesting to determine whether the marginal social benefit of NSF support of these types of expenditures differs by whether they are made to economists at well-endowed universities rather than those with modest or small endowments. Conducting such empirical work is challenging because exogenous changes in the NSF budget would not be easy to find, and because it would require working with confidential data inside NSF on rejected proposals and the subsequent funding and research productivity of these projects. Nevertheless, further progress on this important issue for government support of economic research requires that such efforts proceed.

■ *The author would like to thank David Autor, Katherine Baicker, Carol Corrado, Mark Gertler, Dana Goldman, Gordon Hanson, Anne Krueger, Daniel Newlon, Charles Plott, Jonathan Skinner, James Smith, Phillip Swagel, John Taylor, and Timothy Taylor for comments. None of these individuals should be held responsible for the views expressed herein. Nancy Lutz and Rachel Rosen also provided assistance with NSF Economics Program statistics.*

References

Adams, James. 1990. "Fundamental Stocks of Knowledge and Productivity Growth." *Journal of Political Economy* 98(4): 673–02.

Adams, James, and Zvi Griliches. 1996. "Measuring Science: An Exploration." *Proceedings of the National Academy of Sciences,* November 12, 93(23): 12664–670.

Arora, Ashish, and Alfonso Gambardella. 2005. "The Impact of NSF Support for Basic Research in Economics." *Annals of Economics and Statistics,* July/December, no. 79–80, pp. 91–117.

Arrow, Kenneth. 1962. "Economic Welfare and the Allocation of Resources for Invention." In *The Rate and Direction of Inventive Activity: Economic and Social Factors,* by the National Bureau of Economic Research, 609–625. Princeton University Press.

Bailey, Elizabeth E., and Jeffrey R. Williams. 1988. "Sources of Economic Rent in the Deregulated Airline Industry." *Journal of Law and Economics* 31(1): 173–202.

Bailey, Elizabeth E., David R. Graham, and Daniel R. Kaplan. 1985. *Deregulating the Airlines.* MIT Press.

Baumol, William J., and Wallace E. Oates. 1971. "The Use of Standards and Prices for Protection of the Environment." *Swedish Journal of Economics* 73(1): 42–54.

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

Bettinger, Eric P., Bridget Terry Long, Philip Oreopoulos, and Lisa Sanbonmatsu. 2012. "The Role of Application Assistance and Information in College Decisions: Results from the H&R Block FAFSA Experiment." *Quarterly Journal of Economics* 127(3): 1205–42.

Boskin, Michael J., Ellen R. Dulberger, Robert J. Gordon, Z. Griliches, and D. Jorgenson. 1996. *Toward a More Accurate Measure of the Cost of Living: Final Report to the Senate Finance Committee from the Advisory Commission to Study the Consumer Price Index.* December 4. Washington: US Senate.

Broad, William J. 2014. "Billionaires with Big Ideas Are Privatizing American Science." *New York Times,* March 15.

Campbell, John Y., Robert J. Shiller, and Luis M. Viceira. 2009. "Understanding Inflation-Indexed Bond Markets." *Brookings Papers on Economic Activity,* Spring, 40(1): 79–120.

Cantor, Eric. 2013a. "Making Life Work: Remarks by Majority Leader Eric Cantor." Speech given Febrary 5, 2013, at the American Enterprise Institute, Washington, DC. Video: http://aei.org/events/2013/02/05/making-life-work-remarks-by-majority-leader-eric-cantor.

Cantor, Eric I. 2013b. "Republicans and Science." Letter to the editor. *New York Times,* February 14.

Choi, James M., David Laibson, Brigitte C. Madrian, and Andrew Metrick. 2002. "Defined Contribution Pensions: Plan Rules, Participant Decisions, and the Path of Least Resistance." *Tax Policy and the Economy,* edited by James Poterba. Cambridge, MA: National Bureau of Economic Research, pp. 67–114.

Coburn, Tom A. 2011. *The National Science Foundation: Under the Microscope.* A Report by Tom A. Coburn, April.

Cochrane, John. 2012. "Subsidies for Economists?" *The Grumpy Economist,* John Cochrane's blog, April 10. http://johnhcochrane.blogspot.com/2012/08/subsidies-for-economists.html.

COSSA. 2012. "Washington Update." July 23, vol. 31, issue 14. http://archive.constantcontact.com/fs021/1102766514430/archive/1110549012690.html.

David, Paul A., Bronwyn H. Hall, and Andrew A. Toole. 2000. "Is Public R&D a Complement or Substitute for Private R&D? A Review of the Econometric Evidence." *Research Policy* 29(4–5): 497–529.

Diamond, Arthur M., Jr. 2008. "Science, economics of." *The New Palgrave Dictionary of Economics,* edited by Steven N. Durlauf and Lawrence E. Blume, 2nd edition. New York: Palgrave Macmillan.

Foundation Center. 2014. "Key Facts on U.S. Foundations: 2014 Edition."

Freeman, Richard, and John Van Reenen. 2009. "What If Congress Doubled R&D Spending on the Physical Sciences?" In *Innovation Policy and the Economy,* vol. 9, no. 1, edited by Josh Lerner and Scott Stern, 1–38. University of Chicago Press.

Friedman, Milton. 1981. "An Open Letter on Grants." *Newsweek,* May 18, p. 99.

Friedman, Milton, Thomas Merz et al., David Laband et al., and Zvi Griliches. 1994. "Correspondence: National Science Foundation Grants for Economics." *Journal of Economic Perspectives* 8(1): 199–205.

Griliches, Zvi. 1979. "Issues in Assessing the Contribution of Research and Development to Productivity Growth." *Bell Journal of Economics* 10(1): 92–116.

Griliches, Zvi. 1992. "A Note from the President-Elect." *Journal of Economic Perspectives* 6(4): 3–5.

Grossman, Gene M., and Elhanan Helpman. 1994. "Protection for Sale." *American Economic*

Review 84(4): 833–50.

Jacob, Brian A., and Lars Lefgren. 2011. "The Impact of Research Grant Funding on Scientific Productivity." *Journal of Public Economics* 95(9–10): 1168–77.

Jaffe, Adam B. 1989. "Real Effects of Academic Research." *American Economic Review* 79(5): 957–70.

Jaffe, Adam B. 2002. "Building Programme Evaluation into the Design of Public Research-Support Programmes." *Oxford Review of Economic Policy* 18(1): 22–34.

Kahn, Alfred E. 1971. *The Economics of Regulation: Principles and Institutions.* New York: John Wiley & Sons.

Kehoe, Timothy J., and Kim J. Ruhl. 2013. "How Important is the New Goods Margin in International Trade?" *Journal of Political Economy* 121(2): 358–92.

Laband, David, Michael Piette, Scott Ralson, and Robert Tollison. 1994. "Correspondence: National Science Foundation Grants for Economics." *Journal of Economic Perspectives* 8(1): 201–03.

Li, Danielle, and Leila Agha. 2015. "Big Names or Big Ideas: Do Peer-Review Panels Select the Best Science Proposals?" *Science,* April 23, 348(6233): 434–38.

McAfee, R. Preston, John McMillan, and Simon Wilke. 2010. "The Greatest Auction in History." Chap 7 in *Better Living through Economics*, edited by John J. Siegfried. Harvard University Press.

Mansfield, Edwin. 1991. "Academic Research and Industrial Innovation." *Research Policy* 20(1): 1–12.

Milgrom, Paul, and Robert Weber. 1982. "A Theory of Auctions and Competitive Bidding." *Econometrica* 50(5): 1089–1122.

Nelson, Richard. 1959. "The Simple Economics of Basic Scientific Research." *Journal of Political Economy* 67: 297–306.

Plott, Charles R. 2010. "Overview: Highlights of the Benefits of Basic Science in Economics." In *Better Living through Economics,* edited John Siegfried. Harvard University Press.

Rassenti, Stephen J., and Vernon L. Smith. 1988. "Deregulating Electric Power: Market Design Issues and Experiments." *International Series in Operations Research and Management,* vol. 13, p. 105–20. Dordrecht, Boston, and London: Kluwer Academic.

Roth, Alvin E. 2010. "Deferred-Acceptance Algorithms: History, Theory, Practice." Chap. 9 in *Better Living through Economics* edited by John Siegfried. Harvard University Press.

Roth, Alvin E., Tayfun Sönmez, and M. Utku Ünver. 2005. "Pairwise Kidney Exchange." *Journal of Economic Theory* 125(2): 151–88.

Rothenberg, Marc. 2010. "Making Judgements about Grant Proposals: A Brief History of the Merit Review Criteria at the National Science Foundation." *Technology and Innovation* 12(3): 189–95.

Schuttinga, James A. 2011. "Economics Research at NIH: FY 2009." PowerPoint presentation, May 4.

Scott, Charles, and John Siegfried. 2014. "American Economic Association Universal Academic Questionnaire Summary Statistics." *American Economic Review* 104(5): 678–82.

Stephan, Paula. 2012. *How Economics Shapes Science.* Cambridge: Harvard University Press.

Sweeting, Andrew. 2012. "Dynamic Pricing Behavior in Perishable Goods Markets: Evidence from Secondary Markets for Major League Baseball Tickets." *Journal of Political Economy* 120(6): 133–72.

Taylor, John B. 1980. "Aggregate Dynamics and Staggered Contracts." *Journal of Political Economy* 88(1): 1–23.

Taylor, John B. 2010. "Better Living through Monetary Economics." Chap. 6 in *Better Living through Economics,* edited by John Siegfried. Harvard University Press.

Telser, Lester G. 1960. "Why Should Manufacturers Want Fair Trade?" *Journal of Law and Economics* 3(October): 86–105.

Thaler, Richard H., and Shlomo Benartzi. 2004. "Save More Tomorrow: Using Behavioral Economics to Increase Employee Savings." *Journal of Political Economy* 112(S1): S164–S187.

US Department of Justice and the Federal Trade Commission. 2010. "Horizontal Merger Guidelines." Issued August 19.

White House. 2015. "Executive Order—Using Behavioral Science Insights to Better Serve the American People." https://www.whitehouse.gov/the-press-office/2015/09/15/executive-order-using-behavioral-science-insights-better-serve-american. Accessed October 5, 2015.

Wilcox, David W. 1998. "Policy Watch: The Introduction of Indexed Government Debt in the United States." *Journal of Economic Perspectives* 12(1): 219–227.

Wilson, Robert. 1992. "Strategic Analysis of Auctions." Chap. 8 in *Handbook of Game Theory with Economic Applications*, Vol. 1, edited by Robert Aumann and Sergiu Hart. Amsterdam: North-Holland.

A Skeptical View of the National Science Foundation's Role in Economic Research

Tyler Cowen and Alex Tabarrok

We can imagine a plausible case for government support of science based on traditional economic reasons of externalities and public goods. Yet when it comes to government support of grants from the National Science Foundation (NSF) for economic research, our sense is that many economists avoid critical questions, skimp on analysis, and move straight to advocacy.

In this essay, we take a more skeptical attitude toward the efforts of the National Science Foundation to subsidize economic research. We offer two main sets of arguments. First, a key question is not whether NSF funding is justified relative to laissez-faire, but rather, what is the marginal value of NSF funding given already existing government and nongovernment support for economic research? Second, we consider whether NSF funding might more productively be shifted in various directions that remain within the legal and traditional purview of the NSF. Such alternative focuses might include data availability, prizes rather than grants, broader dissemination of economic insights, and more. Given these critiques, we suggest some possible ways in which the pattern of NSF funding, and the arguments for such funding, might be improved.

Although our discussion here will be phrased in terms of grants for economic research from the National Science Foundation, similar arguments would apply to grants in support of economic research from the National Institutes of Health (which

■ *Tyler Cowen is Holbert L. Harris Chair of Economics and Director, Mercatus Center, and Alexander Tabarrok is Professor of Economics and Bartley J. Madden Chair in Economics at the Mercatus Center, all at George Mason University, Fairfax, Virginia. Their email addresses are tcowen@gmu.edu and Tabarrok@gmu.edu.*

[†] For supplementary materials such as appendices, datasets, and author disclosure statements, see the article page at
http://dx.doi.org/10.1257/jep.30.3.235 doi=10.1257/jep.30.3.235

plays a significant and growing role in funding health care economics and studies of public health performed by economists) and from other agencies. Those interested in a good overview of the economics of science might begin with Diamond (2008).

Evaluating NSF Funding on the Margin

The grants given to economists by the National Science Foundation should be viewed in the context of the portfolio of extensive government support for economic research. About 80 percent of the academic economics sector, measured by the number of students, is accounted for by state universities. Charitable donations to universities and colleges, along with research centers and other nonprofits, are tax-deductible. The government produces a number of datasets widely used in economic research and makes them freely available, including those from the Bureau of Economic Analysis, the Census Bureau, the Bureau of Labor Statistics, Bureau of Justice, the Department of Education, the Federal Reserve, and others. The government also hires economists directly.

Moreover, universities and colleges provide strong incentives for economists and other academics to turn their research into a public good through publication. Consider the mantra of "publish or perish"—what better private incentives could one ask for? Citations to published research are strong predictors of salary for individual professors as well as the prestige of the departments where professors are employed (Ellison 2013; Hilmer, Ransom, and Hilmer 2015). Economists at research universities, in particular, are given high salaries, low teaching loads, plenty of nonstructured time, and access to a highly skilled and motivated labor force at low cost in the form of graduate students. The subsidy that society provides to economists is large, especially once we consider the opportunity cost of highly skilled labor. The obvious question, although one which many economists are reluctant to ask, is how much subsidization of economic research is enough?

A common strategy in defending NSF grants to economists is to point to research that can be linked to substantial real-world policy improvement. But private incentives are strong not only to publish but also to produce research with real-world implications. For example, the promotional NSF (2000) book, *America's Investments in the Future* lauds NSF funding of auction theory and, in particular, the work of Paul Milgrom. Milgrom's work has indeed been spectacular, but the implicit argument that incentives were lacking for work on auction theory seems incorrect. Indeed, few areas in economics have been as privately remunerative as auction theory. As noted on Milgrom's webpage (http://www.milgrom.net/business-activities, accessed April 2, 2016):

> Milgrom has advised bidders in radio spectrum auctions, power auctions, and bankruptcy auctions. One advisee, Comcast and its consortium, SpectrumCo, followed the advice of Milgrom's team in FCC Auction 66 to achieve the most exceptional performance in US spectrum auction history. SpectrumCo saved

> nearly $1.2 billion on its spectrum license purchases compared to the prices paid by other large bidders—such as T-Mobile and Verizon—for comparable spectrum acquired at the same time in the same auction.

Advice on auctions is a highly valuable private good. Hal Varian, Peter Crampton, Preston McAfee, and Susan Athey are just a few of the other notable economists and auction theorists who have in recent years moved to important roles in the private sector. Lest we be misunderstood, we applaud and celebrate such activity. Our point is that the existing public and private incentives for at least some kinds of research are quite strong, and it seems potentially misleading to conclude that, in the absence of NSF grants, private incentives for such work are lacking.

More generally, the classic science-to-technology paradigm suggests that basic research leads to applications. History, however, is full of examples in which the process is reversed and private applications lead to basic research (Kealey 1996). In economics, for example, Koopmans developed some of his key ideas in operations research and resource allocation when working for the British Merchant Shipping Mission in Washington. Advances in finance have often been driven by the demands of the finance industry (Derman 2004). Amazingly, the Vickey–Clarke–Groves auction mechanism was rediscovered by engineers at Google when they were looking for ways to raise revenue efficiently in sponsored search auctions (Varian and Harris 2014). Instead of being subsidized to work in the ivory tower, economists might contribute more to the public good by working at least part of the time directly in the private or government sector—perhaps looking at consumption functions at the Federal Reserve, or at supply and demand coordination at Uber, or on market design at the New York Stock Exchange—and then returning to academia with a heightened sense of which research questions are most useful to pursue.

Crowding Out and Crowding In

If a substantial quantity of economic research is currently being provided by the combination of educational, nonprofit, and for-profit institutions, then additional funding has a lower marginal value. A federal program to fund mosquito control is harder to justify if state and local programs already exist. Furthermore, government funding for economic research may also crowd out other sources of funding. Crowding out will tend to raise the cost of additional public funding because any given increase in net funding will require higher gross funding with the attendant deadweight loss of taxation and also additional administrative costs, including the overhead charged by universities to the National Science Foundation (Noll and Rogerson 1997).

In the broader literature, Hungerman (2005) finds that government provision of welfare crowds out private charity on the order of 25 cents to the dollar. For the National Endowment for the Arts, Dokko (2009) finds crowding out effects of up to 60 cents to the dollar. Such crowding-out does not preclude public provision, but given the welfare costs of taxation—which average perhaps 30 percent of the expenditure (Ballard, Shoven, Whalley 1985; Bohanon, Horowtiz, McClure 2014)—it does reduce the desirable level of public provision.

Table 1
Number of NSF Grants in Economics by Organization

Organization	*Number of grants*	*Percentage of total*	*Cumulative percentage*
National Bureau of Economic Research	212	15.4	15.4
New York University	60	4.3	19.7
Stanford University	56	4.1	23.8
Northwestern University	49	3.5	27.3
Columbia University	48	3.5	30.8
Yale University	48	3.5	34.3
Duke University	46	3.3	37.6
Princeton University	41	3.0	40.6
University of California—Berkeley	40	2.9	43.5
University of Wisconsin—Madison	39	2.8	46.3
University of Pennsylvania	38	2.7	49.0
Harvard University	36	2.6	51.6

Note: NSF grants in field of economics (code 1320) since 2005. From http://www.nsf.gov/awardsearch/advancedSearch.jsp.

Crowding out may vary by the type of public good produced, and we do not know of any specific estimates for NSF economics funding. However, we do know that the NSF allocates most of its funding to high–prestige economists doing mainstream research at wealthy institutions and schools. Over 50 percent of NSF grants since 2005, for example, have gone to just 11 universities and the National Bureau of Economic Research as indicated in Table 1. The NBER itself distributes funding towards top researchers and universities, especially Harvard and MIT. Wachtel (2000) provides an earlier analysis showing NSF grants flow primarily to a small number of prominent institutions, while Feinberg and Price (2004) discuss the role of social capital and connections in the NSF funding process.

If Harvard, MIT, and Stanford do not feel that it is worth paying faculty or providing research support for a certain economics project, should American taxpayers necessarily have a different opinion and feel compelled to fill the gap? And if prominent and well-endowed academic institutions do feel that it is worth paying faculty and providing research support, why should the NSF risk crowding out such support? Those schools have large, already subsidized endowments and also a strong track record in picking research winners.

