International Journal of Educational Development 80 (2021) 102318

Contents lists available at ScienceDirect

International Journal of Educational Development

journal homepage: www.elsevier.com/locate/ijedudev

Addressing cross-national generalizability in educational impact evaluation

Eric A. Hanushek

Hoover Institution, Stanford University, NBER, CESifo, Stanford, CA, 94305, USA

ARTICLE INFO

Keywords:
RCT
Generalizability
Teacher labor markets
Tracking
TVET

ABSTRACT

Evaluation of educational programs has accelerated dramatically in the past quarter century. With this expansion has come clear methodological improvement involving randomized control studies and other approaches for establishing causation that considerably strengthen their internal validity. Such studies are, however, conducted within individual countries with the institutional structure of the schools and the national labor markets, and they are seldom replicated either within or across countries. A natural question is whether the results of an individual high-quality educational evaluation in one country can be reasonably applied in other countries. This paper focuses on existing research into differences across countries that, while generally impossible to incorporate into program evaluations, potentially have direct effects on key elements of policy and on the outcomes that can be expected. In particular, available cross-national studies on a variety of topics suggest using caution when generalizing evaluation results across countries, because student results are likely to vary systematically with a number of fundamental country-level institutional characteristics that are not explicitly considered in within-country evaluation analyses. Unfortunately, there is currently too little replication of basic research studies to provide explicit guidance on when and where cross-national generalizations are possible.

1. Introduction

Countries around the world exhibit widely different educational outcomes, and these differences have direct implications for future economic performance. A clear understanding of what drives these differences could profoundly improve future economic well-being – a fact that has contributed to dramatic expansion in educational policy and program evaluations. This expansion has coincided with significant improvements in methodology that have strengthened the internal validity of the evaluations but that also have made them more expensive. A natural question arising from this is whether the results of a high-quality evaluation in one country can be reasonably transported to another country. Existing evidence from cross-national studies suggest potentially significant and systematic heterogeneity of impact responses across country because of aggregate institutional differences. Unfortunately, we do not currently have sufficient research in most topical areas to indicate precisely when and where this heterogeneity will be most severe.

There has been an explosion of work from people around the globe attempting to evaluate educational practices in different countries.[1] Increasingly, these evaluations rely on randomized control trials (RCTs) and other analytical approaches designed to provide credible means of establishing causality. Importantly, many of these studies address issues of common policy interest in a wide range of countries. What is the effect of the starting age of schools? What is the effect of different class sizes? How do different forms of school accountability affect achievement? What are the implications of greater school choice for overall student outcomes? These issues have been extensively researched over time, including having a professional association devoted to such study (Comparative and International Education Society, or CIES). But the nature of the discussion and the focus of attention has shifted with the emphasis across social science disciplines in "causality" and the subsequent elevation of RCTs and other causal methods in research funding and in the hierarchy of relevant research approaches.

The overall research strategy has been discussed from a variety of perspectives. The move to a broad portfolio of randomized evaluations is consistent with the arguments of Banerjee and Duflo (2011) and reinforced by the award of the 2019 Nobel Prize in Economics to Abhijit Benerjee, Esther Duflo, and Michael Kremer. Yet, as a general strategy, Pritchett and Sandefur (2013, 2015) question how to interpret results

E-mail address: Hanushek@Stanford.edu.

[1] See, for example, the numbers of researchers at the National Bureau of Economic Research and at the CESifo Research Network focused on the economics of education (https://nber.org/programs/ed/ed-mem.html and https://www.cesifo.org/en/research-network/network-members?search_api_fulltext=&sort_by=lastname&namesort=1&f%5B0%5D=bereich%3A37859&f%5B1%5D=inhaltstyp%3Anetzwerkmitglied.

https://doi.org/10.1016/j.ijedudev.2020.102318
Received 13 July 2020; Received in revised form 12 November 2020; Accepted 13 November 2020
Available online 3 December 2020

from experiments that involve different environmental factors and potential response heterogeneity. Heckman and Smith (1995); Deaton (2010), and Deaton and Cartwright (2018) have voiced broad concerns about over-interpreting the results of RCTs and about the singular focus on RCTs as a research strategy into economic development. And, Ravallion (2020) voices concern, among other things, about the impact of the very expensive RCTs on research budget decisions.

This paper addresses issues of research strategy from the perspective of cross-national generalizability. With the movement toward randomized studies has come an increase in the overall cost of individual studies, making it less likely that specific studies are replicated either within or across countries.[2] As a result, there is a keen interest in being able to apply the results obtained elsewhere. How much of what is known about starting age in the Netherlands should someone take home to the U.S.? Does class size reduction in India mean the same as it does in Germany? Does the growth of randomized control trials (RCTs) in developing countries provide us with relevant conclusions for policy in other developing countries, let alone in developed countries?

Most current policy discussions and related policy evaluations take place within separate countries, and thus within the overall institutional structure of each country. The macro-institutions – institutions that apply to the entire country – cannot themselves be easily addressed within most evaluation studies and yet may interact with the way specific programs impact learning. For example, a single application system for choice of schools may have a very different impact on schooling patterns in the Netherlands (where there is a long history of free choice among alternative school providers) than in the U.S. (where there is more recent and more limited choice through charter schools). Or, allowing for more local decision making in schools may have quite a different impact in the U.S., where there has been a national school accountability system under No Child Left Behind (NCLB) and its successor Every Student Succeeds Act (ESSA), than in a country with no standardized accountability system.

If results from a given evaluation are applied across countries (or even within different parts of the original country), the performance in the secondary application is seldom evaluated or used to judge generalizability of the original study. A notable exception is Duflo et al. (2020), where four interventions developed elsewhere are introduced in a large RCT in Ghana and evaluated. In this case, the interventions partially but not entirely held up in the new country, raising the further issue in accumulating this kind of usage evidence of how to treat marginal impacts of interventions that differ in different settings.[3] Such direct studies of generalizability are nonetheless quite rare.

This paper provides some selective evidence to demonstrate that institutional differences across countries almost certainly condition policy outcomes. It is not an attempt to survey or review the extensive evaluations that currently exist. Instead, it provides specific evidence about potential institutional factors that vary across countries and that touch on key aspects of the generalizability of many educational evaluations. Though it is not entirely clear *how* such institutional differences play into policy outcomes, the examples warn against taking the evaluation results from one country to another without carefully considering how fundamental differences in the schools and environment of different countries may impact policy results. The important structural differences in schools and relevant labor markets across countries will be embedded in the program evaluations and will determine how results can be generalized across countries when the institutional environments differ.

The policy dilemma posed by such uncertainties feeds directly back into the development of research and evaluation strategies. Banerjee and Duflo (2011) make a strong case for developing more credible evaluations based on expanded randomized trials, while Ravallion (2020) comes essentially to a polar opposite conclusion. To be sure, the optimization of research design is a very difficult problem that almost certainly does not have a global solution. Instead, it likely involves the details of individual evaluation situations and of the state of existing research.

2. Some background

There is growing availability of test information about student performance in many countries (Bergbauer et al. (2019)). This testing has been developed and used for a variety of purposes, but one primary purpose is to monitor and manage the educational system in each country. The growth of testing has also led to greater evaluation and research within individual countries, including the rise of field experiments that assess cognitive skills of students. While being controversial in a variety of respects (see Heyneman and Lee (2015); Singer et al. (2018); Berman et al. (2020)), the existence of broad international testing provides the starting point for this discussion.

The Programme for International Student Assessment (PISA) and the Trends in International Mathematics and Science Study (TIMSS) now allow researchers to compare student performance across countries.[4] These data have been increasingly used not only to study how student achievement affects national economic and social outcomes, but also to understand the specific factors that might affect student performance (Hanushek and Woessmann (2011)). These international surveys show not only what performance levels are possible by students but also how widely student performance varies across countries. Thus, they also potentially point to policies that might be improved in order to get higher achievement.

Furthermore, they allow us to analyze things that cannot be analyzed within a single country. Structures like labor market institutions, cultural values, and the enforcement of property rights that are roughly constant within a country make it impossible to analyze their effects on behavior and policies using just within-country data. Cross-national observations, on the other hand, provide indications of how different aggregate factors affect policy outcomes – and implicitly how they affect the generalizability of policy existing evaluations.

The use of cross-country performance data also comes at a cost, because it is necessary to deal analytically with the many ways in which countries might differ and might affect educational outcomes. This trade-off is of course central to this paper and will be discussed below.

3. Achievement differences matter

The main focus of this paper is on the interpretation of country-specific evaluation analyses, but it is useful to begin with a more fundamental issue. Such evaluations frequently use test scores to measure the immediate impact of a policy. Do test scores reflect an important object of policy? Some recent discussions of testing, particularly

[2] The structure of social science publications undoubtedly also enters. The most prestigious journals show little interest in publishing pure replications of prior results, thus lowering the supply of researchers who are willing to invest major amounts of time into repeating studies or into doing minor alterations of prior studies. Additionally, funders are reluctant to repeat a high-quality study that has "already provided the answer" to a specific question.

[3] A different approach involves using a series of similar RCTs across countries; see Lucas et al. (2014); Bando et al. (2019).

[4] The OECD has conducted PISA tests in mathematics, science, and reading every three years since 2000 (OECD (2016)). The PISA testing is now conducted in over 70 countries. The International Association for the Evaluation of Educational Achievement (IEA), provider of the current TIMSS, has published assessments in mathematics and science since the mid-1960s. TIMSS is conducted every four years and covered over 50 countries in 2015 (Mullis et al. (2016)). Indeed, the availability of international testing over the past half century has provided considerable motivation for the extensive research in comparative education.

when used for accountability, have essentially argued that standardized tests are not good indicators of schooling outcomes, that they do not really matter, or that they set up bad incentives (e.g., see the various discussions in Amrein and Berliner (2003); Hout and Elliott (2011); Heyneman and Lee (2015); Koretz (2017)

The most direct way to address this question is to look to how the labor market treats differences in test scores. This approach has not been very common, in part because of having limited test data along with labor market outcomes and in part because the lack of test data has not been viewed as a serious issue. Since the seminal work on human capital by Jacob Mincer (1970, 1974), there has been a focus on the level of schooling and the experience of a worker in characterizing differences in human capital.[5] The availability of these measures, and particularly school enrolment and school attainment, has been ubiquitous and has provided some optimism that workers can be compared within and across countries using common census and survey data. In fact, the reliance on years of schooling as a measure of individual skill differences is so common that the term human capital is almost synonymous with school attainment.

Particularly in an international context, the ubiquitous reliance on measures of school attainment is highly suspect. For the direct cross-country comparisons to hold, one must believe that a year of schooling in Brazil has the same learning and skill content as a year in Portugal. Of course, it is just on this point where the PISA assessments provide direct information. In 2015, the average Brazilian fifteen-year-old was over one standard deviation behind an average fifteen-year-old in Portugal. While these differences have long been recognized, the pragmatic appeal of the use of attainment measures leads to their continued usage, not only within countries but also across countries (e.g., see Psacharopoulos and Patrinos (2018)).

Differences in levels of performance by themselves might not be too damaging if, for example, years of schooling was a good index of the cognitive skill differences found in each country's population.

Unfortunately, as noted this is not the case, yielding significant problems with the common reliance on years of schooling. Work on educational production functions, starting with Coleman et al. (1966), has uniformly shown that families and factors outside of the school have a strong influence on individual achievement and skills (see Hanushek (2002)). Importantly, this work has also shown that the effects of families and schools varies significantly across countries (Heyneman and Loxley (1983); Woessmann et al. (2009)). But focusing just on years of schooling ignores any differences in school quality and severely limits policy discussions.

As data become more available, it is increasingly clear that focusing on school attainment leads to significant distortions in the perceived cognitive skills of a population and that direct measures of skills found in standardized assessments provide superior information – at least as related to economic outcomes. Test scores summarize cognitive skill differences in individuals, regardless of what led to the scores. Consistent with the now extensive literature on educational production functions (Hanushek (1979)), test scores do not solely reflect the quality of schools but also relate to families, peers, and other inputs. Nonetheless, they can be viewed as a direct measure of a large component of human capital differences.

The value of directly measuring cognitive skills can be seen in the economic impacts of differences in test scores. The OECD's Programme for International Assessment of Adult Competencies (PIAAC) ascertains the demographic and labor market experiences of a representative sample of adult workers along with giving them literacy and numeracy tests. Using data for survey takers from PIAAC, one can estimate the earnings returns to individual skills.[6] Hanushek et al. (2015, 2017a, 2017b) provide evidence of strong and statistically significant estimates of the impact of cognitive skills within each of the 32 countries that participated in PIAAC. Individuals with greater measured cognitive skills earn more throughout their lifetime; on average, across countries, a one standard deviation higher test score is associated with a 20 percent higher wage over the lifetime. As discussed below, these returns vary substantially across countries.[7] Within any country, there are wide variations in individual earnings – reflecting a variety of other influences on earnings, but the labor market analyses point to the strong average impacts of cognitive skills.

A second way to recognize the value of standardized assessment measures is to examine how aggregate test scores for countries help to explain differences in long run growth rates of GDP per capita. Over the past three decades, economists have intensively examined how to explain country differences in growth rates. Much of this – similar to the analysis of individual earnings – has focused on how a country's human capital measured by school attainment affects growth (Pritchett (2006); Hanushek and Woessmann (2008)). This focus leads to significant problems (see, for example, Levine and Renelt (1992) or Levine and Zervos (1993))

There is, however, strong evidence that achievement scores, rather than school attainment, better explain differences in long run growth of GDP across countries. Specifically, three-quarters of the variation in growth rates of per capita GDP across 50 countries between 1960 and 2000 can be explained by just the initial GDP levels and the skills of the population measured by international assessment scores (Hanushek and Kimko (2000); Hanushek and Woessmann (2015a)).[8] In contrast, years of schooling by itself can explain just one-quarter of the variation in long run growth rates and is insignificant once learning, as measured by these test scores, is included in the analysis.

Moreover, the relationship between growth rates and cognitive skills is strong enough that, according to historical patterns, improvements in the schools of a country have had huge effects on future economic well-being (Hanushek and Woessmann (2015b)). For example, the difference in international test scores between Canada and U.S. would, according to historical data, yield an increase in U.S. long term growth rate of one percent per year. In short, there is strong justification for focusing on student test scores, which in the aggregate are labelled knowledge capital, in evaluating educational policies.

4. "Case studies" in cross-country institutional features

With this background, it is possible to turn to the challenges of

[5] One on-going discussion surrounding analyses of human capital has focused on whether schools produce greater skills or simply select people with more skills (Spence (1973) or more recently Caplan (2018)). The selection or signaling model is more relevant when human capital discussions are centered on years of schooling or school attainment as opposed to cognitive skills measures where there is ample evidence that schools change outcomes.