In addition to crowding out, there is also the possibility of crowding in, which often occurs in science and the arts (Heutel 2014; Cowen 2010). Certification from a centralized governmental authority helps the grant recipient to raise money from other sources. Some of these additional funds may be "new money," but it may also lead to redistribution of the pool of funds. There is also the possibility that crowding in will increase inequality, as even more support is funneled to NSF recipients while other economists or other scientists receive less. In theory, this outcome could be better or worse for encouraging the quality of economic research, but overall economists tend to be relatively critical of "winner-take-all" markets because of increases in inequality, losses in diversity, and increased incentives for rent-seeking (Cook and Frank 1995). In this regard, the long-run funding impact of NSF grants may not be entirely positive.

NSF Fellowships for Graduate Students

The NSF has a longstanding program of supporting graduate students in economics (Freeman, Chang, and Chiang 2005), currently about 30 per year. We have two reservations about this program. First, a preponderance of the fellowships go to graduate students who choose the top-rated schools. Those schools already receive tax subsidies for their sizeable endowments, and already support highly talented graduate students. It is not clear that further public subsidy is warranted. Second, those individuals are usually extremely talented, but we don't know which allocation of their talent would produce the highest social return. If these individuals did not become economists, they might enter other sciences, or business, or the tech world, or perhaps run innovative nonprofits. In which of these areas are the external benefits from creativity the greatest? We do not pretend to know the answer and so the proper assessment of these grants should be agnostic. Citing the quality of the supported individuals chosen only raises the stakes, rather than settling the issue.

Other Open Questions: Opportunity Cost and Elasticity

It seems plausible that NSF Economics funding has a higher marginal value than *some* government programs. We are reasonably confident that NSF funding for economics is a better idea than, say, ethanol subsidies. But defenses of a government program that compare it only to apparently inferior investments are just special pleading. Even if we believe that NSF funding more than "pays for itself," at the relevant margin, the alternative may be other programs which "pay for themselves" even more. For example, is the NSF Economics Program a better investment than, say, speeding the approval processes at the Food and Drug Administration, hiring additional good economists to work at the US Treasury, or funding research into communicable diseases? The answers are far from obvious. It would be quite remarkable if NSF funding for economics were the number one activity at the margin for government funds.

Some of the proposals to reduce NSF funding of economics and other social sciences would explicitly reallocate the funds to other branches of science, so the question of opportunity cost is pertinent. Given the existence of other non-NSF support for economic research, is spending on economic research of higher value than the average NSF expenditure?

A related question is to consider the elasticity of supply for quality economic research, with respect to wages or payment. Over the last few decades, many in the economics profession have concluded that tax cuts for high earners, at current tax rates, have relatively small effects on labor supply; for example, that argument is often cited as one reason why the Bush tax cuts passed into law in 2001 yielded disappointing economic results. Yet when it comes to NSF grants, there is often the implicit presumption that the elasticity of supply for economic research with respect to additional government grants is relatively large. We are agnostic on the elasticity question, but without understanding this issue, it is difficult to evaluate NSF bang for the buck, and thus economists should not be so confident about the efficacy of this funding. [1]

Are NSF Grants the Best Method of Government Support for Economic Science?

Public goods theory tells us that the National Science Foundation should support activities that are especially hard to support through traditional university, philanthropic, and private-sector sources. This insight suggests a simple test: to the extent that the NSF allocates funds to genuine public goods as opposed to subsidies on the margin, we ought to see a large difference in the kinds of projects the NSF supports compared to what the "market" sector supports. But what stands out from lists of prominent NSF grants (like the one provided by Moffitt in this symposium) is how similar they look to lists of "good" research produced by today's status quo. If we take public goods theory seriously, what areas of economics should be supported?

Replication

The NSF could support replication studies on a significant scale. A significant fraction of economic research does not easily replicate (Dewald, Thursby, Anderson 1986; Chang and Li 2015; Duvendack, Palmer-Jones, and Reed 2015; but also, Camerer et al. 2016, who offer a more positive outlook for experimental economics). Replication and reproducibility studies are true public goods that are not rewarded highly by most top journals or by the tenure process at research universities. Consider Zimmermann's (2015) plea for a replication journal:

> There is very little replication of research in economics, particularly compared with other sciences. This paper argues that there is a dire need for studies that replicate research, that their scarcity is due to poor or negative rewards for replicators, and that this could be improved with a journal that exclusively publishes replication studies. I then discuss how such a journal could be organized, in particular in the face of some negative rewards some replication studies may elicit.

Instead of pointing to the prestigious economists whose research they have funded, perhaps the NSF might point to the prestigious research that has been convincingly replicated—or not replicated.

[1] Arora and Gambardella (2005) find that NSF grants have only a slight positive impact on the marginal productivity of well-known researchers, though a higher impact on lesser-known researchers. However, their dataset is from 1985–1990, and we do not consider this to be a decisive result for today.

Datasets

The NSF should encourage the availability of useful, publicly available datasets especially in areas where data is not yet collected in a sufficient manner. Such data are a public good across the entire community of researchers, and collecting data seems to be an underprovided activity (for example, Belter 2014 finds a high value for public datasets in science). Furthermore, in the tenure and promotion process, creating a new and important dataset is not strongly rewarded at most schools, as typical standards for promotion and tenure emphasize the publication of new research in top journals. That is all the more reason why government science funding should pay more attention to the creation or opening up of useful datasets.[2]

One example of an NSF success is the funding supplied to the Panel Study of Income Dynamics. Thousands of articles have sprung from this dataset and influenced discourse on income mobility, taxes, demography, and many other areas of direct policy relevance. The NSF should be proud of its support here, but the next step is to consider whether more funds should go to comparable enterprises. Looking through NSF grants, we do not see that creating or opening up datasets has been a priority, much less a dominant form of expenditure. The NSF does require (without enforcement) that NSF-supported economics researchers should make their data available to the public. That is a good idea, but still quite different from funding datasets themselves.

Furthermore, the NSF does not always have to *create* new datasets. It could also play a role in improving current datasets or increasing the availability of data. Many current databases have proprietary status, to varying degrees. Universities may buy data licenses for their own researchers, but they are less willing to pay to open up the data more generally. The NSF could buy access rights or do so in a selective manner with a license for qualified researchers. Of course, that would mean more money sent to the private companies that own such datasets and less money sent to high-prestige economists, but that is one reason why the NSF should consider such a move.

We have also noticed a trend for more work to be done using administrative datasets, which have the troubling property that they are often difficult and expensive for most researchers to access for replication or original research. We do applaud the work of the National Science Foundation to expand the number of Research Data Centers, which are secure Census Bureau facilities at locations around the country where external researchers who fulfill certain requirements are given access to confidential microdata. More could be done, however, especially as this is an issue of a growing importance (for some of the issues, see Card, Chetty, Feldstein, and Saez 2010; Mervis 2014).

[2] Of course, the National Science Foundation is not the only institution that could encourage researchers who produce public goods. Perhaps there should have been more consideration of a Nobel Prize for Irving Kravis, Robert Summers, and Alan Heston, the creators of the Penn World Tables? Or how about a Nobel Prize for Stephen Davis, John Haltiwanger, Ron Jarmin, and Javier Miranda for their work in developing the Business Dynamics Statistics database?

Support for Projects with High Fixed Costs

It is a well-known proposition from industrial organization that markets may underinvest in product variety when fixed costs are high. In the current context, the implication is that the NSF should focus more on funding research areas with relatively high fixed costs, including high capital costs (all forms of research involve some fixed intellectual costs). More concretely, this argument suggests that stronger candidates for support would include expensive or lengthy randomized control trials, costly field experiments, and forms of experimental economics that require costly lab investments.[3]

Conversely, an emphasis on supporting research with high fixed costs would imply less support and perhaps no support at all for economic theory. Pencil and paper and even computers simply aren't very costly, and furthermore economic theory seems to have made bigger breakthroughs before the 1990s than it has made since then (Hamermesh 2013). An emphasis on research with high fixed costs also implies a lower level of support for empirical economics based on readily available datasets where the regressions are run on a personal computer. Again, that kind of research just doesn't cost very much money, and it is already being funded by research universities through their high salaries and low teaching loads.

Dissemination of Economics Research

Another possible task for the NSF is to encourage broader dissemination of economic research. Steps along these lines might include subsidizing open access journals and also spreading educational resources, including disseminating knowledge about good teaching and communications techniques, and encouraging economists to do more outreach to policymakers.

In recent years, the NSF has supported some teaching and access activities, such as a recent project on teaching economics in community colleges. The researchers surveyed community college economics faculty and organized meetings to address the problem of the isolation of community college instruction from professional standards. Many part-time faculty in community colleges do not even have a graduate degree in economics, and so they are not always well-informed about what they teach. Maier and Chi (2016) survey this project and offer a generally favorable assessment. (This project, like the NSF support for economics graduate students, is actually funded through the educational branch of the NSF rather than the economics section.) Still, in percentage terms this kind of project constitutes only a small amount of what the NSF does in economics.

If we ask ourselves which economics activity is undersupplied as a public good in today's profession, a lot of indicators point in the direction of good teaching rather than quality research. Many poor economic policy decisions stem from a basic neglect of straightforward economic concepts that do not rest on any particularly

[3] In the interests of full disclosure, we should note that our own department, George Mason University, has received some National Science Foundation grants for its work in experimental economics, which does require a costly lab. We have not ourselves been the recipients of such funds.

partisan view: for example, farm subsidies are undesirable economic policy for straightforward reasons; free trade is (usually) good for reasons that have been well understood for over two centuries; rent-seeking problems were analyzed persuasively by Adam Smith; the basic arguments against price controls have been known for over a century; the Fed should not tighten money if a recession is approaching; and (most) tax cuts do not automatically pay for themselves but rather require offsetting expenditure cuts over some time horizon. To be sure, not all questions of economic policy are as simple as those listed here. If the federal government is considering an extension of the Earned Income Tax Credit, it might want some precise estimates of costs and elasticities, of the kind that would require sophisticated research, which could be done by economists inside of government.

Again, our theme is that economists should be willing to face tradeoffs when thinking about NSF Economics funding. One possible tradeoff is that dissemination and outreach regarding well-accepted basic economic insights may be a more valuable public good than the support of marginal cutting-edge research.

Another public good the NSF might fund is simply a study of which of its previous expenditures on economics have had the greatest marginal value-added. Based on conversations with NSF staff in economics, we are not able to identify such a study. One proposal would be to fund such a study and then follow many or all of its recommendations; after all, the NSF presumably believes it is capable of generating useful research results with practical implications.

High Risk, High Gain, and Far Out Basic Research

It's not surprising that the NSF funds mainstream projects similar to what is already being funded because the NSF chooses which projects to fund by committee peer review. Committee peer-review will gravitate towards funding that reviewers think is valuable and high quality. Almost inevitably, giving high-prestige economists a leading role in deciding on NSF grants means funding research that is relatively close, in intellectual terms, to what already is well accepted in the profession. While this procedure may seem self-evidently correct to most high-prestige economists, it seems peculiar to believe that the best mechanism for allocating public goods should be dominated by the preferences of suppliers. Moreover, it seems more likely than not that supporting mainstream research leads to inefficient allocation.

A common argument for government funding of science, originating with Arrow (1966), is that the private sector will underfund some high-risk projects because the private rate of risk aversion and loss aversion is too high relative to the socially optimal rate. The Defense Advanced Research Projects Agency (DARPA), for example, doesn't subsidize private sector research but instead creates small teams to take on "high-risk," "high-gain," "far out" basic research (to use terms that have been prominent in the agency's mission back to its earliest days, see Hafner and Lyon 1996, p. 22). DARPA is widely considered to be the most successful government research funding agency.

Consider whether the economics profession would have benefited from support for a broader range of research in the lead-up to the Great Recession.

Prior to the Great Recession, a number of mainstream economists argued that the economy was undergoing a Great Moderation (Clarida, Gali, and Gertler 2000). Since then, there has of course been a dramatic rethinking of what had been considered settled wisdom. Paul Krugman, one of the world's most recognized economists and a prominent gadfly, says that the past 30 years of macroeconomic research was "spectacularly useless at best, and positively harmful at worst" (as quoted in the *Economist* 2009). More measured reevaluations have occurred under the auspices of Olivier Blanchard and the IMF who held substantial conferences in 2011, 2013, and again in 2015 on the theme of "Rethinking Macro Policy" (Blanchard, Dell'Ariccia, and Mauro 2010, 2013, see also the conference webpage: http://www.imf.org/external/np/seminars/eng/2015/macro3). One does not have to agree with the post-Keynesians, econo-physicists, or the Austrians to see an argument for broadening the NSF portfolio of grants beyond the usual mainstream contributors (as the NSF has occasionally done sometimes in the mathematics section rather than the economics section). Indeed, the fact that one does not agree with radically different approaches is a good case for funding them. Perhaps the NSF should offer greater support for heterodox economics research for the same reason government funding may be necessary to preserve crop varieties against the risks of monoculture.

The focus of NSF funding on low-risk, mainstream projects is by no means restricted to economics. Biochemist and Nobel Laureate Roger Kornberg lamented in 2007 (as quoted in Lee 2007, referring to funding from the National Institutes of Health) that "the funding decisions are ultraconservative. If the work that you propose to do isn't virtually certain of success, then it won't be funded. And of course, the kind of work that we would most like to see take place, which is groundbreaking and innovative, lies at the other extreme." Ironically, better bibliometric citation measures may make this problem worse, given that Wang, Veugelers, and Stephan (2016) find that novel papers are more likely to be published in journals with lower impact factors. The conservatism of the committee review system suggests that we should use a mix of funding mechanisms to increase the variety of projects that are funded.

Innovation Prizes, Not Grants

Incentives for academic research can be provided through prizes as well as grants (Tabarrok 2011; Williams 2012). The NSF focuses on grants, but arguably a greater share of government support of science should take the form of prizes for achievement of a pre-specified goal or task. Prizes impose greater risk on the scientists. But on other side, the government does not have to decide in advance of production of research who deserves an award, and it is often easier to evaluate excellence after an achievement is completed rather than before the research starts. Furthermore, the government pays out if and only if the valuable end is actually achieved. In recent years, many government agencies including NASA, Health and Human Services, the Environmental Protection Agency, the Department of Agriculture, and the Department of Homeland Security have offered prizes to stimulate research.

DARPA is famous for offering prizes and challenges, including the DARPA Grand Challenge for autonomous vehicles. When the first challenge was held in 2004, the best team travelled just seven miles on a 150 mile course. Nevertheless, the Challenge helped to stimulate deep and surprising innovations that led to today's autonomous vehicles. Or consider how much research was stimulated by Robert Axelrod's iterated prisoner's dilemma competition and the surprising win of the tit-for-tat strategy (Axelrod 1984). The Center for Disease Control recently sponsored a prize to find models to successfully predict the timing, peak, and intensity of the flu using demographic, economic, and social-media data (Office of Science and Technology Policy 2015)—so there is precedent for an agency to offer the types of prizes that might be useful in economics. For example, NSF-sponsored challenges might include better forecasting of recessions and predictions of laboratory and field behavior.

As a simple example, the NSF could sponsor a competition for the best question to add to the Current Population Survey, whereby the NSF would pay the US Bureau of the Census for the winner's question to be added. Kleiner and Krueger (2013) suggest that testing, developing, and editing a new question would cost $50,000 in the first year and less in subsequent years.

Prizes can also confer legitimacy when important ideas come from outside the mainstream. The famous "longitude prize" offered by the British government for a method by which a ship at sea could determine its longitude was won not by Isaac Newton or a member of the Royal Society but by a clockmaker, John Harrison (although Leonhard Euler did receive a runner-up award) (Sobel 1995). A heterodox approach to prediction in economics, for example, would gain greater legitimacy if it were to best mainstream approaches in a competition.

Direct Practical Experience versus Research Funding

If the government wants more of the public good of economic research, it could hire economists directly for this purpose. There is a widespread belief, often expressed in the economics and political science literature, that government relies too heavily on expensive private contractors who often pursue their own agendas, and does not use enough direct employees. When economists themselves need to produce or receive additional ideas, they typically resort to in-house production, rather than outsourcing. For instance, an economist might hire a research assistant and assign that person a specific task, or take on a co-author, but it would be unlikely for an economist to commission outside research through an arms-length relationship. In-house research tends to be more practical, focused, and applied. There is also a general perception that the quality gap between the very top economists and middle-tier economists from good schools has been closing due to the spread of high standards and technical proficiency and the globalization of economics study, among other factors. That change too would seem to militate in favor of more direct commissions of economists as government employees and less use of economists in the role of freestanding discretionary contractors.

DARPA is well-known for hiring researchers for a limited period of time to work to achieve a specific goal. Analogously, if the BEA needs a better method of adjusting price indexes for quality change, perhaps they should hire a team of researchers to work directly on this question. Yet in our experience, academic economists are not eager to compare support for academic research with direct hiring of researchers or government economists. Instead they are keener to argue that when it comes to supporting economists, both methods of supporting economic research should be expanded, which is an interesting case of economists neglecting tradeoffs.

Concluding Remarks

In considering the case for grant-based funding of economics research by the National Science Foundation, we find that a number of pertinent questions are rarely asked, let alone clearly answered. Instead, economists often put forward relatively weak arguments that they would likely dismiss if applied to government subsidies not reserved for economists.

For example, one common approach to defending NSF grants for economists is to list the prestigious individuals with whom the program has been associated. In his paper in this journal, Moffitt notes: "The program has supported every Nobel Prize winner in Economics since 1998 and almost every John Bates Clark medal winner since 1961." (NSF economics funding started in 1960). Indeed, the list of grant recipients from NSF Economics is a literal "Who's Who" of the top economists over the last half-century. But we don't find the prestige of NSF recipients to be a good substitute for an estimate of the public benefits of research. Imagine a group of chefs who defended a hypothetical "National Food Foundation" on the grounds that it had provided grants to Alice Waters, Thomas Keller, Grant Achatz, and every winner of a James Beard Award since 1990. If these names are not familiar, rest assured that their published research output and training of students is very impressive. While we would not consider this information irrelevant (better to fund good chefs than bad ones), as economists we would be unimpressed by this case for government funding of chefs. Talk of how these grants brought about innovations in the culinary arts—such as sous vide, molecular gastronomy, and the introduction of quinoa to the American diet—would also not swing the argument. Instead, as economists, we would focus on how food markets would have operated without such grants and what else might have been done with the money.

Economics should think much harder about the marginal benefits of National Science Foundation grants for economics, and for other subjects, in the context of the many other ways in which society funds research, along with how such money should be spent and what the relevant alternatives might be. There is a good case for a significant change in NSF priorities towards replication and reproducibility of research, data access, and teaching. The extent to which NSF grants add to the sum total of economic research, or whether NSF grants are superior to having the government simply hire economists to perform specified research tasks, isn't obvious. But when it comes to government funding, many economists transform into special pleaders who prefer to ignore tradeoffs. This metamorphosis would not have surprised Adam Smith.

■ *For useful comments and discussions, the authors wish to thank Bryan Caplan, Kevin Grier, Gordon Hanson, Garett Jones, Dan Klein, and several members of the staff at the National Science Foundation. Timothy Taylor was especially helpful. The views expressed in this paper are solely those of the authors.*

References

Arora, Ashish, and Alfonso Gambardella. 2005. "The Impact of NSF Support for Basic Research in Economics." *Annales d'Économie et de Statistique* No. 79/80, pp. 91–117.

Arrow, Kenneth J. 1966. "Discounting and Public Investment Criteria." In *Water Research,* edited by Allen V. Kneese and Stephen C. Smith, 13–32. Baltimore, MA: Johns Hopkins Press.

Axelrod, Robert. 1984. *The Evolution of Cooperation.* Basic Books.

Ballard, Charles L., John B. Shoven, and John Whalley. 1985. "General Equilibrium Computations of the Marginal Welfare Costs of Taxes in the United States." *American Economic Review* 75(1): 128–38.

Belter, Christopher W. 2014. "Measuring the Value of Research Data: A Citation Analysis of Oceanographic Data Sets." *PLOS ONE* 9(3): e92590.

Blanchard, Olivier, Giovanni Dell'Ariccia, and Paolo Mauro. 2010. "Rethinking Macroeconomic Policy." IMF Staff Position Note, February 12. https://www.imf.org/external/pubs/ft/spn/2010/spn1003.pdf.

Blanchard, Olivier, Giovanni Dell'Ariccia, and Paolo Mauro. 2013. "Rethinking Macro Policy II: Getting Granular." IMF Staff Discussion Note. April. http://www.imf.org/external/pubs/ft/sdn/2013/sdn1303.pdf.

Bohanon, Cecil E., John B. Horowitz, and James E. McClure. 2014. "Saying Too Little, Too Late: Public Finance Textbooks and the Excess Burdens of Taxation." *Econ Journal Watch* 11(3): 277–96.

Camerer, Colin F., Anna Dreber, Eskil Forsell, Teck-Hua Ho, Jürgen Huber, Magnus Johannesson, Michael Kirchler, Johan Almenberg, Adam Altmejd, Taizan Chan, Emma Heikensten, Felix Holzmeister, Taisuke Imai, Siri Isaksson, Gideon Nave, Thomas Pfeiffer, Michael Razen, and Hang Wu. 2016. "Evaluating Replicability of Laboratory Experiments in Economics." *Science* 351(6280): 1433–36.

Card, David, Raj Chetty, Martin Feldstein, and Emmanuel Saez. 2010. "Expanding Access to Administrative Data for Research in the United States." NSF White Paper.

Chang, Andrew C., and Phillip Li. 2015. "Is Economics Research Replicable? Sixty Published Papers in Thirteen Journals Say 'Usually Not.'" Finance and Economics Discussion Series 2015-083, Board of Governors of the Federal Reserve System, Washington, DC.

Clarida, Richard, Jordi Galí, and Mark Gertler. 2000. "Monetary Policy Rules and Macroeconomic Stability: Evidence and Some Theory." *Quarterly Journal of Economics* 115(1): 147–80.

Cowen, Tyler. 2010. *Good and Plenty: The Creative Successes of American Arts Funding.* Princeton University Press.

Derman, Emanuel. 2004. *My Life as a Quant: Reflections on Physics and Finance.* Hoboken, N.J.: Wiley.

Dewald, William G., Jerry G. Thursby, and Richard G. Anderson. 1986. "Replication in Empirical Economics: The Journal of Money, Credit and Banking Project." *American Economic Review* 76(4): 587–603.

Diamond, Arthur M., Jr. 2008. "Science, economics of." In *The New Palgrave Dictionary of Economics,* 2nd edition, edited by Steven N. Durlauf and Lawrence E. Blume, pp. 328–334. Palgrave Macmillian. (Also, *The New Palgrave Dictionary* *of Economics Online*, June 15, 2016, http://www.dictionaryofeconomics.com/article?id=pde2008_E000222.)

Dokko, Jane K. 2009. "Does the NEA Crowd Out Private Charitable Contributions to the Arts?" *National Tax Journal* 62(1): 57–75.

Duvendack, Maren, Richard W. Palmer-Jones, and W. Robert Reed. 2015. "Replications in Economics: A Progress Report." *Econ Journal Watch* 12(2): 164–191.

Economist. 2009. "The Other-Worldly Philosophers." July 16. http://www.economist.com/node/14030288.

Ellison, Glenn. 2013. "How Does the Market Use Citation Data? The Hirsch Index in Economics." *American Economic Journal: Applied Economics* 5(3): 63–90.

Feinberg, Robert M., and Gregory N. Price. 2004. "The Funding of Economics Research: Does Social Capital Matter for Success at the National Science Foundation?" *Review of Economics and Statistics* 86(1): 245–52.

Frank, Robert H., and Philip J. Cook. 1995. *The Winner-Take-All Society.* Free Press.

Freeman, Richard B., Tanwin Chang, and Hanley Chiang. 2005. "Supporting 'The Best and the Brightest' in Science and Engineering: NSF Graduate Research Fellowships." NBER Working Paper 11623.

Hafner, Katie, and Matthew Lyon. 1996. *Where Wizards Stay Up Late: The Origins of The Internet.* Simon & Schuster.

Hamermesh, Daniel S. 2013. "Six Decades of Top Economics Publishing: Who and How?" *Journal of Economic Literature* 51(1): 162–72.

Heutel, Garth. 2014. "Crowding Out and Crowding In of Private Donations and Government Grants." *Public Finance Review* 42(2): 143–75.

Hilmer, Michael J., Michael R. Ransom, and Christina E. Hilmer. 2015. "Fame and the Fortune of Academic Economists: How the Market Rewards Influential Research in Economics." *Southern Economic Journal* 82(2): 430–52.

Hungerman, Daniel M. 2005. "Are Church and State Substitutes? Evidence from the 1996 Welfare Reform." *Journal of Public Economics* 89(11–12): 2245–67.

Kealey, Terence. 1996. *The Economic Laws of Scientific Research.* New York: Palgrave Macmillan.

Kleiner, Morris M., and Alan B. Krueger, A. B. 2013. "Analyzing the Extent and Influence of Occupational Licensing on the Labor Market." *Journal of Labor Economics* 31(S1): S173–S202.

Kummerfeld, Erich, and Kevin J. S. Zollman. 2015. "Conservatism and the Scientific State of Nature." *British Journal for the Philosophy of Science*, axv013.

Lee, Christopher. 2007. "Slump in NIH Funding Is Taking Toll on Research." *Washington Post*, May 28.

Maier, Mark, and W. Edward Chi. 2016. "Community College Economics Instruction: Results from a National Science Foundation Project." *Journal of Economic Education* 47(1): 84–88.

Mervis, Jeffrey. 2014. "How Two Economists Got Direct Access to IRS Tax Records." *ScienceInsider,* May 20. http://www.sciencemag.org/news/2014/05/how-two-economists-got-direct-access-irs-tax-records.

National Science Foundation. 2000. *America's Investment in the Future.* National Science Foundation.

Noll, Richard G., and William P. Rogerson. 1997. "The Economics of University Indirect Cost Reimbursement in Federal Research Grants." Available at SSRN: http://papers.ssrn.com/abstract=78786.

Office of Science and Technology Policy. 2015. *Implementation of Federal Prize Authority: Fiscal Year 2014 Progress Report.* US Government, Office of the President. Retrieved from https://www.whitehouse.gov/sites/default/files/microsites/ostp/NSTC/fy14_competes_prizes_-_may_2015.pdf.

Rothenberg, Marc. 2010. "Making Judgments about Grant Proposals: A Brief History of the Merit Review Criteria at the National Science Foundation." *Technology & Innovation* 12(3): 189–195.

Sobel, Dava. 1995. "Longitude: The True Story of a Lone Genius Who Solved the Greatest Scientific Problem of His Time." Walker & Company.

Tabarrok, Alexander. 2011. *Launching the Innovation Renaissance: A New Path to Bring Smart Ideas to Market Fast.* TED Books.

Varian, Hal R., and Christopher Harris. 2014. "The VCG Auction in Theory and Practice." *American Economic Review* 104(5): 442–45.

Wachtel, Howard. 2000. "How the National Science Foundation Funds Research in Economics." *Challenge* 43(5): 20–30.

Wang, Jian, Reinhilde Veugelers, and Paula E. Stephan. 2016. Bias against Novelty in Science: A Cautionary Tale for Users of Bibliometric Indicators. NBER Working Paper 22180.

Williams, Heidi. 2012. "Innovation Inducement Prizes: Connecting Research to Policy." *Journal of Policy Analysis and Management* 31(3): 752–776.

Zimmermann, C. 2015. "On the Need for a Replication Journal." Federal Reserve Bank of St. Louis Paper no. 2015-016. Available at SSRN: http://papers.ssrn.com/abstract=2647280.

Journal of Economic Perspectives—Volume 30, Number 3—Summer 2016—Pages 249–274

Can War Foster Cooperation?

Michal Bauer, Christopher Blattman, Julie Chytilová, Joseph Henrich, Edward Miguel, and Tamar Mitts

Warfare leaves terrible legacies, from raw physical destruction to shattered lives and families. International development researchers and policymakers sometimes describe war as "development in reverse" (for example, Collier et al. 2003), causing persistent adverse effects on all factors relevant for development: physical, human, and social capital. Yet a long history of scholarship from diverse disciplines offers a different perspective on one of the legacies of war. Historians and anthropologists have noted how, in some instances, war fostered societal transitions from chiefdoms to states and further strengthened existing states (Carneiro 1970; Flannery and Marcus 2003; Tilly 1985; Choi and Bowles 2007;

■ *Michal Bauer is Assistant Professor of Economics at CERGE-EI (a joint workplace of Center for Economic Research and Graduate Education and Economics Institute of Czech Academy of Sciences) and Charles University, both in Prague, Czech Republic. Christopher Blattman is Associate Professor of International Affairs and Political Science at Columbia University, New York City, New York, and Faculty Research Fellow, National Bureau of Economic Research, Cambridge, Massachusetts. Julie Chytilová is Assistant Professor of Economics, Charles University, and Researcher at CERGE-EI, both in Prague, Czech Republic. Joseph Henrich is Professor of Human Evolutionary Biology, Harvard University, Cambridge, Massachusetts and Senior Fellow, Canadian Institute for Advanced Research (CIFAR), Toronto, Ontario, Canada. Edward Miguel is Oxfam Professor in Environmental and Resource Economics, Department of Economics, University of California, Berkeley, California, and Research Associate, National Bureau of Economic Research, Cambridge, Massachusetts. Tamar Mitts is a PhD candidate in Political Science, Columbia University, New York City, New York. Their email addresses are bauer@cerge-ei.cz, chrisblattman@columbia.edu, chytilova@fsv.cuni.cz, henrich@fas.harvard.edu, emiguel@berkeley.edu, and tm2630@columbia.edu.*

† For supplementary materials such as appendices, datasets, and author disclosure statements, see the article page at
http://dx.doi.org/10.1257/jep.30.3.249 doi=10.1257/jep.30.3.249

Morris 2014; Diamond 1999). Meanwhile, both economists and evolutionary biologists, in examining the long-run processes of institution-building, have also argued that war has spurred the emergence of more complex forms of social organization, potentially by altering people's psychology (Bowles 2008; Turchin 2016).

In this article, we discuss and synthesize a rapidly growing body of research based on a wealth of new data from which a consistent finding has emerged: people exposed to war violence tend to behave more cooperatively after war. We show the range of cases where this holds true and persists, even many years after war. Until recently, a paucity of individual-level data from conflict and post-conflict societies prevented researchers from systematically exploring the legacies of war on social and political behavior. In the last decade, however, interdisciplinary teams of researchers—mainly in economics, anthropology, political science, and psychology—have begun to design research projects specifically to understand how exposure to war violence affects collective action, fairness, cooperation, and other important aspects of social behavior among populations around the globe.

In case after case, people exposed to war violence go on to behave more cooperatively and altruistically, which we will generally call "prosocial" behavior. Table 1, Panel A illustrates the breadth of evidence, referencing studies involving Sierra Leone, Uganda, and Burundi in Africa, as well as the Republic of Georgia, Israel, Nepal, and many other societies. The data come from individual surveys collected in seven countries, plus one paper with comparable data from 35 European countries. This evidence covers both civil and interstate wars, and includes a wide array of wartime violence experiences, ranging from personal exposure in which individuals themselves were targeted or directly witnessed violence, to more indirect exposure in which family members were killed or injured.

The evidence suggests that war affects behavior in a range of situations, real and experimental. People exposed to more war-related violence tend to increase their social participation by joining more local social and civic groups or taking on more leadership roles in their communities. They also take actions intended to benefit others, such as altruistic giving, in experimental laboratory games. Our meta-analysis also suggests the effects of wartime violence are persistent and fairly consistent across cases. Moreover, we see little systematic difference by the type of violence experienced (including crime victimization, as examined by a related body of studies), or across studies with different empirical strategies. The results appear to hold for men and women, as well as children and adults exposed to violence, and are remarkably similar for both the victims and perpetrators of violence. Finally, the impacts of exposure do not diminish with time; indeed, if anything, the opposite seems to be true.

Violence may also affect in-group prosocial behavior most of all: that is, participation with, and altruism towards, members of one's own village or identity group. Too few studies define "out-groups" consistently (or at all), so this in-group bias remains somewhat speculative. Nonetheless, it and some of the other patterns we observe are consistent with a broad literature on human behavior and evolutionary biology emphasizing that parochial altruism is a widespread evolved response to external threats. The increased local cooperation we document might help to explain why some post-conflict countries experience what seem to be almost miraculous economic and social recoveries. Yet if people become more parochial and less cooperative with out-group members, this behavioral response could also harden social divisions, contribute to conflict cycles, and help explain the well-known pattern that many post-conflict countries soon return to violence.

Understanding the effects of war in all its complexity, including on postwar patterns of individual behavior and institution-building, is of broad importance. Nearly half of all nations in the world have experienced some form of external or internal armed conflict in the past half century (Blattman and Miguel 2010). According to the World Bank, about two billion people live in countries deemed fragile (Burt, Hughes, and Milante 2014). The findings discussed here emphasize that war is not only one of the most consequential forces for economic development and the emergence of state institutions, but also appears to have complex and multifaceted effects on postwar populations, society, and politics.

Case Evidence on the Effects of Exposure to Wartime Violence

To make the discussion more concrete, we begin by highlighting the case of Sierra Leone, a post-conflict society for which there is an unusual wealth of evidence: three studies by three sets of authors, each with different study populations. The Sierra Leone case also illustrates the synergy of diverse measurement and research methods, including survey reports, study of behavior in lab experimental tasks, and observational data.

The Sierra Leone Civil War

A brutal, countrywide civil war afflicted Sierra Leone from 1991 to 2002. The Revolutionary United Front (RUF), a small group of militants who first entered Sierra Leone from Liberia, inspired a violent rebellion which was nominally directed against the corruption and ineffectiveness of the government. The reach and duration of the war were fueled by access to alluvial diamonds and opportunities to loot civilian property. Many communities organized local fighting groups to protect themselves from the violence of the rebels. Neither ethnic nor religious divisions played a central role in this war: both the RUF and the Sierra Leone army were explicitly multi-ethnic. An internationally-brokered peace agreement was signed in 2003 after a large deployment of United Kingdom and United Nations troops. The war killed more than 50,000 civilians and temporarily displaced roughly two million people—nearly half of the country's population. Armed groups mutilated and raped thousands of civilians. Few people escaped some form of assault or other violence. Nonetheless, there was wide variation in the degree of exposure and victimization.

The period since the end of the civil war has seen an almost miraculous recovery. While Sierra Leone remains one of the poorest countries in the world, it has experienced over a decade of peace and has held several rounds of national and local elections, with alternation of political power among the major political parties at the national level. Until the Ebola outbreak during 2014, the local economy had improved in each year since the end of the conflict, often with rapid growth rates and high levels of foreign direct investment (Casey, Glennerster, and Miguel forthcoming).