[6] The PIAAC surveyed 5,000 or more adults in 32 countries in either 2012 or 2015. (Data for Indonesia, an additional sampled country, included just Jakarta and are not used). See http://www.oecd.org/skills/piaac/.

[7] One standard questions is why people with more education earn more? The simplest answer is not that they are more dexterous or that they can work faster on the production line. It is that they are better able to adapt to changes (Nelson and Phelps (1966); Welch (1970)). One of the first real tests of that hypothesis comes from comparing the rates of return to cognitive skills with the annual growth rate in GDP. Indeed, the faster the growth rate in GDP, where presumably jobs are changing more rapidly, the higher the return to skills (Hanushek et al. (2017)).

[8] Initial GDP levels are included to reflect the fact that countries starting behind can grow fast by copying what countries near the technological frontier do, while technologically leading countries have to invent new production processes and new technologies. There is, of course, debate about whether the impact of differences in cognitive skills are causally related to cross-sectional growth rates. The analysis in Hanushek et al. (2017a) provides strong evidence for a causal interpretation, but open questions remain. And, some have questioned the strength of the relationship (Komatsu and Rappleye (2017)).

generalizing from various country evaluation studies that typically look at policy impacts on test scores.[9] It will be, of course, difficult to arrive at general conclusions about when, where, and to what extent generalizations can be made across the variety of policy evaluations found in different countries. On the other hand, some insights can be gained by looking at evidence on cross-country achievement differences and how they are affected by macro-institutional factors.

The thought experiment is simple: if armed with a study with high internal validity, say from a well-structured RCT or a particularly compelling natural experiment, what conclusions can be transferred to policy in a different country? This issue has become increasingly relevant because field experiments tend to be quite expensive, particularly when done in high income countries. Thus, for a variety of reasons, relatively more experiments have been conducted in developing countries where they are easier and cheaper to run. But also, because of the expense, they are seldom replicated in different settings.

This paper considers a set of "case studies" that assess how macro-institutional factors interact with specific aspects of country school systems in affecting student outcomes. These case studies are neither representative nor exhaustive. They arise directly out of various recent analyses that provide causal evidence on how major institutional features of the schools in different countries interact with overall student outcomes. This accumulating evidence about the importance of a range of macro-institutional factors motivates this consideration of generalizability.

The macro-institutional factors examined through case studies are: use of testing; policies of local school autonomy in decision making; varying country labor markets for skilled workers; overall country differences in the selection of teachers; emphasis on vocational versus general education; and early tracking in schools. The intention is not to provide the details behind each of the case studies. Instead, the aim is to summarize their results and discuss how macro-institutional factors relate to the sought after marginal policy impacts, and thus enter into how generalizable specific evaluation efforts are actually.

Importantly, these macro-institutional factors will not affect all of the many micro-evaluations equally, since programmatic elements will interact most severely with varying background factors. Also, while this discussion will not address within-country heterogeneity, many of these issues could interact with attempts to generalize evaluations within countries. This possibility is most obvious in the case of federalist countries like Brazil, Germany or the U.S. where the individual states drive most of education policy and both affect background institutional features and are affected in turn by them. These issues are, however, beyond the scope of this paper.

4.1. Case study 1: testing[10]

The previous discussion described the significance of test scores in explaining economic outcomes, but testing itself is the subject of policy discussions and of research. Student testing, which comes in a variety of forms across countries, provides information that can be used in different ways. It can, for example, support accountability systems, be used to compare school performance, be used to assess teacher performance, or become the basis of student promotion and placement. While the multiple uses of tests are not mutually exclusive, it is possible to sort out the impacts of the several major alternatives.

The extent and purpose of student testing have become areas of heated debate in many countries, both developed and developing. Some express the view that high-stakes tests – meaning assessments that enter into reward and incentive systems for some individuals – are inappropriate (Koretz (2017)). Others argue that testing and accountability systems are essential for the improvement of educational outcomes (World Bank (2018)) and, by extension, for the improvement of economic outcomes (Hanushek and Woessmann (2015a); Hanushek et al. (2015)). As both national and international testing enters more deeply into decision making, it also becomes more subject to controversy and to public discussion (Heyneman and Lee (2015); Finn and Hanushek (2020)).

Most applications of student assessments have not been adequately evaluated, largely because testing has been introduced in ways that make identification of impacts difficult. National testing programs often lack suitable comparison groups, creating fundamental analytical issues. A key question is, when are student assessments used in ways that promote higher achievement?

The six waves of the PISA assessments between 2000 and 2015 permit country-level panel estimation that relies on within-country, over-time analysis of country changes in assessment practices. Bergbauer et al. (2019) combine data across 59 countries to estimate how varying testing situations and applications affect student outcomes.[11] The results indicate that using standardized tests to compare outcomes across schools and students produces greater student outcomes (as measured by PISA scores) than those systems that simply report the results of standardized tests. They also produce greater achievement results than systems relying on localized or subjective information that cannot be readily compared across schools and classrooms. Systems that use tests to evaluate teachers have little or negative impact on student achievement.

Moreover, information pertaining both to schools and to students results in greater student learning (i.e., higher national PISA scores). General comparisons of standardized testing at the school level appear to lead to somewhat stronger results than direct rewards to students that come through sorting across educational opportunities and subsequent careers. However, rewards to both are significant.

Most interestingly from an international perspective is the finding that country-level testing is most important for school systems that are performing poorly. It appears that school systems with strong testing results know more about how to boost student performance and are less in need of strong external information systems. Fig. 1 shows confidence intervals for the estimated impact of the different kinds of accountability systems as a function of the initial levels of student performance. This figure indicates that standardized external comparisons have declining impacts according to the overall level of country achievement, and that the impacts cease being significantly different from zero at about 500 points, the mean student score on PISA for the OECD countries.

Comparative testing appears to create incentives for better performance and allows rewarding those who are contributing most to educational improvement efforts. It may thus interact directly with other educational policies. The interaction of testing with the overall functioning of the school system suggests that any interactions of educational policies with testing and accountability policies may also differ markedly with the macro-institutional structure of the schools. Yet, in terms of assessing policy evaluations across countries, any such interactions with accountability policies are generally not identified in the policy evaluations. Much of testing holds for entire nations and

[9] For the reasons just discussed, evaluation studies that focus on years of schooling, school completion rates, and the like have obvious limitations when one considers generalizing across countries, and such studies are not considered here.

[10] The underlying analysis for this section can be found in Bergbauer et al. (2019).