Table 1
Studies of War Exposure and Cooperation

Paper	*Country*	*Conflict*	*Data collection*	*Sample*	*Comparable survey measures*	*Comparable experimental measures*	*Time since war exposure*	*Published*	*Data available*
A: Papers eligible for the meta-analysis									
1. Annan, Blattman, Mazurana, and Carlson (2011)	Uganda	Lord's Resistance Army (LRA) insurgency (1986–2006)	2005–2007	Representative sample of youth, some of whom were conscripted by LRA; $N = 613$	Groups, community, trust, voting, interest in politics	–	~7 years	✓	✓
2. Bauer, Cassar, Chytilová, and Henrich (2014)	Georgia and Sierra Leone	Georgia: war with Russia over South Ossetia (2008) Sierra Leone: civil war (1991–2002)	Georgia: 2009 Sierra Leone: 2010	Georgia: children; $N = 565$ Sierra Leone: adult population; $N = 586$	Georgia: groups Sierra Leone: groups, community, trust, voting, interest in politics	Both countries: Allocation tasks (mini-dictator games)	Georgia: 6 months Sierra Leone: 8 years	✓	✓
3. Bauer, Fiala, and Levely (2014)	Uganda	Lord's Resistance Army insurgency (1986–2006)	2011	Young men, some of whom were conscripted by LRA; $N = 337$	Groups, community, trust, voting	Trust game	5 years		✓
4. Bellows and Miguel (2006, 2009)	Sierra Leone	Civil war (1991–2002)	2005 and 2007	Nationally representative sample; $N = 10{,}496$	Groups, community, trust, voting, interest in politics	–	3–5 years	✓	✓
5. Blattman (2009)	Uganda	Lord's Resistance Army (LRA) insurgency (1986–2006)	2005–2006	Young men, some of whom were conscripted by LRA; $N = 741$	Groups, community, voting	–	~5 years	✓	✓
6. Cassar, Grosjean, and Whitt (2013)	Tajikistan	Civil war (1992–1997)	2010	Adult population; $N = 426$	Groups, community, trust, voting	Trust game	13 years	✓	✓
7. Cecchi, Leuveld, Voors, and van der Wal (2015)	Sierra Leone	Civil war (1991–2002)	2010	Youth male street football players; $N = 162$	–	Dictator game	8 years	✓	✓
8. De Luca and Verpoorten (2015a)	Uganda	Lord's Resistance Army insurgency (1986–2006)	2000, 2005, 2012	Nationally representative sample; $N = 4{,}671$	Groups, trust	–	12 years	✓	✓
9. De Luca and Verpoorten (2015b)	Uganda	Lord's Resistance Army insurgency (1986–2006)	2000, 2005, 2012	Nationally representative sample; $N = 4{,}671$	Community, voting, interest in politics	–	12 years	✓	✓
10. Gilligan, Pasquale, and Samii (2014)	Nepal	Civil war (1996–2006)	2009–2010	Household heads; $N = 252$	Interest in politics	Dictator game, Trust game, Public goods game	3 years	✓	✓
11. Gneezy and Fessler (2012)	Israel	Israel–Hezbollah war (2006)	2005–2007	Senior citizens; $N = 50$	–	Ultimatum game, Trust game	1 year	✓	✓

12. Grosjean (2014)	35 countries in Europe, the Caucasus, and Central Asia	WWII (1939–45);
		Yugoslav wars (1991–95);
		Kosovo war (1998–99);
		Tajik civil war (1992–97);
		Chechen wars (1994–2009);
		Kyrgyzstan clashes (2010)
			2010	Nationally representative samples; $N = 38{,}864$	Groups, trust, voting, interest in politics	–	5 months–65 years	✓	✓

(continued on next page)

Table 1

Studies of War Exposure and Cooperation *(continued)*

Paper	*Country*	*Conflict*	*Data collection*	*Sample*	*Comparable survey measures*	*Comparable experimental measures*	*Time since war exposure*	*Published*	*Data available*
13. Grossman, Manekin, and Miodownik (2015)	Israel	Israeli–Palestinian conflict (1967+)	2013	Former soldiers who enlisted between 1998–2003 and 2004–2009; *N* = 2,334	Voting, interest in politics	–	1–12 years	✓	✓
14. Rohner, Thoenig, and Zilibotti (2013)	Uganda	Lord's Resistance Army insurgency (1986–2006)	2000 and 2008	Nationally representative sample; *N* = 2,431	Trust	–	8 years	✓	✓
15. Voors et al. (2012)	Burundi	Civil war (1993–2005)	2009	Household heads, *N* = 287	Groups, community, voting	Allocation tasks (social value orientation experiment)	4–6 years	✓	✓
16. Voors and Bulte (2014)	Burundi	Civil war (1993–2005)	2007	Adult population; *N* = 874	Groups, trust	–	4 years	✓	✓
B: Papers ineligible for the meta-analysis									
17. De Juan and Pierskalla (2016)	Nepal	Civil war (1996–2006)	2003	Nationally representative sample; *N* = 8,822	Trust in national government	–	0–7 years	✓	
18. Hartman and Morse (2015)	Liberia	Civil war (1989–2003)	2013	Adult population; *N* ~ 1,600	Willingness to host refugees	–	10 years		
19. Shewfelt (2009)	Indonesia, Bosnia and Hercegovina, United States (Vietnam veterans)	Indonesia: insurgency in Aceh (1976–2005) B&H: civil war (1992–1995) United States: Vietnam war (1955–1975)	Indonesia: 2007 Bosnia: 2006 United States: 1986	Indonesia: *N* = 1,752 Bosnia: nationally representative sample; *N* = 3,580. United States: male Vietnam theater veterans; *N* = 1,171	Indonesia: groups, community, trust, voting Bosnia: groups, voting, interest in politics United States: groups	–	2–11 years		
C: Papers studying other forms of violence									
20. Bateson (2012)	70 countries	Crime victimization	Americas: 2010 Africa: 2008–2009 Europe: 2000 Asia: 2005–2008	Latin America: 39,238 United States and Canada: 3,000. Africa: 27,713 Europe: 17,088 Asia: 16,725	Groups, community, voting, interest in politics	–		✓	✓
21. Becchetti, Conzo, and Romeo (2014)	Kenya	Kenyan crisis, post-election violence (2007–2008)	2010	Nairobi slum-dwellers; *N* = 404	–	Trust game		✓	
22. Hopfensitz and Miquel-Florensa (2014)	Colombia	Colombian conflict (1964+)	2012	Coffee farmers; *N* = 260	Community	Public goods game			
23. Rojo-Mendoza (2014)	Mexico	Crime victimization	2011	Nationally representative sample; *N* = 7,416	Groups, interest in politics	–			

Note: In the comparable survey measures column, "groups" means "social group participation" and "community" means "community leadership and participation."

All three studies from Sierra Leone identified the same essential pattern: plausibly exogenous variation in exposure to war-related violence was associated with greater social participation and prosocial behavior. The earliest study in this literature, by Bellows and Miguel (2006, 2009), analyzed patterns of local collective action and individual political engagement using a large-scale nationally representative survey dataset on more than 10,000 Sierra Leone households gathered three to five years after the conflict's end. To measure exposure to war-related violence, they constructed an index from responses to three questions: Were any members of your household killed during the conflict? Were any members injured or maimed during the conflict? Were any members made refugees during the war? Victimization rates were high; for instance, 44 percent of respondents reported a household member being killed during the conflict. They found that people whose households directly experienced war violence displayed much higher levels of civic and political engagement compared to nonvictims: they were more likely to report attending community meetings (by 6.5 percentage points), to vote in elections (by 2.6 percentage points), to join social and political groups, and to participate in school committees and "road brushing," a local infrastructure maintenance activity.

To move past relying on self-reports of behavior, researchers have also carried out incentivized lab-in-field experimental games in Sierra Leone, in order to more directly assess whether war-related violence causes changes in social preferences or in beliefs about others' behavior, albeit in controlled and artificial situations. This experimental evidence complements observational survey evidence, and thus may contribute to a better understanding of competing theories.

Table 1 summarizes the games that were implemented in each study. Different types of experimental games help to distinguish between different factors. In simple allocation tasks, such as a Dictator game or a Social Value Orientation experiment, decisionmakers anonymously allocate rewards between themselves and another person. Because the recipient is passive and the interaction is one-shot and anonymous, beliefs about the reaction of the other player should not in principle affect sharing decisions. Choice situations in which participants not only maximize their own rewards but also take into account the welfare of recipients are taken as measures of social preferences, such as altruism, inequality aversion, or adherence to social norms.

In a second class of games, including the Ultimatum game or Trust game, the recipient is not passive and choices are made sequentially. These tasks are designed to uncover willingness to reciprocate (by rewarding kind acts and punishing unfair behavior) as well as beliefs about cooperative behavior of others. In an Ultimatum game, the first player is given a sum of money to divide with another player. If the second player accepts the division, then both receive the money; but if the second player rejects the division, neither player receives anything. The second player's choices, in particular, rejections of low offers, reveal whether that second player is willing to sacrifice earnings in order to punish unfair behavior, while beliefs about whether others have such fairness motivations should be reflected in the choices of the first player. In a Trust game, the amount given by the first player to the second player is tripled, and then the second player can decide whether to give some of the money back to the first player. Transfers of the first player reveal trust—that is, beliefs about whether other players will cooperate by returning some of the money—while back transfers made by the second player provide a measure of reciprocity.

Finally, in a Public Goods game, multiple players decide simultaneously (without knowing about the choices of others) whether to contribute to a public good. The private return from contributing is negative, but the total group payoff to contributing is positive because the return to other players combined is substantial. This game thus reveals individual willingness to cooperate or to free ride (that is, hoping that other players will contribute to the public good). The identities of the other players can also vary in these games, in particular by whether players are interacting with those from a group with whom they have some reason to identify, such as an ethnic or social group.[1]

Bauer, Cassar, Chytilová, and Henrich (2014) ran various allocation games, sometimes referred to as mini-Dictator games, designed to distinguish selfishness from altruism and inequality aversion, in northwestern Sierra Leone. They experimentally manipulated the identity of an otherwise anonymous recipient to shed light on whether violence increases prosocial behavior only towards people at the local level, or whether the effects on prosocial behavior are more generalized. In the in-group condition, the partner was from the same village as the decision maker, and in the out-group condition the partner was from a "distant village." Compared to nonvictims, people who were directly exposed to conflict-related violence were less selfish (by 23 percentage points) and more inequality averse (by 25 percentage points) towards in-group members eight years after experiencing war-related violence. Effects were especially large among those exposed to violence during their childhood and adolescence. There were no comparable effects on behavior towards out-group members.

Elsewhere in Sierra Leone, once again eight years post-conflict, Cecchi, Leuveld, Voors, and van der Wal (2015) found similar results among young street soccer players (aged 14–31 years) using both experimental and observational approaches. Players made anonymous choices in the Dictator game, and those who had been exposed to more intense conflict-related violence behaved more altruistically towards their teammates (the in-group) but not towards the out-group (their match opponents). Direct observation of behavior during soccer matches also revealed that the more violence-exposed players were more likely to receive a yellow or red

[1] In considering the contribution of these behavioral experiments, an important question is the degree to which links between such measures and the formation of real world institutions and cooperation has been made. Work establishing these links is limited. However, Rustagi, Engel, and Kosfeld (2010) show that communities in Ethiopia with more prosocial individuals, as measured using behavioral games, more effectively form real world cooperatives to monitor forest exploitation, more energetically monitor for free-riders (forest exploiters), and end up cooperating more effectively to manage harvests; these findings hold when the frequency of prosocial individuals is instrumented using the distance from market towns. The results suggest that if these villages were "shocked" (for example, by war) in a way that suddenly increased the frequency of prosocial individuals (as measured by experiments), they might become better at constructing local institutions to address real public goods problems.

(penalty) card during the game, suggesting that a violent conflict not only elevated in-group prosocial behavior but may also have exacerbated out-group antagonism.[2]

A common feature of this body of research—for Sierra Leone and the other studies discussed below—is that analysis is based on a comparison of individuals who suffered different degrees of war violence. These data do not allow the estimation of impacts on society as a whole since no suitable counterfactual exists.

Other Country Cases: Uganda, Burundi, Georgia, Nepal, and Others

Another much-studied country case is Uganda, with six papers listed in Table 1. Blattman (2009) examines the case of northern Uganda, where for 20 years the rebel group the Lord's Resistance Army (LRA) forcibly recruited tens of thousands of young people. The study attempted to account for confounders and other econometric identification concerns, using rebel raiding patterns as a source of plausibly exogenous variation in armed recruitment. The paper used a prewar sample, tracked survivors, and attempted to account for nonsurvivors, reducing concerns about bias due to selective attrition. An average of five years after temporary conscription into the LRA, the experience led to substantial increases in postwar social participation, in this case, self-reported voting and community leadership (though not social group membership).

Studies from other post-conflict societies in Africa and elsewhere have documented similar patterns. Notably, Voors et al. (2012) implemented a Social Value Orientation experiment (similar to a Dictator game) among adults in rural Burundi to study consequences of the 1993–2003 civil conflict there between the Tutsi-dominated army and Hutu rebels. Nine years after the war, individuals who personally experienced war-related violence, or who lived in attacked communities, behaved more altruistically towards neighbors in the experimental tasks, and were also more likely to report being involved in local community organizations.

Bauer, Cassar, Chytilová, and Henrich (2014) conducted an experimental study in the Republic of Georgia that paralleled their Sierra Leone study. The data were gathered among a sample of children six months after the brief August 2008 war with Russia over South Ossetia. As in Sierra Leone, the authors found evidence of differential treatment towards in-group and out-group members: participants who were more affected by the conflict were less selfish and more inequality averse towards in-group members (their classmates) as compared to their less-affected peers, but there was no such effect on behavior towards out-group members.

In a study of Nepalese society, Gilligan, Pasquale, and Samii (2014) found that members of communities with greater exposure to violence during the 1996–2006 civil war between governmental forces and Maoist revolutionaries exhibited greater levels of cooperation when interacting with each other: three years post-conflict, they were more trustworthy in a Trust game, more willing to contribute to the

[2] While not directly comparable due to a lack of data on in-group cooperation, Miguel, Saiegh, and Satyanath (2011) show that professional soccer players (in the major European leagues) who lived in conflict settings as children are also more prone to committing violent card fouls against the opposing team during matches.

common pot in the Public Goods game, and they reported being more active in community organizations.

In Israel, meanwhile, results from Ultimatum and Trust games indicate that living in a society with an active ongoing conflict (the Israel–Hezbollah conflict of 2006) temporarily increased the willingness of senior citizens to punish noncooperators and to reward cooperation (Gneezy and Fessler 2012). An aspect of this study is that it relied on a comparison of choices made before, during, and after the conflict and thus does not account for any time effects that occurred contemporaneously with the conflict.

In a study in Tajikistan, more than a decade after its 1992–1997 civil war, Cassar, Grosjean, and Whitt (2013) explored the effects of war-related violence on trust and cooperation. The war in Tajikistan has been described as a power struggle pitting former communists against a highly fractionalized group of challengers with diverse ideologies (including Islamist groups, ethnic nationalists, and prodemocratic reformers). During this civil war, a complex network of rivalries emerged within local communities during the fighting, often resulting in neighbors fighting neighbors (intragroup conflict). This contrasts with the above-mentioned studies, in which violence was typically perpetrated by people from outside of the affected communities (intergroup conflicts). In experimental games, Cassar, Grosjean, and Whitt (2013) matched subjects with another (anonymous) individual from the same village, and thus with some probability with someone from an antagonistic group. It turns out that the exposure to violence during the civil war was associated with a decrease in trust (measured by the first mover transfers in the Trust game). Interestingly, these negative effects were quite heterogeneous and appear to have depended on the nature of infighting within local communities: effects were particularly negative in regions where opposing groups were residentially intermixed and where local allegiances were thus split, indicating that exposure to violence reduced cooperative behavior when people thought they may interact with members of an opposing group in the conflict. Yet the authors also found evidence of elevated participation in local groups and associations among the war exposed, as in other studies. In the case of local group participation, individuals presumably had some ability to choose with whom they would interact (in contrast to the games, where matching was random), and so this result is also consistent with war exposure raising levels of prosocial behavior towards in-group members, although alternative interpretations remain possible.

The broad pattern of war exposure stimulating greater cooperation also holds in large-scale national surveys across multiple countries. Grosjean (2014) linked comparable nationally representative surveys from the Life in Transition Survey project, which gathered data from 35 countries in central and eastern Europe, the Baltic states, southeastern Europe, the former Soviet Union, and Mongolia in 2010. Nearly 40,000 individuals answered questions about their own and their parents' and grandparents' war exposure, with the relevant recall period covering World War II (1939–1945), as well as the civil wars in the former Yugoslavia (in the 1990s), the Tajik civil war (1992–1997), Chechen wars (1994–2009), and the Kyrgyzstan clashes in 2010. The incidence of World War II exposure was very high: the average proportion of respondents who reported that they or their parents/grandparents were injured or killed was nearly 30 percent overall. Grosjean then focused on within-country variation in exposure to war violence. The results show a positive link between past experiences related to violent conflict and contemporary participation in community groups, collective action, and membership in political parties—although there was also a negative effect on trust in central government institutions.[3]

Disentangling Correlation and Causation

An obvious econometric concern is the possibility that the correlation between war exposure and cooperation is driven by some omitted variable that has a confounding effect, rather than reflecting a causal impact. For instance, more cooperative people might be more likely to participate in collective action, including civil defense forces or armed organizations that represent their groups during wartime, and thus more likely to live in a family that experiences some form of direct war victimization. Or perhaps attackers systematically target people who are likely to be more cooperative in nature, such as leading families or wealthy and influential citizens. If true, statistical tests would overstate the effect of war victimization on later civic participation and social capital. Attrition poses another potential challenge for causal identification if the least prosocial or cooperative people are also more likely to die, migrate, or be displaced and not return home.

Given the impossibility of randomized experiments involving targeted violence, studies in this area have taken various analytical steps to mitigate some of the most worrisome confounders. For example, Bellows and Miguel (2009) use three strategies in their study of Sierra Leone. First, they control for local fixed effects, typically at the village level, thus removing potential regional and local omitted variables, and show that within-village variation in violence exposure helps to explain patterns of within-village cooperation. In some settings, the qualitative evidence suggests violence is relatively indiscriminate in nature within a village, which is supported by statistical tests documenting the weak relationship between observable prewar characteristics and the likelihood of falling victim to violence. Second, the researchers attempt to control for local confounders with an extensive set of prewar characteristics, such as wealth or whether victimized households were more central to local politics. González and Miguel (2015) expand on this issue, discuss limitations of the original Bellows and Miguel (2009) analysis, and present alternative ways of accounting for the possible selection into war violence exposure. Third, they estimate effects among subsamples for which victimization was likely to

[3] Some evidence suggests that the effects of experiencing war-related violence may be more persistent if experienced during childhood and adolescence, in line with a broader literature on critical periods in the formation of preferences and noncognitive skills (Heckman 2006; Almås Cappelen, Sørensen, and Tungodden 2010; Bauer, Chytilová, and Pertold-Gebicka 2014; Kosse, Deckers, Schildberg-Hörisch, and Falk 2014). In Sierra Leone, Bauer, Cassar, Chyilová, and Henrich (2014) find the strongest effects on social preferences among those who were children or adolescents during the civil war. Similarly in Uganda, Bauer, Fiala, and Levely (2014) show that effects are driven mainly by those who soldiered during childhood or early adolescence.

be less systematic: for example, for individuals who were children too young to have been prewar community leaders, or for individuals living in areas where fighters were unlikely to have detailed knowledge of the local area, in which case indiscriminate violence seems more likely.

These three strategies describe nearly every study in our sample. All make some form of a conditional unconfoundedness assumption, and control (where such data exist) for possible confounders. Every war is different, of course, and so there is no universal set of confounders. But each paper makes a plausible case that the remaining variation in violence is largely idiosyncratic. Despite these efforts, none of these empirical strategies can fully eliminate concerns about bias from selection and omitted variables. As we show in the meta-analysis, the results are nonetheless relatively consistent across different studies and approaches to causal identification, arguably generating more confidence that the estimated relationships are causal.

Meta-analysis

The existence of so many new papers tackling the same core question with similar data permits us to formalize some of the cross-paper comparisons with a formal reanalysis.

We identified 23 published and unpublished papers that estimate the effects of violence on social behavior, and report them in Table 1. Of these, 19 focus on war violence (as opposed to violence in the form of crime or during elections) and we focus our analysis on these war-related papers here. Of these, 16 studies meet two additional criteria for our reanalysis: the dependent variable was some measure of social participation, cooperation, or prosociality; and the individual data were available online or from the authors.[4] We perform a meta-analysis of these 16 studies using the original data, calculating the average effect of war violence on cooperation as a weighted mean across studies. The online appendix available with this paper at http://e-jep.org summarizes details of the formal literature search, inclusion and exclusion criteria, and discusses the statistical methods and results in greater detail.

Outcome Measures

Outcome measures vary across studies, and not all outcomes are gathered in every paper. To simplify comparisons, we employ the data from each study to construct a standardized index of outcomes that has a mean of zero and unit

[4] We excluded one paper for which data are unavailable, and excluded two papers that examine behaviors that are not comparable to other studies (such as trust in the national government, or willingness to host refugees). Panel B in Table 1 provides information on these three studies. In addition, we identified four related studies focusing on other types of exposure to violence (such as crime, electoral violence, or displacement) in Panel C. We explored the robustness of our results to including some of these additional studies in the meta-analysis, and find qualitatively similar patterns. The results are available in the online appendix.

standard deviation. The outcome variables generally fall into six categories, as follows (and we summarize them for each study in Table 1):

1) Social group participation. This variable captures participation in local social clubs, sports teams, or community organizations. Some studies report the number of groups in which an individual participates, and we standardize the summed measure. If a study uses a binary indicator for group participation and no data is available for the number of groups, we standardize the binary measure.

2) Community leadership and participation. This variable includes indicators for community leadership and engagement, such as participating in local meetings, volunteering for community work, and/or being a community leader or mobilizer. We sum the available indicators for each study and standardize.

3) Trust. For each study, we sum the available trust variables (such as "How much do you trust members of your village?") and standardize the sum. Since trust in in-group and out-group members might differ, we also create separate variables for these subgroups. We define in-group members as people from the same family, village, class, and ethnic group. Out-group members are classified as individuals from other ethnic groups or parts of the country.

4) Prosocial behavior in experimental games. Measures of prosocial behavior vary by study (see Table 1), ranging from altruistic and inequality-averse behavior in allocation tasks (such as the Dictator game), trust and reciprocal behavior in a Trust game, punishment of unfair offers in an Ultimatum game, and contributions in a Public Goods game. As the scale of each outcome measure varies by game and study, we standardize each outcome, where higher (positive) values correspond to more prosocial behavior. We also distinguish between prosocial behavior toward in-group and out-group members for studies that manipulated the identity of the experimental counterpart accordingly.

5) Voting. This variable measures voting in local and national elections. We sum the number of elections in which participants were registered to vote, planned to vote, or voted, and standardize the summed measure.

6) Knowledge of and interest in politics. This measure combines binary indicators for familiarity with political figures or events and more general interest in a country's politics. For each study, we sum these indicators and standardize the summed measure.

To enhance comparability, as well as address the multiple comparison problem, we also create a summary index of all cooperation measures. In particular, for each study, we generate a mean effect across all available outcomes (following the approach of Kling and Liebman 2004; Kling, Liebman, and Katz 2007), where the indices are calculated from the standardized outcome measures of each study.

Statistical Approach

We replicate each study's original research design, taking the study's identification strategy, measure of violence exposure, control variables, and observation weights at face value.[5] Each study has a different empirical strategy for identifying

[5] There is one small exception to this statement: namely, if a paper uses a continuous measure of violence, we convert it to an indicator for comparability with other studies and ease of interpretation. In the

the impact of war violence exposure, and as noted above, most papers assume conditional unconfoundedness—namely, that after adjusting for any observed variables (including location fixed effects in many cases) that would help to determine violence, the remaining exposure to violence can be treated as random.

Violence is rarely truly random, of course, and not all the plausible determinants of violence are observed. Thus, the plausibility of the econometric identification assumptions vary from paper to paper, and these causal claims must be taken with some caution. To analyze this issue more systematically, we code studies by their analytical approach, and document the details in the online Appendix. For example, some studies possess prewar data on victims, some have a long list of "substantive" control variables that go beyond basic demographics to control for the specific confounders (such as wealth or status) that arguably could drive victimization risk.

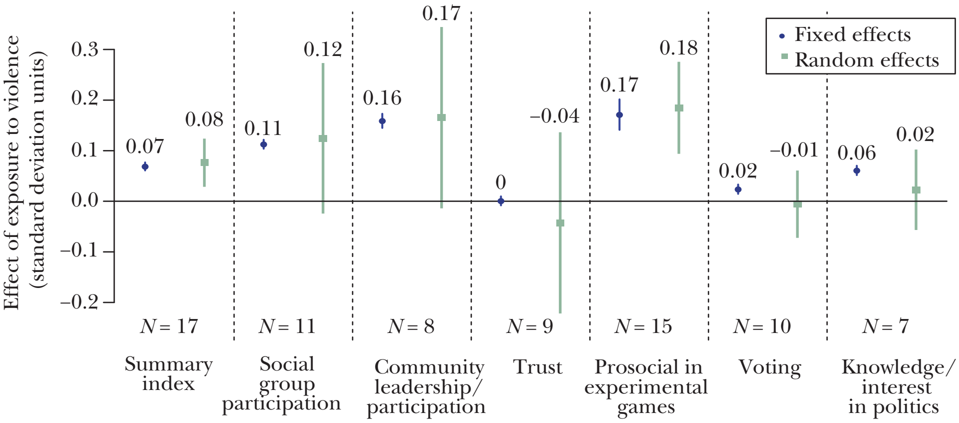
First, however, we estimate overall effects of violence on prosocial behavior. We use both fixed effects and random effects models for this meta-analysis, though note that this terminology has a somewhat different meaning in a meta-analysis than it would when referring to the use of fixed or random effects in a regression model in a single study. In a meta-analysis, a fixed effect refers to whether the effects of the independent variable are indicative of a single stable underlying parameter, while a random effect allows the effect of the variable to differ across contexts in possibly idiosyncratic ways. To put it another way, a fixed effect meta-analysis model is based on the assumption that there is a common effect across all the studies, and thus effectively assumes that studies are drawn from the same population, with larger sample studies thus receiving much more weight in the analysis. In contrast, random effects models allow the true effect magnitude to vary across studies, perhaps because the nature of war violence effects is context-specific. In this case, the studies included in the meta-analysis are simply thought of as a sample from the broader distribution of effects, and smaller sample studies receive relatively more weight than they do in the fixed effects meta-analysis.

In this meta-analysis, the random effects model is arguably preferable on conceptual grounds, since the nature and effects of war violence are likely to be heterogeneous across contexts, but we also report the results of fixed effects approaches, as is common in the related meta-analysis literature, in order to assess robustness to statistical modeling assumptions.[6] Below we also explicitly model the heterogeneity in effect estimates as a function of observed study factors (for example, duration since war exposure), in order to better characterize the nature of context-dependence, something random effects meta-analysis alone is unable to shed light on.

appendix, we also consider alternative independent variables: standardized continuous measures; indicators of the respondent's direct or personal exposure to violence; and indicators of indirect exposure to violence (for example, through the household or community's exposure; these include, for example, having household members killed or injured, or being in a community that was targeted by violence). Results, reported in Appendix Table A17, are qualitatively similar using alternative approaches.

[6] The online Appendix available with this paper also considers a third approach, following Stanley and Jarrell (1989), to include studies without published data. To do so, we use *t*-statistics as a standardized measure of effect size. As can be seen in Table A18, we find qualitatively similar results.

Figure 1
Meta-Analysis Results, War Exposure, and Cooperation



Notes: The figure plots the meta-analysis results reported in Table 2. The effect of exposure to violence on each outcome is estimated using fixed-effects (circles) and random-effects (squares) meta-analysis models. Results are reported in standard deviation units. The vertical lines denote 95 percent confidence intervals. *N* denotes the number of studies/games included in the meta-analysis for each outcome.

Results

Figure 1 displays the average effect of war violence on the standardized indexes, as well as on the overall summary index of all cooperative and prosocial behaviors. There is some variation in the number of studies that capture particular aspects of cooperative behavior, as indicated in the figure, with $N = 17$ studies contributing to the summary index. We present both the fixed- and random-effects average treatment effects with 95 percent confidence intervals. Table 2 reports the corresponding coefficients, standard errors, and *p* values.[7]

Overall, exposure to war violence is associated with a positive and statistically significant increase in the summary index, with a coefficient of 0.07–0.08 standard deviation units and statistical significance for both the fixed effects (p value < 0.01) and random effects (p value < 0.01) approaches. We interpret this as a rejection of the null of no effect, and substantial evidence of positive effects, albeit with only moderate magnitude.

When considering different types of outcomes, the standard errors in the random effects models are much larger than in the fixed effects case, which is not uncommon in a meta-analysis. Precision is increasing in both the number of subjects per study as well as the number of studies, and so the effects are least precise where we have a small number of studies (as in the case of trust). Taken together, there is substantial evidence of an increase in several dimensions of cooperation and

[7] In the online Appendix available with this paper, Figures A4 to A25 present the study-by-study estimates that make up the meta-analysis, for each outcome. The count for the summary index is 17 (and not 16, the total number of analyzed studies) because the Bauer, Cassar, Chytilová, and Henrich (2014) paper has data from two countries, as we thus consider them as two estimates here.

Table 2
Meta-analysis Results: Estimated Population Effects of Exposure to Violence across Studies

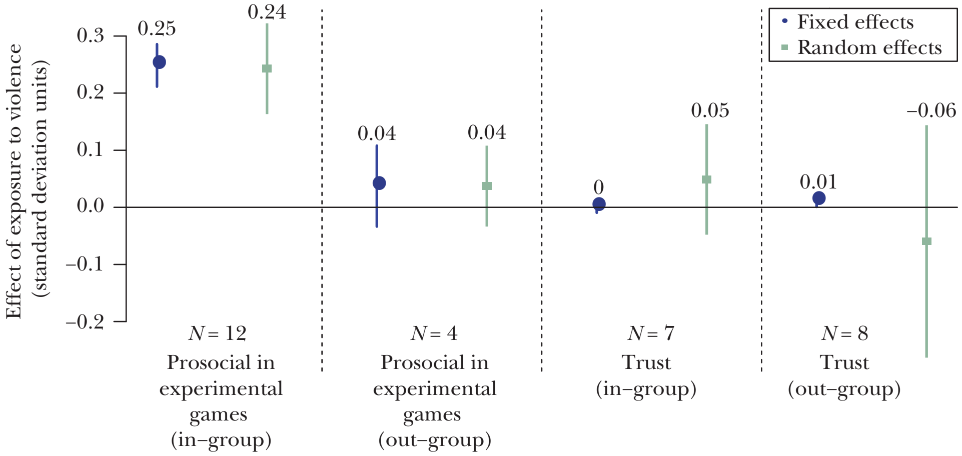
Outcome (Standardized)	*Estimate*	*Fixed effects* (1)	*Random effects* (2)
Summary index (mean effects)	Coefficient	0.07***	0.08***
	Standard error	0.00	0.02
	p-value	< 0.01	< 0.01
Social groups participation	Coefficient	0.11***	0.12
	S.E.	0.00	0.08
	p-value	< 0.01	0.10
Community leadership/ participation	Coefficient	0.16***	0.17*
	Standard error	0.01	0.09
	p-value	< 0.01	0.07
Trust	Coefficient	0.00	–0.04
	Standard error	0.00	0.09
	p-value	0.87	0.64
Prosocial behavior in experimental games	Coefficient	0.17***	0.18***
	Standard error	0.02	0.05
	p-value	< 0.01	< 0.01
Voting	Coefficient	0.02***	–0.01
	Standard error	0.00	0.03
	p-value	< 0.01	0.86
Knowledge/interest in politics	Coefficient	0.06***	0.02
	Standard error	0.00	0.04
	p-value	< 0.01	0.57

Notes: Meta-analysis results for each outcome are reported in the rows. Column (1) reports results from a fixed-effects model; column (2) reports results from a random-effects model. In a meta-analysis, a fixed effect refers to whether the effects of the independent variable are indicative of a single stable underlying parameter, while a random effect allows the effect of the variable to differ across contexts in possibly idiosyncratic ways. The coefficient represents the estimated population effects of exposure to violence across studies, measured in standard deviation units. This analysis excludes exposure to crime violence. ***, **, and * indicate statistical significance at the 1 percent, 5 percent, and 10 percent levels, respectively.

prosocial behavior with exposure to war violence. The fixed effect estimates are positive and statistically significant for participation in social groups, community leadership and participation, prosocial behavior in experimental games, voting, and knowledge of politics (all with p value < 0.01). However, the effect of exposure to war violence on trust is close to zero. The random effect estimates are positive and significant for prosocial behavior in experimental games, and marginally significant for community leadership and participation in social groups, while effects are not distinguishable from zero for the other categories.

In Figure 2, we examine behavior towards in-groups versus out-groups, focusing on the papers and outcomes with appropriate data. For experimental game measures of prosocial behavior, there are positive and significant impacts of war exposure on behavior towards in-group members in both the fixed effect and random effect models, with substantial gains of 0.24 to 0.25 standard deviation units and statistically significant findings (p value < 0.01). In contrast, effects are

Figure 2
Meta-analysis Results, In-Group versus Out-Group Effects



Note: The figure plots the meta-analysis results, broken down by behavior towards in-group and out-group. The effect of exposure to violence on each outcome is estimated using fixed-effects (circles) and random-effects (squares) meta-analysis models. Results are reported in standard deviation units. The vertical lines denote 95 percent confidence intervals. *N* denotes the number of studies/games included in the meta-analysis for each outcome. A meta-regression test for the difference in behavior towards in-group and out-group shows that for games, the difference is significant under both fixed-effects and random-effects model assumptions. For trust, we do not find a significant difference in attitudes towards in-group and out-group members.

smaller in magnitude (at 0.04 standard deviation units) for behavior towards out-group members and not statistically significant in either model. While there is no indication of negative effects towards the out-group, there is significantly less prosocial behavior towards them than towards the in-group. For the stated trust measures, there are no statistically significant effects overall or towards in-group or out-group individuals separately, nor do we find a significant difference between effects on in-group and out-group members, although it is worth recalling that there are relatively few studies with the detailed trust questions needed to undertake this analysis.

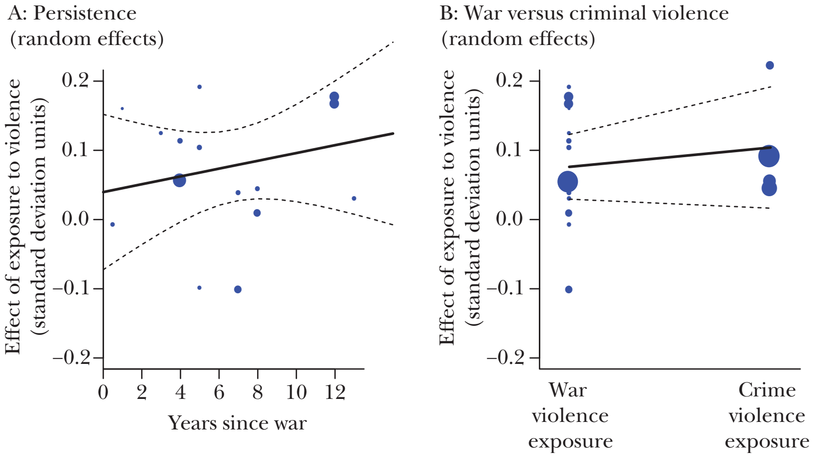
Patterns across Studies

It is informative to examine how circumstances, settings, or study characteristics correlate with the estimated effects of violence on prosocial behavior, although standard errors are relatively wide given the *N* = 17 estimates in hand.

First, we see no evidence that the effects of war violence on prosocial behavior decline over time. We regressed the estimated effect from each study on the length of time between the end of the conflict and the study measures (Table 1 reports the time since war exposure for each study). Figure 3A illustrates the results in a meta-analytic scatterplot; the figure shows the observed effects estimated for individual studies (measured as a standardized index of all cooperation outcomes) plotted against the length of time (in years) between the end of the conflict and the timing

Figure 3

The Effect on Cooperation of War Violence Exposure over Time, and of War versus Crime-Related Violence



Notes: Figure 3A presents the meta-analytic scatterplot of the observed effects estimated for individual studies, where the dependent variable is an index of all cooperation outcomes, plotted against the length of time between the end of the conflict and the timing of each study. Figure 3B plots the observed effects against an indicator of war/crime violence exposure. The point sizes are proportional to the inverse of the standard errors, which means that studies with larger samples tend to have visually larger points. The predicted average effects are included (with corresponding 95 percent confidence intervals), calculated from the random effects meta-analysis model. Grosjean (2014) is dropped from analysis in panel A because of high variability in the years-since-war across the 35 countries studied.

of each study. The resulting regression line has a small positive slope of 0.01 that is not statistically significant in the random effects model (although the fixed effects estimate, reported in the online Appendix, is significant).

Second, in Figure 3B, we compare the war violence studies to data from a study that examines exposure to criminal violence and prosocial behavior in multiple countries (Bateson 2012), and obtain similar average effects. Indeed, the estimated effects from crime studies are, if anything, somewhat larger: the difference in average effect size is 0.03, although it is again not significant in the random effects model. Of course, the difference between war violence and criminal violence is often hard to distinguish, especially if crime involves victims and perpetrators arrayed across a salient social cleavage or carried out by gangs, so some crime incidents could also have an organized intergroup dimension; the data do not allow us to say. But the evidence at least suggests that the "war" aspect may not be at the core of whatever is causing this phenomenon.

Third, the estimated effects are fairly consistent across the various empirical strategies used in the emerging literature. As discussed above, we coded variables that capture different aspects of the research designs, including the use of prewar data, substantive controls, community fixed effects, instrumental variables, and sensitivity analyses. The results, reported in the online Appendix (Figure A3), show that the empirical strategy does not significantly predict variation in the magnitude of the effects across studies in a random effects model, although some study-level covariates are significant in the fixed effects model. For instance, we find that estimates from studies that control for prewar individual covariates are of larger magnitude, and estimates from studies that employ sensitivity analyses are somewhat smaller in magnitude (and these patterns are significant in the fixed effects meta-analysis), although we have not identified a definitive explanation for these differences.

Fourth, we examine whether the way in which violence exposure is measured—on the personal and household level, as opposed to the community, municipality, or district level—might explain the variation in the magnitudes of the effects. We find that studies using measures of personal exposure have smaller coefficients, on average, than studies using more aggregated measures of exposure. We also categorize each study based on whether those exposed to violence were civilians, as opposed to combatants, and find that exposure to violence as a civilian is associated with larger effects.

Theoretical Explanations

The research to date has done a far better job of establishing the effect of war violence on later cooperation than of explaining it. Most papers propose at least one economic, evolutionary, or psychological theory consistent with the observed patterns, but few are able to directly test alternative theoretical predictions of specific models, and the existing pattern of results does not strongly favor any single theoretical perspective. Here, we try to organize the various explanations into a somewhat more coherent conceptual framework.