[11] The analysis in Bergbauer et al. (2019) uses country-level panel analysis, extensive micro-level controls, and direct analysis of alternative country policies to separate the causal impact of testing regimes as opposed to other country differences. The estimates, which rely upon country changes in testing policies over time, thus have a plausible causal interpretation

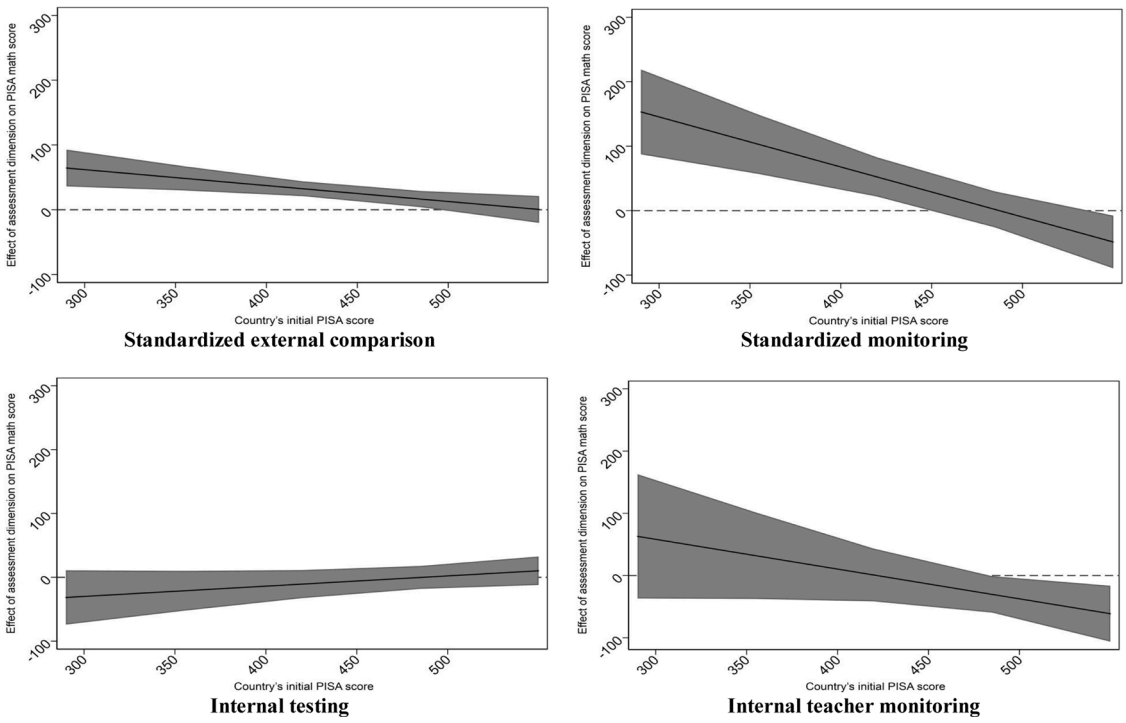

Fig. 1. Effect of student assessments on math performance by initial achievement levels.
Notes: Average marginal effects of student assessments on PISA math score by initial country achievement, with 95 percent confidence intervals.
Source: Bergbauer et al. (2019).

cannot be readily extracted from evaluation results.[12]

4.2. Case study 2: local autonomy in decision making[13]

Local autonomy has been a policy discussed intensively in both developing and developed countries. Interestingly, while many countries have decentralized processes such as the hiring of teachers or the choosing of curriculum over time, others have actually made decision-making more centralized. As described below, these heterogeneous policies may itself reflect the complicated research results.

Autonomy in school decision making 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 may in fact hurt student achievement in low-performing systems that lack basic standards and local capacity. Existing micro-studies of autonomy in decision making include a variety of randomized studies, but there still exists considerable variation in results. Reviews by Patrinos (2011) and Galiani and Perez-Truglia (2014) of decentralized decision making in developing countries suggest that methodology of the underlying studies is important: A clear focus on identification (such as the use of random control trials or various instrumental-variable applications), while currently limited, influences the results of program evaluations but cannot explain all of the different results. The review by Arcia et al. (2011) concludes that "the empirical evidence from Latin America shows very few cases in which SBM [school based management] has made a significant difference in learning outcomes (Patrinos (2011)), while in Europe there is substantial evidence showing a positive impact of school autonomy on learning (Eurydice (2007))." Cross-sectional evidence from international achievement tests concerning school autonomy has similarly been quite mixed (Hanushek and Woessmann (2011)), but these studies may also be particularly plagued by identification issues.

In the first use of the international PISA tests as a country panel, Hanushek et al. (2013) combine different waves of the international assessments by pooling the individual data of over one million students in 42 countries in the four PISA waves from 2000 to 2009. To avoid bias from unobserved cross-country differences such as those arising from culture and other government institutions, they incorporate country fixed effects in their estimation using the individual level data.[14] They exploit the fact that many countries have reformed their school systems to become more or less autonomous over time.

They find that school autonomy has a significant effect on student achievement but that this effect varies systematically with the level of economic and educational development: The effect of greater school autonomy on student achievement is strongly positive in developed and high-performing countries, but strongly negative in developing and low-performing countries. Countries with otherwise strong institutions gain considerably from decentralized decision-making in their schools, while countries that lack such a strong existing structure may actually be hurt by decentralizing decision-making.[15]

Hanushek et al. (2013) also 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 overall results across countries are shown in Fig. 2. Impacts of greater autonomy, measured on the vertical axis, rise with the level of economic development and are shifted higher with accountability.

Significantly, the amount of local decision making is frequently a

[12] Note that concerns about generalizing across states in federalist system may be particularly salient when it comes to testing, which often involves individual state policies.

[13] The underlying analysis for this section can be found in Hanushek et al. (2013).

[14] The analytical approach in this study motivated the previous analysis of the impacts of testing.

[15] Note that this interaction of institutional development and estimated impact of autonomy is entirely consistent with the empirical finding of differences between Latin America and Europe described above.

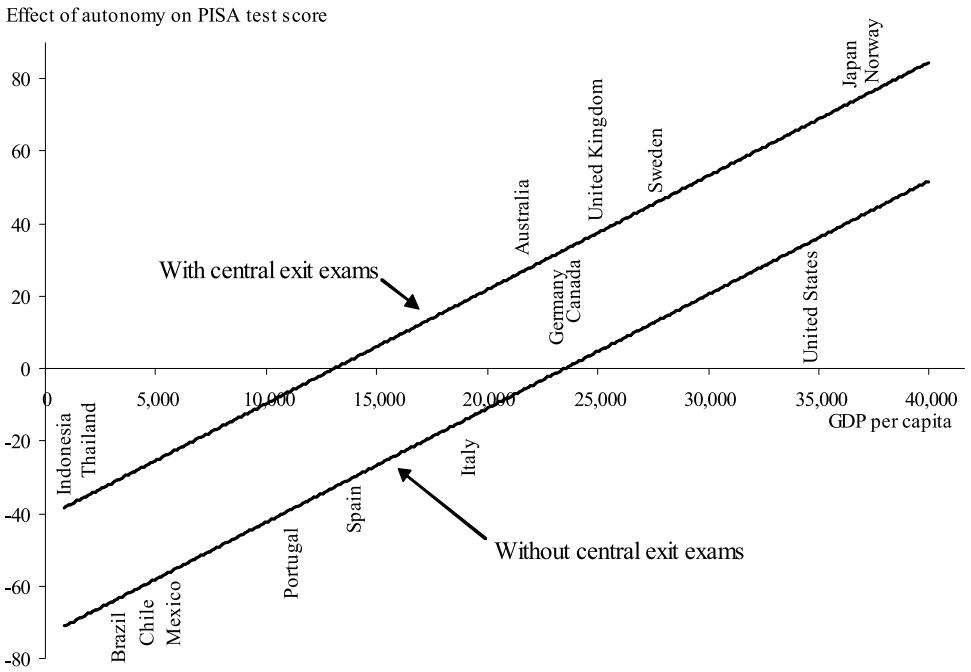

Fig. 2. Effect of autonomy reforms on student achievement by level of development.
Notes: Estimated effect of academic-content autonomy (scaled 0–1) on PISA math test score (scaled with standard deviation 100) depending on initial GDP per capita (in 2000) and on the existence of central exit exams, estimated in a panel model of PISA tests 2000-2009. Example countries illustrate initial level of GDP per capita. Own depiction based on Hanushek et al. (2013), Table 9.
Source: Hanushek and Woessmann (2015a).

feature that is common across the schools within a country. The heterogeneity of institutional effects again shows that understanding the institutional structures in a country's school system is important for assessing the generalizability of the more common within-country evaluations of specific policies and programs.