Changes in Constraints, Economic Payoffs, and Beliefs

Interestingly, almost none of the studies in Table 1 proposes an explanation rooted in the logic of neoclassical economics—that is, an explanation in which social participation or prosocial behavior becomes the optimal choice after war due to the effects of violence on people's economic incentives, constraints, and beliefs. Even so, it is possible that violence affects behavior in this way.

Several economic channels may be relevant. First, greater cooperation may arise from the greater value of social insurance. War frequently destroys household assets, and may make victims of violence more dependent on local informal systems of risk-sharing and insurance, especially among kin and neighbors, thus increasing the return to investments in social capital. Moreover, during wartime, investments in various types of physical and human capital may have been too risky, too constrained, or too expensive relative to investments in social capital. Those most victimized (or most at risk of violence) would thus have an incentive to make larger social capital investments, which could be reflected after war in group memberships, community leadership, and other forms local participation. Second, cooperative behavior could emerge from motives of personal safety and protection. During and after war, property rights and personal security would likely be endangered, and investments in local social capital could be a valuable form of self-protection—for example, in the case of mutual assistance patrols of the neighborhood or village against intruders). It is also possible that the rapid economic recovery many postwar societies experience—such as Sierra Leone after its civil war, or many of the European cases studied in Grosjean (2014) after World War II—could produce the effects we document, if improving economic circumstances tend to generate more social cooperation.

War-related experience may also induce changes in people's beliefs that make prosocial behavior more (or less) persistent. If a sufficiently large number of community members experience the war "shock" at the same time, the entire community could be driven to a more prosocial equilibrium. In this situation, war-affected individuals would appear particularly prosocial soon after the war, but in the long run they would not be distinguishable from the rest of their community because all community members would converge to the new equilibrium. Alternatively, assuming that only a subset of a community experienced the shock of war at the same time, then perhaps the community as a whole does not shift to a new equilibrium. Instead, the prosocial behavior of war-affected individuals might decline over time as their beliefs converge back to the prevailing reality in their communities.

Changes in Parochial Norms and Preferences

Social scientists commonly seek to explain variation in individual social and political activity by pointing to variation in altruism, ethical norms, intrinsic motives to serve the public good, and other "social preferences." Some researchers have suggested that exposure to war-related violence may shape these underlying preferences.

In particular, evolutionary theories suggest that changes due to war violence might lead to favoring one's own group rather than social and political action in general. More specifically, evolutionary researchers from several disciplines have argued that our species' long history of intergroup competition may have favored adaptive psychological responses that promote the success of an individual's group relative to other groups—especially relative to antagonistic out-groups (Alexander 1987; Boyd, Gintis, Bowles, and Richardson 2003; Darwin 1871[1981]; Henrich 2004). This idea has spurred two theoretical variants, one rooted in purely genetic evolution, and a second that considers the interaction between cultural and genetic evolution.

In the purely genetic version, intergroup competition directly favors prosociality toward in-group members and the derogation of those in competing groups (Bowles 2006; Choi and Bowles 2007; Haidt 2012; Wilson 2012). The prediction from this approach is that intergroup competition—and especially war, an extreme form of such competition—will increase individuals' prosocial behavior toward in-group members. These effects are expected to shift people's social preferences—their intrinsic motivations—to make them more parochially prosocial.

In the culture-gene coevolutionary variant, intergroup competition favors cultural practices in the form of social norms or institutions that promote success in intergroup competition (Henrich and Boyd 2001; Richerson and Boyd 2001). Meanwhile, operating within groups, natural selection favors psychological reactions that motivate stronger adherence to these local social norms, institutional practices, and cultural beliefs in favor of culturally defined in-groups. This psychological response to intergroup competition is favored because cultural evolution has long selected cooperative combinations of norms, institutions, and beliefs—so greater norm adherence, including a greater willingness to punish norm-violators, should promote competitive success.

To the degree that local norms prescribe cooperative behavior, individuals more exposed to intergroup competition—including war—should reveal greater prosociality. Since norms are eventually internalized as motivations (or preferences), this approach predicts a shift in preferences similar to that noted above for the purely genetic version. However, unlike in the genetic version, this war exposure could also increase adherence to other norms: for example, if local social norms derogate homosexuality, favor attendance at religious rituals, or promote belief in a particular god, then more war-exposed individuals also ought to be more inclined to derogate homosexuality, attend rituals, and believe in the relevant deity (Henrich 2016).

To study changes in parochial norms and preferences, it is essential to assess what the relevant in- and out-groups are. For example, the experience of a civil war that pits one ethnic group against another might strengthen coethnic prosociality, while corroding the between-ethnic group social capital that could be necessary for later nation-building. Conversely, the experience of an external aggressor attacking a population that already possesses a national identity might bond that entire population even more tightly together and potentially enhance the opportunities for constructing effective national-level institutions in the postwar period. In both cases, and more speculatively, war experience would harden people's parochial prosociality, but the downstream consequences for social stability might depend on how the in-group is interpreted, and what role the relevant out-group plays in social and political life going forward.

Changes in General Preferences and Other Psychological Explanations

A final set of theories and articles propose that preferences for participation and prosociality shift more generally, rather than for or against a particular group. For example, there is substantial evidence that war violence is linked to symptoms of depression and distress, which include a general malaise and lack of desire to engage with people, avoidance of places or people that remind one of the traumatic event, difficulty in maintaining close relationships, an inability to experience positive emotions, negative feelings about oneself or others, and hopelessness about the future (Ehlers and Clark 2000; Galovski and Lyons 2004). Most victims of wartime violence do tend to recover from these symptoms with time, but an important minority continues to experience moderate to severe symptoms for many years, or even the rest of their lives. When people speak of the harmful effects of war on social and political activity, they often have this kind of lasting psychological damage in mind. What is striking is that, in spite of the well-documented effects of violence on distress and depression for some individuals, the emerging empirical evidence reveals an increase in average cooperation and community participation.

Along the same lines of generalized preference change, other psychologists have documented the opposite reaction to violence, a phenomenon they have labeled "post-traumatic growth." Working with the survivors of serious accidents, rape, or other near-death experiences, psychologists have noted that some people respond to trauma by reflecting on and reevaluating their lives, especially in terms of what they regard as important and valuable, such as family and relationships; this research is based largely on case studies. For instance, some victims report a greater valuing of life, more meaningful relationships with others, greater personal hardiness, a realization of new possibilities, and increased spirituality (Tedeschi, Park, and Calhoun 1998; Tedeschi and Calhoun 2004). After war violence, it is possible to imagine victims changing their priorities in life and placing renewed value on relationships with family and community, and even changing other-regarding preferences. Such changes need not be parochial in nature; the existing literature in this area is silent on this point.

Yet another perspective on preference change comes from the political science literature on rebellion. Some ethnographers studying who joins rebel movements (and why) have argued that the experience of injustice, particularly war-related violence, increases individual preferences for collective action. Wood (2003), studying insurgents in El Salvador, noted that people tended to join or support the rebel movement in response to government violence against them or their family members. She argues that material considerations (such as destruction of property or aspirations of land distribution) played little role in who joined. Rather, Wood argues that the injustice of being the subject of violence instilled a "pleasure in agency"—an increase in the intrinsic value in collective action and associational life.

Political scientists use the intrinsic pleasure of participation or expression to explain a variety of behaviors, perhaps most importantly to explain why people expend time and energy to vote, and these intrinsic motives are referred to as "expressive preferences" (for example, Brennan and Lomasky 1997). As with the economists' closely related concept of social preferences, it is not clear what drives these expressive preferences, or how they respond to experience or investment. Some ethnographers have argued that injustices instill a desire for revenge and a pleasure in punitive action (for example, Petersen 2001). Wood's (2003) work in El Salvador has powerful parallels to psychological narratives of post-traumatic growth. On the other hand, since the participation Wood observes is inherently parochial, it is possible that these expressive preferences are also sometimes parochial and could have similar evolutionary origins.

What Does the Evidence Suggest?

Each of the above theories is intuitive and plausible, but empirical support is, so far, relatively limited. Nonetheless, the patterns in the emerging literature do weigh against certain interpretations and lend some support to others. Our reading is that the evidence favors the idea that war violence influences individual social preferences or adherence to existing social norms, and there is suggestive evidence that these changes may be parochial in nature.

For instance, several patterns suggest skepticism towards neoclassical economic explanations. First, the evidence from anonymous behavioral games seems to suggest that something beyond a straightforward calculated response to costs and benefits is occurring. Second, some studies document effects even among young children, and children are more likely to be influenced by prevailing norms and social preferences than by economic cost–benefit considerations or constraints. Third, the war violence effects we document endure long after the conflicts end and even when postwar prosperity and security have improved relative to the prewar (or immediate postwar) situation. Finally, if it were simply a matter of postwar household economic circumstances driving cooperation, one might expect that improving living standards driven by external assistance programs would have a similar effect on local cooperative behavior, but there is little evidence of such a relationship. For example, in a randomized controlled trial in postwar Sierra Leone, Casey, Glennerster, and Miguel (2012) show that large amounts of aid increased local incomes and market activity but did not translate into improvement in a wide range of measures of village meeting participation, social capital, and cooperation.

Nor do we see much evidence consistent with the view that a change in beliefs about the behavior of others is key. Such a view would have two empirical implications: first, that behavioral differences between war-affected people and others are driven by possibly ephemeral differences in information and beliefs, and second, there may not be any enduring long-run differences between the war-affected individuals and the rest of the community (although there may be persistent differences between entire communities subjected to war and those that were not, if a new local equilibrium emerges). Yet neither of these is borne out in the data. War-exposed individuals do not expect others to be more cooperative in survey questions on trust, they behave more prosocially even in games in which beliefs about the behavior of others should not matter, and the behavioral differences between more- and less-war-exposed members of the same community are not ephemeral: they appear to last for many years after conflict ends.

There are at least three reasons, meanwhile, to suggest that war violence may lead to changing social preferences. First, several studies document behavioral changes in experiments that were specifically designed to identify social preferences or adherence to social norms, while controlling for other motivations. Second, the body of qualitative studies and case evidence from the political analysis of conflict, described above, documents self-reported changes in preferences following war victimization. Third, several studies document a change in in-group prosociality, but not out-group prosociality—a form of social preference change predicted by the theory.

Ultimately, there is still insufficient evidence to conclude decisively in favor of one theory over another, but the generation of such evidence is a clear direction for future research.

Conclusion

In less than a decade, nearly 20 observational studies have emerged on the same basic question in different settings, 16 of which are sufficiently similar and have publicly available data such that they can be jointly reanalyzed. This in itself is a striking accomplishment: not only did a few provocative early papers promote a flurry of replications and extensions around the world, but in nearly every case the data have been made freely available online or shared with us directly by the authors, even for unpublished papers. This replication and openness, and the synthesis it permits here, generate some important and perhaps surprising conclusions about violence, psychology, and the formation of social capital, conclusions that differ in some cases from the arguments in the individual papers themselves.

Most of the papers in this emerging literature agree on one central matter: that the data strongly reject the common view that communities and people exposed to war violence will inevitably be deprived of social capital, collective action, and trust. Across the 16 studies from economics, anthropology, political science, and psychology, the average effect on a summary index of cooperation is positive and statistically significant, if moderate in magnitude.

Looking across many studies, however, systematic patterns emerge which were not readily apparent in any single article. For instance, despite early indications that political behavior might also be as positively affected as prosociality (Blattman 2009), this increase in political engagement is not borne out in several more recent studies (for example, Voors et al. 2012; Cassar, Grosjean, and Whitt 2013; Bauer, Fiala, and Levely 2014). Another example comes from the lab experiments, which more often than not have been showing that the prosociality that emerges is focused on in-group interactions but not on behavior towards out-groups. This evidence for parochial altruism, while preliminary, matters because war might enhance intragroup cooperation and facilitate post-conflict reconstruction while simultaneously raising the risk of future social divisions and renewed intergroup conflict.

The most important next step will be for researchers to focus on establishing the reach and generality of this parochial altruism finding. Does it withstand scrutiny, and can we decisively rule out generalized changes in prosocial preferences, or more standard economic arguments? This necessitates a sharper focus on behaviors towards out-group members that belong to the antagonistic group in the war, which is not the case in most existing studies.

Another important direction is to examine other forms of physical insecurity, including crime, state repression, natural disaster, life-threatening accidents, and domestic abuse. In particular, the distinction between wartime violence and urban crime may not be large in certain cases, especially where widespread organized crime takes on characteristics of civil conflict, such as the cases of Mexican or Colombian drug trafficking organizations. Early evidence does indeed suggest that our findings on violence and cooperation could generalize to a wider range of situations. The meta-analysis finds that those who have experienced crime-related violence are also more likely to display cooperative behavior, just like war victims. There are parallels in related literatures, including findings that victims of crime are more likely to participate in community and political meetings, be interested in politics, and engage in group leadership (Bateson 2012). Other emerging evidence exploring the effects of post-election violence (Becchetti, Conzo, and Romeo 2014), and earthquake and tsunami damage (Caló-Blanco et al. 2015; Cassar, Healy, and Von Kessler 2011; Rao et al. 2011) also mimics the main finding of this paper, namely that survival threats tend to enhance local cooperation. We expect that work in these areas will yield new insights about what psychological, economic, and social mechanisms could lead those who experience violence to shift to more cooperative behavior.

The core empirical finding we identify—that exposure to wartime conflict fosters cooperative behavior—resonates with the experience of rapid postwar political, social, and economic recovery in many war-torn societies, as well as their tendency to implement egalitarian social policies, including progressive taxation and gender equality reforms (Tripp 2015; Scheve and Stasavage 2010, 2012). While the human costs of war are horrific, there may at least be some reason for optimism once the violence ends.

■ *We have received valuable comments from Gordon Hanson, Ulrike Malmendier, Enrico Moretti, Suresh Naidu, and Timothy Taylor. JH thanks CIFAR. MB and JC thank the Czech Science Foundation (Bauer P402/12/G130, Chytilová 13-20217S) and the People Programme (Marie Curie Actions) of the EU's Seventh Framework Programme under REA grant (609642). We also thank the many scholars who generously shared their data, sometimes prepublication.*

References

Alexander, Richard D. 1987. *The Biology of Moral Systems.* New York: Aldine De Gruyter.

Almås, Ingvild, Alexander W. Cappelen, Erik Ø. Sørensen, and Bertil Tungodden. 2010. "Fairness and the Development of Inequality Acceptance." *Science*, May 28, 328(5982): 1176–78.

Annan, Jeannie, Christopher Blattman, Dyan Mazurana, and Khristopher Carlson. 2011. "Civil War, Reintegration, and Gender in Northern Uganda." *Journal of Conflict Resolution* 55(6): 877–908.

Bateson, Regina. 2012. "Crime Victimization and Political Participation." A*merican Political Science Review* 106(3): 570–87.

Bauer, Michal, Alessandra Cassar, Julie Chytilová, and Joseph Henrich. 2014. "War's Enduring Effects on the Development of Egalitarian Motivations and In-Group Biases." *Psychological Science* 25(1): 47–57.

Bauer, Michal, Julie Chytilová, and Barbara Pertold-Gebicka. 2014. "Parental Background and Other-Regarding Preferences in Children." *Experimental Economics* 17(1): 24–46.

Bauer, Michal, Nathan Fiala, and Ian V. Levely. 2014. "Trusting Former Rebels: An Experimental Approach to Understanding Reintegration after Civil War." CERGE-EI Working Paper 512. Available at SSRN: http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2417617.

Becchetti, Leonardo, Pierluigi Conzo, and Alessandro Romeo. 2014. "Violence, Trust, and Trustworthiness: Evidence from a Nairobi Slum." *Oxford Economic Papers* 66(1): 283–305.

Bellows, John, and Edward Miguel. 2006. "War and Institutions: New Evidence from Sierra Leone." *American Economic Review* 96(2): 394–99.

Bellows, John, and Edward Miguel. 2009. "War and Local Collective Action in Sierra Leone." *Journal of Public Economics* 93(11–12): 1144–57.

Blattman, Christopher. 2009. "From Violence to Voting: War and Political Participation in Uganda." *American Political Science Review* 103(2): 231–47.

Blattman, Christopher, and Edward Miguel. 2010. "Civil War." *Journal of Economic Literature* 48(1): 3–57.

Bowles, Samuel. 2006. "Group Competition, Reproductive Leveling, and the Evolution of Human Altruism." *Science* 314(5805): 1569–72.

Bowles, Samuel. 2008. "Being Human: Conflict: Altruism's Midwife." *Nature*, November 20, 456(7220): 326–27.

Boyd, Robert, Herbert Gintis, Samuel Bowles, and Peter J. Richerson. 2003. "The Evolution of Altruistic Punishment." *Proceedings of the National Academy of Sciences* 100(6): 3531–35.

Brennan, Geoffrey, and Loren Lomasky. 1997. *Democracy and Decision: The Pure Theory of Electoral Preference.* Cambridge University Press.

Burt, Alison, Barry Hughes, and Gary Milante. 2014. "Eradicating Poverty in Fragile States: Prospects of Reaching The 'High-Hanging' Fruit by 2030." Policy Research Working Paper 7002, World Bank. http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2479667.

Caló-Blanco, Aitor, J. Kovarik, Frederike Mengel, and J. Gabriel Romero. 2015. "Natural Disasters and Social Cohesion." Unpublished paper.

Carneiro, Robert L. 1970. "A Theory of the Origin of the State." *Science* 169(3947): 733–38.

Casey, Katherine, Rachel Glennerster, and Edward Miguel. 2012. "Reshaping Institutions: Evidence on Aid Impacts Using a Preanalysis Plan." *Quarterly Journal of Economics* 127(4): 1755–1812.

Casey, Katherine, Rachel Glennerster, and Edward Miguel. Forthcoming. "Healing the Wounds: Learning from Sierra Leone's Post-War Institutional Reforms." In *African Successes: Government and Institutions*, vol. 1, edited by Sebastian Edwards, Simon Johnson, and David N. Weil. University of Chicago Press.

Cassar, Alessandra, Pauline Grosjean, and Sam Whitt. 2013. "Legacies of Violence: Trust and Market Development." *Journal of Economic Growth* 18(3): 285–318.

Cassar, Alessandra, Andrew Healy, and Carl Von Kessler. 2011. "Trust, Risk, and Time Preferences after a Natural Disaster: Experimental Evidence from Thailand." Unpublished paper.

Cecchi, Francesco, Koen Leuveld, Maarten Voors, and Lizzy van der Wal. 2015. "Civil War Exposure and Competitiveness: Experimental Evidence from the Football Field in Sierra Leone." http://www.tilburguniversity.edu/upload/e4d323d7-eef9-45fc-a589-2ca175bee824_cecchi.pdf.