4.3. Case study 3: the market for skilled labor[16]

Teachers and administrators are often the ones most directly involved in carrying out an educational intervention, but many other "human inputs" enter into the process as well. Because most educational interventions require the work of people, it is not difficult to believe that the success of an intervention depends somewhat on the quality of the people involved in implementing it. While any such people differences might be adequately accounted for through randomization within a specific evaluation, differences in levels across countries in general are not an analytic factor that can be directly considered.

It is generally difficult to compare the quality of workers internationally, but the PIAAC data permit analyzing how labor markets vary around the sample of 32 countries described above. The easiest summary is to estimate a "modified Mincer Model" as described in the prior section, where skills are measured by the PIAAC literacy and numeracy assessments. From this, one can look at how the rewards to worker skill vary across countries and what factors could be underlying this variation.

Fig. 3 portrays the range of returns to math skills across countries.[17] The most obvious result is that these returns vary widely – from 11 percent higher wages for one standard deviation higher math scores in Greece to 45 percent in Singapore. The U.S. has returns of 25 percent for one standard deviation higher math scores (Hanushek et al., 2017a, 2017b).

The subtler facet of this figure is that the returns vary systematically with differences in the structure of the labor force and the characteristics of the economy. Though attaching a causal interpretation is not possible, there is a distinct relationship between returns to skills and significant country differences in union density, employment protection, and the portion of the population in the public sector (Hanushek et al. (2015)).

These results have been extended in several directions by Hampf et al. (2017) which considers not only measurement errors in the PIAAC tests but also possible biases from omitted variables and reverse causation. These consistently show significant differences in the returns to skill across countries.

Given that labor markets operate in such very different ways across countries, these different rewards for skills and the subsequent impact on workers can clearly lead to significant variations across countries in impacts of a given policy that is dependent on a substantial people element.

4.4. Case study 4: differences in teacher cognitive skills[18]

From the PIAAC sample, it is also possible to identify all individuals employed as teachers and then compare the test scores of teachers across countries. The cross-country differences in measured teacher skills are very large. Fig. 4 provides a comparison of the numeracy skills of the college educated population in each country. The bars represent the interquartile range of test scores for college graduates. The vertical line in each bar shows where the median teacher falls in the cognitive skill distribution of college graduates.

The figure makes clear that two things are important in determining the skills of teachers in any country. One is the quality of the pool of potential teachers. If a country has a better pool of college graduates (i.e., the bar for the interquartile range is farther to the right), it is likely to have teachers with greater numeracy skills. Second is where teachers are drawn out of that pool of college graduates. Finland has roughly the best pool, but it also draws the median teacher from the 62nd percentile of the distribution of college graduates. Farther down the figure is the U.S.:

[16] The underlying analysis for this section can be found in Hanushek et al. (2015, 2017b).

[17] The returns to skills are the coefficient estimates on numeracy score (standardized to a standard deviation of 1 within each country) in a regression of log gross hourly wage on numeracy, gender, and a quadratic polynomial in age for the sample of full-time employees aged 35–54. Hollow bars in the figure indicate first-round PIAAC countries, black bars indicate second-round PIAAC countries.

[18] The underlying analysis for this section can be found in Hanushek et al. (2019).

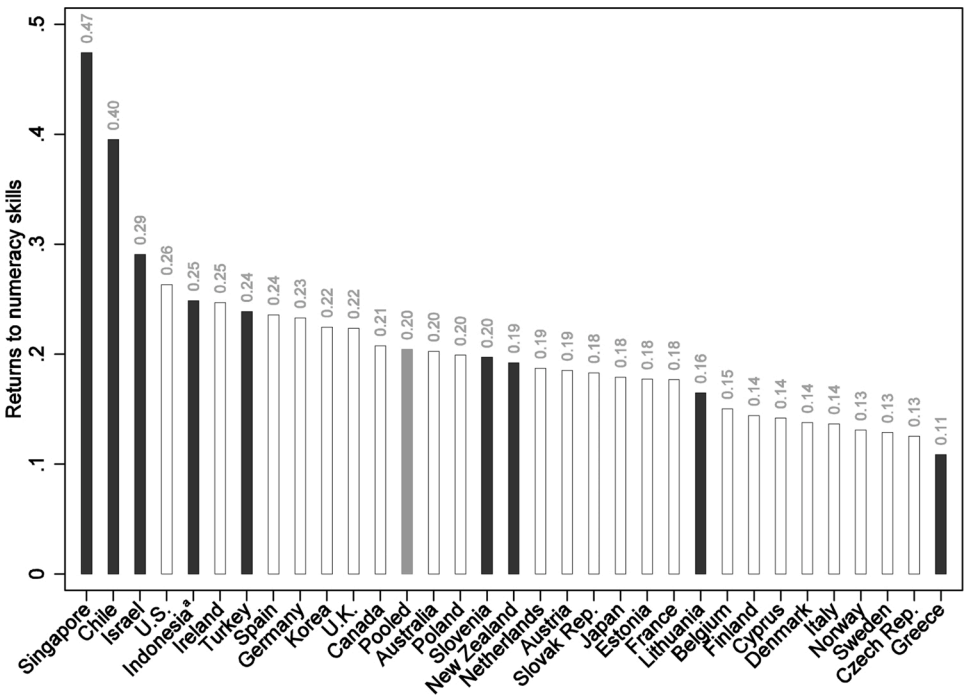

Fig. 3. Returns to Cognitive Skills.
Notes: Coefficient estimates on numeracy score (standardized to std. dev. 1 within each country) in a regression of log gross hourly wage on numeracy, gender, and a quadratic polynomial in age, sample of full-time employees aged 35–54. Regressions weighted by sampling weights. Pooled specification includes country fixed effects and gives same weight to each country. Hollow bars indicate first-round countries, black bars indicate second-round countries. [a]Jakarta only.
Source: Hanushek et al. (2017a) .

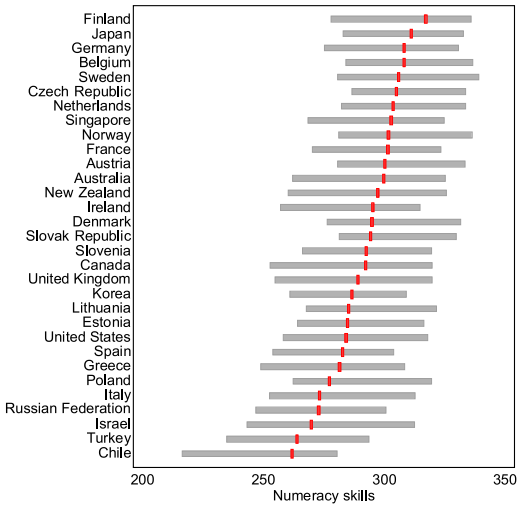

Fig. 4. Numeracy Scores of Teachers.
Note: Gray bars give the 25–75 percentile range of college graduates; Red marker indicates score of the median teacher.
Source: Hanushek et al. (2019).

it has a poorer pool of college graduates than Finland, and it also draws teachers from the 47th percentile of the college graduates.

If one relates the test scores of teachers to student achievement scores, it becomes apparent that smarter teachers yield smarter kids.[19] Teacher cognitive skills do not determine all of the differences in teacher effectiveness, but they are significant – explaining a substantial portion of the variation in average PISA scores across countries.