Choi, Jung-Kyoo, and Samuel Bowles. 2007. "The Coevolution of Parochial Altruism and War." *Science* 318(5850): 636–40.

Collier, Paul, V. L. Elliott, Håvard Hegre, Anke Hoeffler, Marta Reynal-Querol, and Nicholas Sambanis. 2003. *Breaking the Conflict Trap: Civil War and Development Policy.* World Bank Policy Research Report. World Bank and Oxford University Press.

Darwin, Charles. 1871 [1981]. *The Descent of Man, and Selection in Relation to Sex.* Princeton University Press.

De Juan, Alexander, and Jan Henryk Pierskalla. 2016. "Civil War Violence and Political Trust: Microlevel Evidence from Nepal." *Conflict Management and Peace Science* 33(1): 67–88.

De Luca, Giacomo, and Marijke Verpoorten. 2015a. "Civil War and Political Participation: Evidence from Uganda." *Economic Development and Cultural Change* 64(1): 113–41.

De Luca, Giacomo, and Marijke Verpoorten. 2015b. "Civil War, Social Capital and Resilience in Uganda." *Oxford Economic Papers* 67(3): 661–86.

Diamond, Jared. 1999. *Guns, Germs, and Steel: The Fates of Human Societies.* W. W. Norton & Company.

Ehlers, Anke, and David M. Clark. 2000. "A Cognitive Model of Posttraumatic Stress Disorder." *Behaviour Research and Therapy* 38(4): 319–45.

Flannery, Kent V., and Joyce Marcus. 2003. "The Origin of War: New 14C Dates from Ancient Mexico." *PNAS* 100(20): 11801–5.

Galovski, Tara, and Judith A. Lyons. 2004. "Psychological Sequelae of Combat Violence: A Review of the Impact of PTSD on the Veteran's Family and Possible Interventions." *Aggression and Violent Behavior* 9(5): 477–501.

Gilligan, Michael J., Benjamin J. Pasquale, and Cyrus Samii. 2014. "Civil War and Social Cohesion: Lab-in-the-Field Evidence from Nepal." *American Journal of Political Science* 58(3): 604–19.

Gneezy, Ayelet, and Daniel M. T. Fessler. 2012. "Conflict, Sticks and Carrots: War Increases Prosocial Punishments and Rewards." *Proceedings of the Royal Society B: Biological Sciences* 279(1727): 219–23.

González, Felipe, and Edward Miguel. 2015. "War and Local Collective Action in Sierra Leone: A Comment on the Use of Coefficient Stability Approaches." *Journal of Public Economics* 128 (August): 30–33.

Grosjean, Pauline. 2014. "Conflict and Social and Political Preferences: Evidence from World War II and Civil Conflict in 35 European Countries." *Comparative Economic Studies* 56(3): 424–51.

Grossman, Guy, Devorah Manekin, and Dan Miodownik. 2015. "The Political Legacies of Combat: Attitudes Toward War and Peace Among Israeli Ex-Combatants." *International Organization* 69(4): 981–1009.

Haidt, Jonathan. 2012. *The Righteous Mind: Why Good People Are Divided by Politics and Religion.* New York: Pantheon Books.

Hartman, Alexandra C., and Benjamin S. Morse. 2015. "Wartime Violence, Empathy, and Intergroup Altruism: Evidence from the Ivoirian Refugee Crisis in Liberia." http://cega.berkeley.edu/assets/miscellaneous_files/119_-_HartmanMorseViolenceEmpathy-May_2015_-_ABCA.pdf.

Heckman, James J. 2006. "Skill Formation and the Economics of Investing in Disadvantaged Children." *Science* 312(5782): 1900–1902.

Henrich, Joseph. 2004. "Cultural Group Selection, Coevolutionary Processes and Large-Scale Cooperation." *Journal of Economic Behavior & Organization, Evolution and Altruism* 53(1): 3–35.

Henrich, Joseph. 2016. *The Secret of Our Success: How Culture is Driving Human Evolution,* *Domesticating Our Species, and Making Us Smarter.* Princeton University Press.

Henrich, Joseph, and Robert Boyd. 2001. "Why People Punish Defectors: Weak Conformist Transmission Can Stabilize Costly Enforcement of Norms in Cooperative Dilemmas." *Journal of Theoretical Biology* 208(1): 79–89.

Hopfensitz, Astrid, and Josepa Miquel-Florensa. 2014. "Investigating Social Capital in Colombia: Conflict and Public Good Contributions." Unpublished paper, Toulouse School of Economics (TSE). http://citeseerx.ist.psu.edu

Kling, Jeffrey R., and Jeffrey B. Liebman. 2004. "Experimental Analysis of Neighborhood Effects on Youth." John F. Kennedy School of Government, Harvard University, May. http://www.ksg.harvard.edu/jeffreyliebman/klingliebman2004.pdf.

Kling, Jeffrey R., Jeffrey B. Liebman, and Lawrence F. Katz. 2007. "Experimental Analysis of Neighborhood Effects." *Econometrica* 75(1): 83–119.

Kosse, Fabian, Thomas Deckers, Hannah Schildberg-Hörisch, and Armin Falk. 2014. "Formation of Human Prosociality: Causal Evidence on the Role of Social Environment." http://www.sfbtr15.de/fileadmin/user_upload/redaktion/events/Tagung_Bonn_April_2015/C8_Falk_SchildbergHoerisch.pdf.

Miguel, Edward, Sebastián M. Saiegh, and Shanker Satyanath. 2011. "Civil War Exposure and Violence." *Economics & Politics* 23(1): 59–73.

Morris, Ian. 2014. *War! What Is It Good For? Conflict and the Progress of Civilization from Primates to Robots.* Macmillan.

Petersen, Roger D. 2001. *Resistance and Rebellion: Lessons from Eastern Europe.* Cambridge University Press.

Rao, Li-Lin, Ru Han, Xiao-Peng Ren, Xin-Wen Bai, Rui Zheng, Huan Liu, Zuo-Jun Wang, Jin-Zhen Li, Kan Zhang, and Shu Li. 2011. "Disadvantage and Prosocial Behavior: The Effects of the Wenchuan Earthquake." *Evolution and Human Behavior* 32(1): 63–69.

Richerson, Peter J., and Robert Boyd. 2001. "The Evolution of Subjective Commitment to Groups: A Tribal Instincts Hypothesis." *Evolution and the Capacity for Commitment,* edited by Randolph M. Nesse, 186–220. New York, NY: Russel Sage Foundation.

Rohner, Dominic, Mathias Thoenig, and Fabrizio Zilibotti. 2013. "Seeds of Distrust: Conflict in Uganda." *Journal of Economic Growth* 18(3): 217–52.

Rojo-Mendoza, Reynaldo T. 2014. "From Victims to Activists: Crime Victimization, Social Support, and Political Participation in Mexico." Unpublished paper.

Rustagi, Devesh, Stefanie Engel, and Michael Kosfeld. 2010. "Conditional Cooperation and Costly Monitoring Explain Success in Forest Commons Management." *Science* 330 (6006): 961–65.

Scheve, Kenneth, and David Stasavage. 2010. "The Conscription of Wealth: Mass Warfare and the Demand for Progressive Taxation." *International Organization* 64(4): 529–61.

Scheve, Kenneth, and David Stasavage. 2012. "Democracy, War, and Wealth: Lessons from Two Centuries of Inheritance Taxation." *American Political Science Review* 106(01): 81–102.

Shewfelt, Steven Dale. 2009. *Legacies of War: Social and Political Life after Wartime Trauma.* Dissertation, Yale University. http://gradworks.umi.com/33/61/3361597.html.

Stanley, T. D., and Stephen B. Jarrell. 1989. "Meta-Regression Analysis: A Quantitative Method of Literature Surveys." *Journal of Economic Surveys* 3(2): 161–70.

Tedeschi, Richard G., and Lawrence G. Calhoun. 2004. "Posttraumatic Growth: Conceptual Foundations and Empirical Evidence." *Psychological Inquiry* 15(1): 1–18.

Tedeschi, Richard G., Crystal L. Park, and Lawrence G. Calhoun. 1998. *Posttraumatic Growth: Positive Changes in the Aftermath of Crisis.* Routledge.

Tilly, Charles. 1985. "War Making and State Making as Organized Crime." Chap 5 in *Bringing the State Back In*, edited by Peter B. Evans, Dietrich Rueschemeyer, and Theda Skocpol. Cambridge University Press.

Tripp, Aili Mari. 2015. *Women and Power in Postconflict Africa.* Cambridge University Press.

Turchin, Peter. 2016. *Ultrasociety: How 10,000 Years of War Made Humans the Greatest Cooperators on Earth.* Beresta Books.

Voors, Maarten J., and Erwin H. Bulte. 2014. "Conflict and the Evolution of Institutions: Unbundling Institutions at the Local Level in Burundi." *Journal of Peace Research* 51(4): 455–69.

Voors, Maarten J., Eleonora E. M. Nillesen, Philip Verwimp, Erwin H. Bulte, Robert Lensink, and Daan P. Van Soest. 2012. "Violent Conflict and Behavior: A Field Experiment in Burundi." *American Economic Review* 102(2): 941–64.

Wilson, Edward O. 2012. *The Social Conquest of Earth.* New York: Liveright Pub.

Wood, Elisabeth Jean. 2003. *Insurgent Collective Action and Civil War in El Salvador.* Cambridge University Press.

Journal of Economic Perspectives—Volume 30, Number 3—Summer 2016—Pages 275–282

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 e-mail at taylort@macalester.edu, or c/o *Journal of Economic Perspectives*, Macalester College, 1600 Grand Ave., Saint Paul, Minnesota, 55105.

A Selection of Symposia

Earl Pomeroy and Jim McCrery are former Congressmen who in the past headed the Social Security Subcommittee of the House Ways and Means Committee. They have collaborated on producing a volume of 16 essays by a range of authors called *SSDI Solutions: Ideas to Strengthen the Social Security Disability Insurance Program.* Patricia Owens offers "An Overview of Social Security Disability Insurance (SSDI)" focused on the dim financial prospects for the program as currently constituted. The chapters cover proposals about early intervention, program administration, interactions with other programs, structural reform, and international comparisons. In that last category, Robert Haveman discusses "Approaches to Assisting Working-Aged People with Disabilities: Insights from Around the World." Steps taken in other countries to reform disability insurance include: "The introduction of more

■ *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.30.3.275 doi=10.1257/jep.30.3.275

stringent vocational criteria into the eligibility determination process…;" "Instead of relying on an applicant's own doctors, responsibility for assessing capability has been assigned to government agencies;" "Changing the emphasis in the disability pension program toward a 'rehabilitation before benefits' model involving the requirement that benefit applicants have undertaken rehabilitation efforts, as well as requiring employers to pursue workplace accommodation;" "Limiting the duration of disability pension payments to a fixed period (say, three years), with the need to reapply and reestablish eligibility after that period in order to continue benefit receipt;" "Increasing work incentives for benefit recipients through wage or employment subsidies or disregarding earnings in calculating benefits for recipients who combine work and disability benefit receipt." 2016. Published by Committee for a Responsible Federal Budget. At http://ssdisolutions.org/book. This volume is a useful complement to the three-paper "Symposium on Disability Insurance" in the Spring 2015 issue of this journal.

The proceedings of the 14th Bank of International Settlements Annual Conference, held in June 2015 in Lucerne, Switzerland, have now been published. John Kay delivered the keynote address on the subject: "Finance Is Just Another Industry" (BIS Papers no. 84). "We need a finance sector to manage our payments, finance our housing stock, restore our infrastructure, fund our retirement and support new business. But very little of the expertise that exists in the finance industry today relates to the facilitation of payments, the provision of housing, the management of large construction projects, the needs of the elderly or the nurturing of small businesses. The process of financial intermediation has become an end in itself. … High salaries and bonuses are awarded not for fine appreciation of the needs of users of financial services, but for outwitting competing market participants. In the most extreme manifestation of a sector which has lost sight of its purposes, some of the finest mathematical and scientific minds on the planet are employed to devise algorithms for computerised trading in securities which exploit the weaknesses of other algorithms for computerised trading in securities." The fourth of the four papers that follow is by Andrew W. Lo, "Moore's Law vs. Murphy's Law in the Financial System: Who's Winning?" (BIS Working Papers no. 564). "Breakthroughs in computing hardware, software, telecommunications and data analytics have transformed the financial industry, enabling a host of new products and services such as automated trading algorithms, crypto-currencies, mobile banking, crowdfunding and robo-advisors . However, the unintended consequences of technology-leveraged finance include firesales, flash crashes, botched initial public offerings, cybersecurity breaches, catastrophic algorithmic trading errors and a technological arms race that has created new winners, losers and systemic risk in the financial ecosystem. These challenges are an unavoidable aspect of the growing importance of finance in an increasingly digital society. Rather than fighting this trend or forswearing technology, the ultimate solution is to develop more robust technology capable of adapting to the foibles in human behaviour so users can employ these tools safely, effectively and effortlessly." May 2016. Kay's keynote speech and the four other papers are available at http://www.bis.org/publ/bppdf/bispap84.htm.

The *Review of Environmental Economics and Policy* has published a three-paper symposium on the subject of "The EU Emissions Trading System: Research Findings and Needs." The lead paper, by A. Denny Ellerman, Claudio Marcantonini, and Aleksandar Zaklan, discusses "The European Union Emissions Trading System: Ten Years and Counting." "The EU ETS is a classic cap-and-trade system. As of 2014, the EU ETS covered approximately 13,500 stationary installations in the electric utility and major industrial sectors and all domestic airline emissions in the EU's twenty-eight member states, plus three members of the closely associated European Economic Area … The great surprise of the second phase of the EU ETS was that, as phase III started in 2013, the price paid to emit carbon was less than €5, not the €30 or more that had been indicated by 2013 futures prices in 2008 and that was generally expected at that time. This development has created a lively debate about the future of the EU ETS and its role in climate policy. This debate can be summarized as being between those who view the current, much-lower-than-expected price as indicating serious flaws in the EU ETS and those who argue that the low price shows that the system is working exactly as it should given all that has happened since 2008 (i.e., reduced expectations for economic growth in the Eurozone, increased electricity generation from renewable sources, the significant use of offsets), including the possibility that abatement may be cheaper than initially expected. Fundamentally, this debate reflects differing views of the objectives of climate policy itself: whether the objective is solely to reduce GHG [greenhouse gas] emissions or also (and perhaps principally) to transform the European energy system. Although no one is suggesting that emissions have exceeded the cap, or that they will do so, current prices do not seem likely to lead to the kind of technological transformation that would greatly reduce Europe's reliance on fossil fuels." Winter 2016. The three-paper symposium runs from pp. 89–148.

John H. Cochrane and John B. Taylor have edited a collection of six essays on *Central Bank Governance and Oversight Reform.* In Chapter 4, Kevin M. Warsh distills insights from his time as a member of the Federal Reserve Board of Governors and from an insider's perch at the Bank of England to look at the functioning of monetary policy-making committees in "Institutional Design: Deliberations, Decisions, and Committee Dynamics." He points out that successful committees typically don't have too many participants, and those who do participate bring independent information and a willingness to dissent. But in the Fed Open Market Committee, Warsh notes: "By statute, the FOMC includes twelve voting members. … Policy deliberations, however, occur in a much larger institutional setting. Nineteen people convene in the discussion (voters and non-voters alike) and a total of about sixty people are in attendance, including a range of subject-matter experts on key aspects of the economic and financial landscape. While the Reserve Bank presidents are supported by large, independent staffs of economists to help inform their forecasts and policy judgments, I would note that the economic models and forecasting tools are substantially similar across the Federal Reserve System. … By both FOMC tradition and practice, the bar for lodging a dissenting vote is high. Neither Chairman Greenspan nor Chairman Bernanke ever cast a vote in the minority. In contrast, the governor of the Bank of England was outvoted on nine occasions since 1997. … Meade and Stasavage (2008) find evidence that the Fed's post-1993 transcript policy led to deterioration in the quality of FOMC deliberations. … The existence of public transcripts, even with a lag, caused FOMC participants to voice less dissent in the meetings themselves and to be less willing to change policy positions over time. For example, the number of dissenting opinions expressed by voting members fell from forty-eight (between 1989 and 1992) to twenty-seven (between 1994 and 1997)." Hoover Institution. 2016. At http://www.hoover.org/research/central-bank-governance-and-oversight-reform.

Smorgasbord

Indivar Dutta-Gupta, Kali Grant, Matthew Eckel, and Peter Edelman provide an overview of *Lessons Learned from 40 Years of Subsidized Employment Programs.* "Subsidized employment programs have successfully raised earnings and employment. This effect is not universal across programs or target populations, but numerous rigorously evaluated interventions offer clear evidence that subsidized employment programs can achieve positive labor market outcomes. Some of these effects derive from the compensation and employment provided by the subsidized job itself, but there also is evidence that well-designed programs can improve outcomes in the competitive labor market after a subsidized job has ended. … Fundamentally, subsidized jobs and paid work experience programs provide a source of both income and work experience. A number of experimentally-evaluated subsidized employment programs have in turn reduced family public benefit receipt, raised school outcomes among the children of workers, boosted workers' school completion, lowered criminal justice system involvement among both workers and their children, improved psychological well-being, and reduced longer-term poverty; there may be additional effects for some populations, such as increases in child support payment and improved health, which are being explored through ongoing experiments." The report also notes that two major federal government studies involving subsidized employment are now underway: the "Subsidized and Transitional Employment Demonstration (STED), 2010–2017" run by the US Department of Health and Human Services and the Enhanced Transitional Jobs Demonstration study run by the US Department of Labor. Georgetown Center on Poverty and Inequality. Spring 2016. At https://www.law.georgetown.edu/news/press-releases/report-by-georgetown-center-on-poverty-and-inequality-lessons-learned-from-40-years-of-subsidized-employment-programs.cfm.

The UK government set up a Review on Antimicrobial Resistance, funded by the Wellcome Trust and the UK Department of Health, and chaired by Jim O'Neill, which produced the report, *Tackling Drug-Resistant Infections Globally: Final Report and Recommendations.* "An externality is the cost or benefit to a third party for a decision over which they have no control. … Antibiotic consumption fits in this category: individuals take and may benefit from the antibiotics but the resistance to which they contribute impacts all of society. … We estimate that by 2050, 10 million lives a year and a cumulative 100 trillion USD of economic output are at risk due to the rise of drug-resistant infections if we do not find proactive solutions now to slow down the rise of drug resistance. Even today, 700,000 people die of resistant infections every year. Antibiotics are a special category of antimicrobial drugs that underpin modern medicine as we know it: if they lose their effectiveness, key medical procedures (such as gut surgery, caesarean sections, joint replacements, and treatments that depress the immune system, such as chemotherapy for cancer) could become too dangerous to perform. Most of the direct and much of the indirect impact of AMR [anti-microbial resistance] will fall on low and middle-income countries. It does not have to be this way. … The economic impact is also already material. In the US alone, more than two million infections a year are caused by bacteria that are resistant to at least first-line antibiotic treatments, costing the US health system 20 billion USD in excess costs each year." May 2016. At http://amr-review.org.