Importantly, the pattern of teacher selection and of teacher cognitive skills varies across time and across macro-institutional policy regimes. Two things can be identified as determining the skill differences of teachers across countries and over time. First, if women have more opportunities outside of teaching, the quality of teachers measured by test scores is lower. Historically in the U.S., women were concentrated in teaching and nursing, but this obviously changed over time and impacted the supply of teachers (in terms of cognitive skills). The relationship between the skills of teachers and the employment opportunities of women holds over time and across countries.

The second determining factor of teacher quality is the premium paid for being a teacher in a given country. Using PIAAC data it is possible to estimate a simple earnings function based on cognitive skills, experience, and gender (similar to that done in the previous section). Holding these attributes constant, the estimated earnings function indicates the earnings of an average teacher compared to a similar worker in other occupations. As shown in Fig. 5, teachers in the U.S. earn 22 percent less than similar workers in other professions.

The teacher wage premium essentially predicts where the median teacher will fall in the distribution of cognitive skills in Fig. 4.

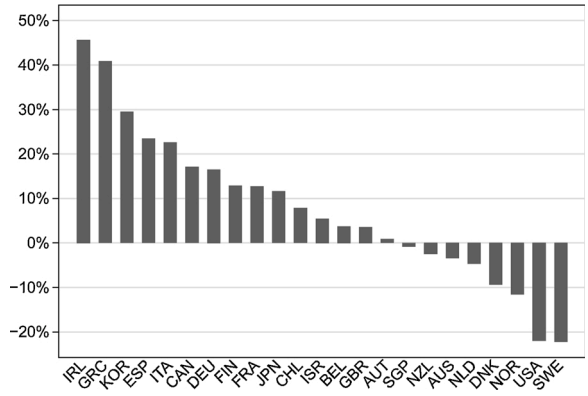

Fig. 5. Teacher Wage Premiums around the World.
Notes: Estimated higher earnings for teachers given their test scores, experience level, and gender.
Source: Hanushek et al. (2019).

[19] The analysis pursues a variety of approaches designed to support a causal interpretation of the country differences in teacher cognitive skills. These include while holding constant individual student fixed effects and analyzing a range of specifications and of placebo tests.

Furthermore, these wage premiums carry over into the achievement of students.

Thus, another way that evaluations of educational programs could differ across countries is varying teacher quality. This concern is actually a specialized issue about how labor markets differ across countries as identified in Case Study 3. The potential interaction of policy impacts and teacher quality enters into the potential transfer of policy results across countries but cannot be readily included in the individual studies themselves.

4.5. Case study 5: vocational versus general education[20]

A major organizational decision that is mostly made at the national level is the balance between general education and more vocationally oriented education. These discussions intensified after the 2008 recession when youth unemployment skyrocketed in many countries, leading to discussions about whether education more directed to the demands of industry would help smooth the school-to-work transition. The fact that Germany, a country with some of the most extensive vocational schooling, weathered the recession better than most European countries added fuel to a push toward more vocational schools.

Countries have actually made very different choices about the extent of vocational education. Germany, Switzerland, Austria, and Denmark, for example, have developed extensive apprenticeship programs where students split their time between formal schooling and work/training at firms. On the other hand, at least until recently the United States has essentially dismantled its vocational education system, largely on the argument that the vocational skills would become quickly obsolete with technological change. But, like many other countries, the U.S. began reconsidering vocational training when the Trump administration proposed re-igniting the vocational system as a way of giving employable skills to youth that had done poorly in the traditional schools.[21]

Analysis of the impacts of vocational education have mainly been aimed at understanding its impact on job entry of youth. Such analysis has been difficult, however, because students entering into vocational programs typically look different from those staying in general education. This fact makes development of an adequate control group difficult.

Hanushek et al. (2017b) address the comparison issue in the context of broadening the focus to consider life-cycle employment effects of vocational training. A central concern with vocational education is that those trained in very specific skills may not be able to adapt to changing demands for skills as production technologies changes. Data from the International Adult Literacy Survey (IALS) – an early precursor of the PIACC sample – provide detailed information about skills and labor market attributes of representative samples of adult workers in 11 countries with varying intensity of vocational training. From these data, it is possible to compare the employment patterns of workers with different types of education. To address the concern of selection into different types of education, they employ a difference-in-differences framework, comparing labor-market outcomes across different ages for people with general and vocational education. Under the assumption that conditional selectivity into education types does not vary over time, this approach allows them to identify how relative labor-market outcomes of different education types vary with age cohorts.

The pattern of employment for the most vocationally intensive countries (Germany, Switzerland, and Denmark in their data set) shows an initial employment advantage to vocational training but one that declines over the life-cycle and that tends to reverse at ages in the late 40′s (see Fig. 6). This result is confirmed with the PIAAC data by Hampf and Woessmann (2017).

The important aspect for this discussion is that the life-cycle employment patterns differ dramatically across countries and that these patterns follow the intensity of vocational training. Countries with less intensive vocational education see less differentiation in the life-cycle employment patterns. In contrast to apprenticeship countries, the U.S. with limited vocational education sees little life-cycle employment difference based on type of education. Thus, for example, analysis of employment outcomes within an individual country associated with an educational intervention may not generalize to countries that have inherently different structures to their vocational training. But the impact of these aggregate institutional differences on the labor market results cannot be ascertained within evaluations conducted in an individual country.

4.6. Case study 6: early tracking[22]

Countries also 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, but some countries such as Austria and Germany track students into different-ability schools as early as age 10. Many other countries maintain a comprehensive school system (although perhaps with some streaming within schools) through the end of high school. A common concern is that early tracking, perhaps because of the relative increase of parental influences or because of peer effects, may increase inequality as lower-achieving groups are tracked into lower-ability schools. But similar to the other case studies, this phenomenon cannot be readily analyzed within any given country when the structure of schools is set nationally.

Hanushek and Woessmann (2006) employ an identification strategy that compares achievement changes from primary to later schooling across tracked and untracked countries. Different country-level

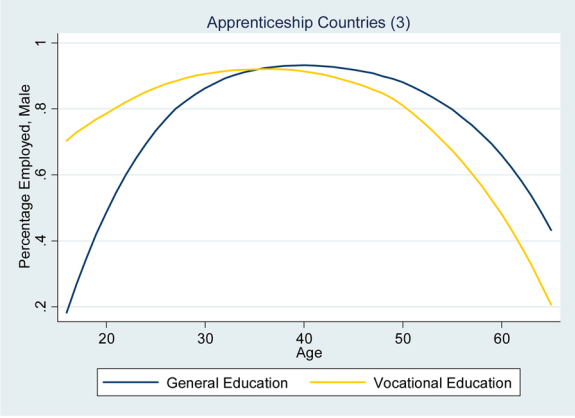

Fig. 6. Life-cycle Employment Rates by Education Type for Apprenticeship Countries.
Notes: Smoothed plots of employment rates by age for "apprenticeship" countries (Denmark, Germany, and Switzerland). Sample includes all males who finished secondary education or the first stage of tertiary education and are not currently enrolled in school.
Source: Hanushek et al. (2017b)

[20] The underlying analysis for this section can be found in Hanushek et al. (2017b).

[21] See, for example, "Remarks by President Trump in Meeting with Cabinet Members," July 18, 2018 (https://www.whitehouse.gov/briefings-statements/remarks-president-trump-meeting-cabinet-members/), accessed January 7, 2018.

[22] The underlying analysis for this section can be found in Hanushek and Woessmann (2006).

assessments provide information about inequality in scores at different grade levels.[23] Using a differences-in-differences model, they find that early tracking significantly increases the inequality in countries' achievement outcomes.

Fig. 7 shows how the standard deviation of scores changes between primary and secondary schools across countries that participated in the 2003 PISA and PIRLS tests. The largest increases in standard deviations are found in Germany, Greece, the Czech Republic, and Italy – all countries with early tracking. The largest decreases in standard deviations are found in Turkey, New Zealand, Canada, and the United States – all countries with no early tracking. Hanushek and Woessmann (2006) do not find a consistent effect of early tracking on the level of achievement. Interestingly, simple cross-sectional estimation with its attendant concerns about missing variable bias does not indicate the association of tracking with educational inequality found in the difference-in-differences analysis, showing the importance of careful attention to identification issues.

Building on these findings, Ruhose and Schwerdt (2016) find that early tracking negatively affects migrant students of the first generation as well as those second-generation migrant students who do not speak the host-country language at home.

Again, because the amount of tracking is largely a national educational decision, schools (and policies) within countries are conditioned by these structural factors. The differences in student inequality is conditioned by underlying (but unmeasured) institutional features.

5. Some implications

Returning to the simple thought experiment, if armed with a study with high internal validity, say from a well-structured RCT or a particularly compelling natural experiment, what conclusions will transfer to policy in a different country?

The previous case studies have a common theme. Looking across countries, there are significant national institutional factors that systematically affect student outcomes and that set the overall environment for schools but that are not readily incorporated into any program evaluation. Clearly each issue raised previously will vary in importance when consideration goes to specific educational evaluations, but the key variations discussed are relevant to a wide range of potential policy applications. This finding of significant structural variations sends up a series of warning signals about how small-scale evaluations can be generalized to other settings.

This discussion pulls together a number of strands of literature that directly relate to educational policy evaluations. The underlying theme is that a wide range of macro-institutional factors are likely to affect the micro-level evaluations, but these institutions cannot readily be considered in the micro-level evaluations.

One of the clearest interactions involves personnel issues versus programmatic issues. Evaluation analyses – whether RCTs or consideration of a natural experiment derived from policy changes – are generally most convincing when it is possible to describe the treatment as a binary condition where a program either exists or doesn't exist. But the implementation of programs involves personnel, often teachers. The randomization of the evaluation might when successful guard against selection issues of the program personnel (except when personnel issues are a central identified part of the program). The results will nonetheless be conditioned by the overall quality and character of the teacher labor market that determines the skills of the program and nonprogram teachers, the reaction of these people to varying incentives, and like. These are things that are both difficult to describe within a given country and hard to compare across countries.

Fig. 7. Early Tracking and Inequality.
Notes: Standard deviations of test scores for countries with early tracking (black with solid lines) and without early tracking (red with dotted lines). The fourth grade variation on the 2003 Progress in International Reading Literacy Study (PIRLS) is linked to the eighth grade variation on the 2003 PISA reading assessment.
Source: Hanushek and Woessmann (2006).

One inference might be that results can be roughly carried to different countries within similar development levels. For example, Ghana and India are both lower middle income countries (by World Bank classification), so programs developed in India might be reasonably applied in Ghana (e.g., see Duflo et al. (2020)). But the prior case studies suggest caution even there because of the significant labor market differences for teachers and nonteachers seen across OECD countries.

It might be tempting to think of the world as being linear and additive such that the macro-institutions do not affect the marginal policy choices being evaluated for some intervention at the micro-level. If so, it might be possible to transport the lessons about marginal policy effects to other environments.

But the kinds of macro-institutions considered here appear to go deeper. Because the institutions condition the impact of, say, salary policies that enter into personnel inputs in a wide range of policies, it is not obvious that they can be ignored. Policies whose effects interact with the level of economic development or the initial quality of the school system as a whole have marginal effects that are both more complicated than would be found in any single environment and that are very difficult to analyze or understand even within a given environment.

This suggests that learning about policy choices may be more expensive than previously thought, at least if only based on RCTs that need to be widely replicated. While careful, well-constructed experiments may yield very powerful findings within a given institutional structure, the findings may not travel well to other institutional structures. With the limited replication of RCTs and related evaluations across institutional settings, attempts to generalize results across countries appear very risky. Moreover, to the extent that the institutional

[23] Available tests for varying years include data for several pairs of achievement tests of the Progress in International Reading Literacy Study (PIRLS), the Trends in International Mathematics and Science Study (TIMSS), and the PISA study. PIRLS is an assessment of reading skills conducted by the IEA.

structure is constant across all of the subjects of the policy evaluation, it is not obvious how one guards against possible interactions with the marginal policy effects.

The existing enthusiasm for RCT and rigorous focus on the identification of causal structure are well-founded within the setting of specific evaluations. But the use of results for policy purposes in other international settings requires deepening the information about how they perform under differing institutional structures. This in turn raises a some serious research strategy issues because of the expense and time-commitments of RCTs (Pritchett and Sandefur, 2013; Ravallion, 2020).

Acknowledgments

Laura Talpey, Doug Besharov, an anonymous referee, and participants at the Conference on Rigorous Impact Evaluation in Europe provided many helpful comments and suggestions. This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