The Office of Economic Policy at the US Department of the Treasury has published "Non-compete Contracts: Economic Effects and Policy Implications." "Non-compete agreements are contracts between workers and firms that delay employees' ability to work for competing firms. Employers use these agreements for a variety of reasons: they can protect trade secrets, reduce labor turnover, impose costs on competing firms, and improve employer leverage in future negotiations with workers. However, many of these benefits come at the expense of workers and the broader economy. Recent research suggests that a considerable number of American workers (18 percent of all workers, or nearly 30 million people) are covered by non-compete agreements. The prevalence of such agreements raises important questions about how they affect worker welfare, job mobility, business dynamics, and economic growth more generally. This report presents insights from economic theory and evidence on the economic effects of non-compete agreements. It goes on to discuss policy implications, starting a discussion about how such agreements could be used in a way that balances the interests of firms with those of workers and society as a whole." March 2016. Ryan Nunn was principal drafter of the report. https://www.treasury.gov/resource-center/economic-policy/Documents/UST%20Non-competes%20Report.pdf.

A Council of Economic Advisers report discusses *Economic Perspectives on Incarceration and the Criminal Justice System.* "Researchers who study crime and incarceration believe that the true impact of incarceration on crime reduction is small, with a 10 percent increase in incarceration decreasing crime by just 2 percent or less … Additional incarceration may be particularly ineffective in reducing crime when incarceration rates are already high. When incarceration rates are high, further incarceration entails incapacitating offenders who are on average lower risk, which means that their incarceration will yield fewer public safety benefits." "Cost-benefit analyses of incarceration weigh the direct costs of incarcerating an individual against the social value of crimes that may have been averted due to incarceration. Lofstrom and Raphael (2013) examine a 2011 policy change in California that resulted in the realignment of 27,000 State prisoners to county jails or parole. They find that realignment had no impact on violent crime, but that an additional year of incarceration is associated with a decrease of 1 to 2 property crimes, with effects strongest for motor vehicle theft. Applying estimates of the societal cost of crime, the authors calculate that while the cost of a year of incarceration is $51,889 per prisoner in California, the societal value of the corresponding reduction in motor vehicle thefts is only $11,783, yielding a loss of $40,106 per prisoner. Notably, this net loss per prisoner would be larger if the study considered the additional costs of collateral consequences, such as lost earnings or potential increases in re-offending due to incarceration. These estimates highlight the fact that there are more cost-effective ways of reducing crime than incarceration, such as investing in law enforcement, education, and policies that expand economic opportunity." April 2016. https://www.whitehouse.gov/sites/default/files/page/files/20160423_cea_incarceration_criminal_justice.pdf. The report is a useful companion to "Crime, the Criminal Justice System, and Socioeconomic Inequality," by Magnus Lofstrom and Steven Raphael in the Spring 2016 issue of this journal.

Lana Conforti describes "The first 50 years of the Producer Price Index: Setting Inflation Expectations for Today." "March 3, 2016, marked the 125th anniversary of the PPI—one of the oldest economic time series compiled by the federal government. The index, known as the Wholesale Price Index (WPI) until 1978, was established as part of a U.S. Senate resolution on March 3, 1891, the last day of the last session of the 51st U.S. Congress. This Congress was famously known as the 'Billion-Dollar Congress,' because of its expensive initiatives, such as expanding the Navy and creating pensions for families of military members who served in the Civil War. It operated in an era of industrialization, immigration, and economic growth. Two of its most well-known bills were the Sherman Antitrust Act, which sought to protect consumers from certain anticompetitive business practices that tended to raise prices (e.g., monopolies and cartels), and the McKinley Tariff Act of 1890, which raised duties on imports with the goal of protecting domestic industries from foreign competition. Born out of the necessity to measure the impact of such economic policies, the resolution marking the origin of the PPI read thus: 'Resolved, The Committee on Finance be, and they are hereby, authorized and directed, by subcommittee or otherwise, to ascertain in every practicable way, and to report from time to time to the Senate, the effect of the tariff laws upon the imports and exports, the growth, development, production, and prices of agricultural and manufactured articles, at home and abroad….' In response to this resolution, Senator Nelson W. Aldrich, who later played a role in the establishment of the Federal Reserve System, authored a report on *Retail Prices and Wages* in July 1892." *Monthly Labor Review*, published by the US Bureau of Labor Statistics. June 2016. http://www.bls.gov/opub/mlr/2016/article/the-first-50-years-of-the-producer-price-index.htm.

Abhijit Banerjee and Esther Duflo are editing a *Handbook of Field Experiments*, forthcoming from Elsevier. Working paper versions of most of the 17 chapters are posted online at https://www.povertyactionlab.org/handbook-field-experiments. Banerjee and Duflo write: "Taken together, these papers offer an incredibly rich overview of the state of literature. This page collects together all the working paper versions of the chapters, and will also link to the final versions as they become available."

Economies in Africa

African Economic Outlook 2016 is the latest version of an annual report produced by the African Development Bank, the OECD Development Centre and the United Nations Development Programme. The report provides overview of the economic situation in the nations of Africa as well as chapters on the theme of "Sustainable Cities and Structural Transformation." "The African continent is urbanising fast. The share of urban residents has increased from 14% in 1950 to 40% today. By the mid-2030s, 50% of Africans are expected to become urban dwellers … However, urbanisation is a necessary but insufficient condition for structural transformation. Many countries that are more than 50% urbanised still have low-income levels. Urbanisation per se does not bring economic growth, though concentrating economic resources in one place can bring benefits. Further, rapid urbanisation does not necessarily correlate with fast economic growth: Gabon has a high annual urbanisation rate at 1 percentage point despite a negative annual economic growth rate of –0.6% between 1980 and 2011. In addition, the benefits of agglomeration greatly depend on the local context, including the provision of public goods. … Congestion, overcrowding, overloaded infrastructure, pressure on ecosystems, higher costs of living, and higher labour and property costs can offset the benefits of concentrating economic resources in one place. These negative externalities tend to increase as cities grow. This is especially true if urban development is haphazard and public investment does not maintain and expand essential infrastructure. Dysfunctional systems, gridlocks, power cuts and insecure water supplies increase business costs, reduce productivity and deter private investment. In OECD countries, cities beyond an estimated 7 million inhabitants tend to generate such diseconomies of agglomeration." May 2016. Available at http://www.africaneconomicoutlook.org.

Finance & Development has published nine readable articles on the theme of "Africa: Growth's Ups and Downs." The lead article by Stephen Radelet, "Africa's Rise—Interrupted?" provides an overall perspective: "At a deeper level, although high commodity prices helped many [African] countries, the development gains of the past two decades—where they occurred—had their roots in more fundamental factors, including improved governance, better policy management, and a new generation of skilled leaders in government and business, which are likely to persist into the future. … Overall growth is likely to slow in the next few years. But in the long run, the outlook for continued broad development progress is still solid for many countries in the region, especially those that diversify their economies, increase competitiveness, and further strengthen institutions of governance. … The view that Africa's surge happened only because of the commodity price boom is too simplistic. It overlooks the acceleration in growth that started in 1995, seven years before commodity prices rose; the impact of commodity prices, which varied widely across countries (and hurt oil importers); and changes in governance, leadership, and policy that were critical catalysts for change." June 2016. Available at http://www.imf.org/external/pubs/ft/fandd/2016/06/index.htm.

Discussion Starters

Peter Sands argues for "Making it Harder for the Bad Guys: The Case for Eliminating High Denomination Notes." "Our proposal is to eliminate high denomination, high value currency notes, such as the €500 note, the $100 bill, the CHF1,000 [Swiss franc] note and the £50 note. Such notes are the preferred payment mechanism of those pursuing illicit activities, given the anonymity and lack of transaction record they offer, and the relative ease with which they can be transported and moved. By eliminating high denomination, high value notes we would make life harder for those pursuing tax evasion, financial crime, terrorist finance and corruption. … To get a sense of why this might matter to criminals, tax evaders or terrorists, consider what it would take to transport US$1m in cash. In US$20 bills, US$1m in cash weighs roughly 110lbs and would fill 4 normal briefcases. One courier could not do this. In US$100 bills, the same amount would weigh roughly 22lbs and take only one briefcase. A single person could certainly do this, but it would not be that discrete. In €500 notes, US$1m equivalent weighs about 5lbs and would fit in a small bag. … It should be no surprise that in the underworld the €500 note is known as a 'Bin Laden.'" Mossavar-Rahmani Center for Business & Government at the Harvard Kennedy School, Working Paper #52, February 2016, https://www.hks.harvard.edu/centers/mrcbg/publications/awp/awp52.

The Johns Hopkins-Lance Commission on Drug Policy and Health, composed of 22 experts from a wide range of disciplines and professions in low-income, middle-income, and high-income countries, has published "Public Health and International Drug Policy." "The war on drugs and zero-tolerance policies that grew out of the prohibitionist consensus are now being challenged on multiple fronts, including their health, human rights, and development impact. … The disconnect between drug-control policy and health outcomes is no longer tenable or credible. … This challenge is significant, because policy responses to drugs negatively affect human lives and human rights and contradict evidence-based public health approaches. As noted by former UN Secretary-General Kofi Annan, 'Drugs have destroyed many people, but wrong policies have destroyed many more.'" Published at the website of the *Lancet*, March 24, 2016. Available (with free registration) at http://www.thelancet.com/journals/lancet/article/PIIS0140-6736(16)00619-X/fulltext.

Compliments of the American Economic Association

Webcasts of Selected Sessions from the 2016 AEA Annual Meeting . . .

Now available on the AEA Website

January 3, 2016

- **Historical Perspectives on Financial Crisis, Banks and Regulation**

  Presiding: *Gary Richardson*

  Crisis and Collapse in the Long Run: Some Microeconomic Evidence
  Raghuram Rajan and Rodney Ramcharan

  What Ends Banking Panics? *Gary Gorton and Ellis Tallman*

  Interbank Markets and Banking Crises: New Evidence on the Establishment and Impact of the Federal Reserve *Mark Carlson and David Wheelock*

  Commercial Bank Leverage and Regulatory Regimes: Comparative Evidence from the Great Depression and Great Recession *Christoffer Koch, Gary Richardson and Patrick Van Horn*

- **United States Economy: Where To From Here?**

  Presiding: *Dominick Salvatore*

  U.S. Macro Policy in the Future *Olivier J. Blanchard*

  Dealing with Long Term Deficits *Martin Feldstein*

  Central Banking: What's Next? *Stanley Fischer*

  How to Restore Equitable and Sustainable Economic Growth in the United States
  Joseph Eugene Stiglitz

  Can We Restart the Recovery All Over Again? *John B. Taylor*

  Discussant: *Dominick Salvatore*

- **AEA/AFA Joint Luncheon: Why Are Money Markets Different?**

  Bengt Holmstrom

  introduced by *Patrick Bolton*

- **Critiquing Robert J. Gordon's Rise and Fall of American Growth (Panel Discussion)**

  Presiding: *Robert Shiller*

  Gregory Clark
  Nicholas Crafts
  Benjamin Friedman
  James T. Robinson

- **Richard T. Ely Lecture "Restoring Rational Choice: The Challenge of Consumer Finance"**

  John Y. Campbell

  introduced by *Robert Shiller*

January 4, 2016

- **AEA Presidential Address: Behavioral Economics: Past, Present and Future**
  Richard Thaler
  introduced by Robert Shiller

- **Gender at Work: Evidence from Experimental Economics**
  Presiding: *Catherine Eckel*
  Knowing When to Ask: The Cost of Leaning In *Christine Exley, Muriel Niederle and Lise Vesterlund*
  A University-Wide Field Experiment on Gender Differences in Job Entry Decisions *Anya Samek*
  Born to Lead? Gender Differences in Incentive Provision and Its Evaluation *Alexandra Van Geen and Olga Shurchkov*
  Gender Differences in Negotiation by Communication Method *Adam Greenberg and Ragan Petrie*
  Stress and the Gender Difference in Willingness to Compete *Thomas Buser, Anna Dreber Almenberg and Johanna Mollerstrom*
  Discussants: *Christine Exley, Anat Bracha, Katherine Coffman and Alexandra Van Geen*

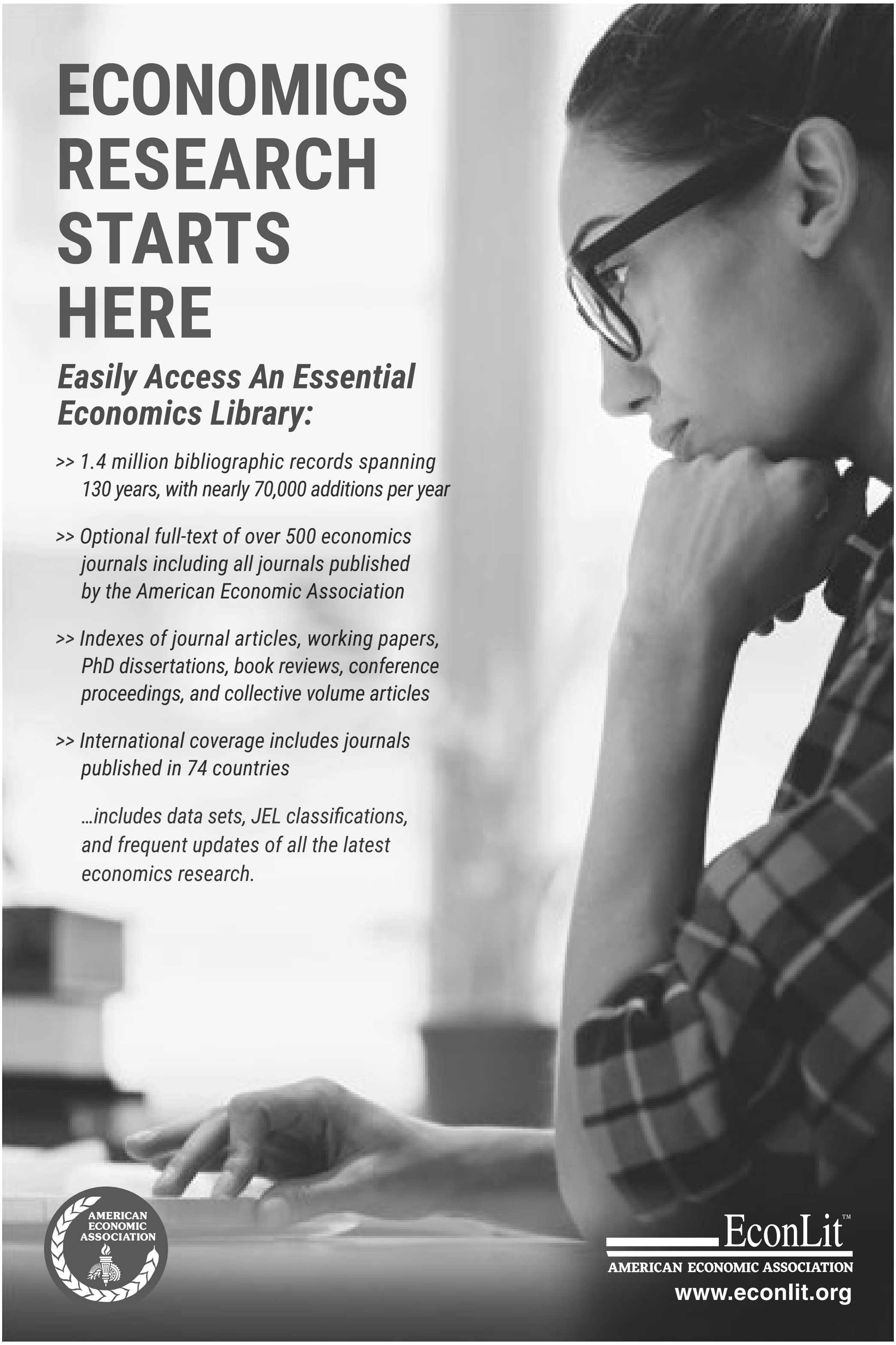
- **60 Million Refugees**
  Presiding: *Robert J. Shiller*
  Refugees, Asylum Seekers and Policy *Timothy J. Hatton*
  Rethinking Protection of those Displaced by Humanitarian Crises *Susan F. Martin*
  What Global Principles Should Govern National Migration Policies? *Jeffrey D. Sachs*
  The Economic Impact of Syrian Refugees on Host Countries: Quasi-Experimental Evidence from Turkey *Semih Tumen*
  Discussants: *Joseph Altonji, George J. Borjas, David Jaeger and Giovanni Peri*

- **AEA Nobel Laureate Luncheon**
  Honoree*: Jean Tirole*
  Presiding: *Robert J. Shiller*
  Roland Benabou
  Drew Fudenberg
  Bengt Holmstrom
  Eric Maskin

- **Digitization and Innovation**
  Presiding: *Shane Greenstein*
  Data in Action: Data-Driven Decision-Making in U.S. Manufacturing *Erik Brynjolfsson and Kristina McElheran*
  Copyright Enforcement in Stock Photography *Hong Luo and Julie Mortimer*
  Agglomeration of Invention in the Bay Area: Not Just ICT *Chris Forman, Avi Goldfarb and Shane Greenstein*
  Discussants: *Kathryn Shaw, Megan MacGarvie and MaryAnn Feldman*

- **AEA Awards Ceremony**
  Robert Shiller

Visit www.aeaweb.org/webcasts/2016
2016 AEA Continuing Education webcasts also available
(AEA Members only)



The American Economic Association

Correspondence relating to advertising, business matters, permission to quote, or change of address should be sent to the AEA business office: aeainfo@vanderbilt.edu. Street address: American Economic Association, 2014 Broadway, Suite 305, Nashville, TN 37203. For membership, subscriptions, or complimentary *JEP* for your e-reader, go to the AEA website: http://www.aeaweb.org. Annual dues for regular membership are $20.00, $30.00, or $40.00, depending on income; for an additional fee, you can receive this journal, or any of the Association's journals, in print. Change of address notice must be received at least six weeks prior to the publication month.

FSC
www.fsc.org
MIX
Paper from
responsible sources
FSC™ C101537

The Journal of
Economic Perspectives

Summer 2016, Volume 30, Number 3

AMERICAN
ECONOMIC
ASSOCIATION