References

Amrein, Audrey L., Berliner, David C., 2003. Does accountability work? Education Next 3 (3 (Fall)), 8.
Arcia, Gustavo, Macdonald, Kevin, Patrinos, Harry A., Porta, Emilio, 2011. School autonomy and accountability. System Assessment and Benchmarking for Education Results. World Bank, Washington, DC.
Bando, Rosangela, Näslund-Hadley, Emma, Gertler, Paul, 2019. Effect of Inquiry and Problem Based Pedagogy on Learning: Evidence From 10 Field Experiments in Four Countries. National Bureau of Economic Research, Cambridge, MA. NBER Working Paper No. 26280 (September).
Banerjee, Abhijit V., Duflo, Esther, 2011. Poor Economics: a Radical Rethinking of the Way to Fight Global Poverty. Public Affairs, New York.
Bergbauer, Annika B., Hanushek, Eric A., Woessmann, Ludger, 2019. Testing. National Bureau of Economic Research, Cambridge, MA. NBER Working Paper No. 24836 (revised November 2019).
Berman, Amy I., Edward, H.Haertel, James, W.Pellegrino (Eds.), 2020. Comparability of Large-Scale Educational Assessments: Issues and Recommendations. National Academy of Education, Washington, DC.
Caplan, Bryan., 2018. The Case Against Education: Why the Education System Is a Waste of Time and Money. Princeton University Press, Princeton, NJ.
Coleman, James S., Campbell, Ernest Q., Hobson, Carol J., McPartland, James, Mood, Alexander M., Weinfeld, Frederic D., York, Robert L., 1966. Equality of Educational Opportunity. U.S. Government Printing Office, Washington, D.C.
Deaton, Angus., 2010. Instruments, randomization, and learning about development. J. Econ. Lit. 48 (2 (June)), 424–455.
Deaton, Angus, Cartwright, Nancy, 2018. Understanding and misunderstanding randomized controlled trials. Soc. Sci. Med. 210 (August), 2–21.
Duflo, Annie, Kiessel, Jessica, Lucas, Adrienne, 2020. External Validity: Four Models of Improving Student Achievement. National Bureau of Economic Research, MA. NBER Working Paper No. 27298 Cambridge (June).
Eurydice, 2007. School Autonomy in Europe: Policies and Measures. Eurydice, Brussels. December.
Finn Jr., Chester E., Hanushek, Eric A., 2020. Test-based accountability in distressed times. State Edu. Stand. 20 (3 (September)), 13–17.
Galiani, Sebastian, Perez-Truglia, Ricardo, 2014. School management in developing countries. In: Glewwe, Paul (Ed.), Education Policy in Developing Countries. University of Chicago Press, Chicago, pp. 193–241.
Hampf, Franziska, Woessmann, Ludger, 2017. Vocational vs. General education and employment over the life cycle: new evidence from PIAAC. CESifo Econ. Stud. 63 (3), 255–269.
Hampf, Franziska, Wiederhold, Simon, Woessmann, Ludger, 2017. Skills, earnings, and employment: exploring causality in the estimation of returns to skills. Large-scale Assess. Educ. 5 (12), 1–30.
Hanushek, Eric A., 1979. Conceptual and empirical issues in the estimation of educational production functions. J. Hum. Resour. 14 (3 (Summer)), 351–388.
Hanushek, Eric A., 2002. Publicly provided education. In: Auerbach, Alan J., Feldstein, Martin (Eds.), Handbook of Public Economics, Volume 4. North Holland, Amsterdam, pp. 2045–2141.
Hanushek, Eric A., Kimko, Dennis D., 2000. Schooling, labor force quality, and the growth of nations. Am. Econ. Rev. 90 (5 (December)), 1184–1208.
Hanushek, Eric A., Woessmann, Ludger, 2006. Does educational tracking affect performance and inequality? Differences-in-differences evidence across countries. Econ. J. 116 (510 (March)), C63–C76.
Hanushek, Eric A., Woessmann, Ludger, 2008. The role of cognitive skills in economic development. J. Econ. Lit. 46 (3), 607–668.
Hanushek, Eric A., Woessmann, Ludger, 2011. The economics of international differences in educational achievement. In: Hanushek, Eric A., Machin, Stephen, Woessmann, Ludger (Eds.), Handbook of the Economics of Education, Volume 3. North Holland, Amsterdam, pp. 89–200.
Hanushek, Eric A., Woessmann, Ludger, 2012. Do better schools lead to more growth? Cognitive skills, economic outcomes, and causation. J. Econ. Growth 17 (4), 267–321.
Hanushek, Eric A., Woessmann, Ludger, 2015a. The Knowledge Capital of Nations: Education and the Economics of Growth. MIT Press, Cambridge, MA.
Hanushek, Eric A., Woessmann, Ludger, 2015b. Universal Basic Skills: What Countries Stand to Gain. Organisation for Economic Co-operation and Development, Paris.
Hanushek, Eric A., Link, Susanne, Woessmann, Ludger, 2013. Does school autonomy make sense everywhere? Panel estimates from PISA. J. Dev. Econ. 104, 212–232.
Hanushek, Eric A., Schwerdt, Guido, Wiederhold, Simon, Woessmann, Ludger, 2015. Returns to skills around the world: evidence from PIAAC. Eur. Econ. Rev. 73, 103–130.
Hanushek, Eric A., Schwerdt, Guido, Wiederhold, Simon, Woessmann, Ludger, 2017a. Coping with change: international differences in the returns to skills. Econ. Lett. 153, 15–19.
Hanushek, Eric A., Schwerdt, Guido, Woessmann, Ludger, Zhang, Lei, 2017b. General education, vocational education, and labor-market outcomes over the life-cycle. J. Hum. Resour. 52 (1), 48–87.
Hanushek, Eric A., Piopiunik, Marc, Wiederhold, Simon, 2019. The value of smarter teachers: international evidence on teacher cognitive skills and student performance. J. Hum. Resour. 54 (4 (Fall)), 857–899.
Heckman, James J., Smith, Jeffrey A., 1995. Assessing the case for social experiments. J. Econ. Perspect. 9 (2 (Spring)), 85–110.
Heyneman, Stephen P., Bommi, Lee, 2015. International large-scale assessments: uses and implications. In: Ladd, Helen F., Goertz, Margaret E. (Eds.), Handbook of Research in Education Finance and Policy. Routledge, New York and London, pp. 104–118.
Heyneman, Stephen P., Loxley, William, 1983. The effect of primary school quality on academic achievement across twenty-nine high and low income countries. Am. J. Sociol. 88 (6 (May)), 1162–1194.
Hout, Michael, Elliott, Stuart W. (Eds.), 2011. Incentives and Test-Based Accountability in Education. National Academies Press, Washington, DC.
Komatsu, Hikaru, Rappleye, Jeremy, 2017. A new global policy regime founded on invalid statistics? Hanushek, woessmann, pisa, and economic growth. Comp. Educ. 53 (2), 166–191.
Koretz, Daniel., 2017. The Testing Charade: Pretending to Make Schools Better. University of Chicago Press, Chicago.
Levine, Ross, Renelt, David, 1992. A sensitivity analysis of cross-country growth regressions. Am. Econ. Rev. 82 (4 (September)), 942–963.
Levine, Ross, Zervos, Sara J., 1993. What we have learned about policy and growth from cross-country regressions. Am. Econ. Rev. 83 (2 (May)), 426–430.
Lucas, Adrienne M., McEwan, Patrick J., Ngware, Moses, Oketch, Moses, 2014. Improving early-grade literacy in East Africa: experimental Evidence from Kenya and Uganda. J. Policy Anal. Manag. 33 (4 (Fall)), 950–976.
Mincer, Jacob, 1970. The distribution of labor incomes: a survey with special reference to the human capital approach. J. Econ. Lit. 8 (1 (March)), 1–26.
Mincer, Jacob, 1974. Schooling, Experience, and Earnings. NBER, New York.
Mullis, Ina V.S., Martin, Michael O., Foy, Pierre, Hooper, Martin, 2016. TIMSS 2015 International Results in Mathematics. Boston College, Boston.
Nelson, Richard R., Phelps, Edmund, 1966. Investment in humans, technology diffusion and economic growth. Am. Econ. Rev. 56 (2 (May)), 69–75.
OECD, 2016. PISA 2015 Results: Policies and Practices for Successful Schools. Organisation for Economic Co-operation and Development, Paris.
Patrinos, Harry A., 2011. School-based management. In: Bruns, Barbara, Filmer, Deon, Patrinos, Harry A. (Eds.), Making Schools Work: New Evidence on Accountability Reforms. The World Bank, Washington, D.C, pp. 87–140.
Pritchett, Lant, 2006. Does learning to add up add up? The returns to schooling in aggregate data. In: Hanushek, Eric A., Welch, Finis (Eds.), Handbook of the Economics of Education. North Holland, Amsterdam, pp. 635–695.
Pritchett, Lant, Sandefur, Justin, 2013. Context matters for size: why external validity claims and development practice don't mix. J. Glob. Dev. 4 (2 (December)), 161–197.
Pritchett, Lant, Sandefur, Justin, 2015. Learning from experiments when context matters. Am. Econ. Rev. 105 (5), 471–475.
Psacharopoulos, George, Patrinos, Harry Anthony, 2018. Returns to investment in education: a decennial review of the global literature. Educ. Econ. 26 (5 (September)), 445–458.
Ravallion, Martin, 2020. "Should the Randomistas (Continue to) Rule?" NBER Working Paper No. 27554. National Bureau of Economic Research, Cambridge, MA. July.
Ruhose, Jens, Schwerdt, Guido, 2016. Does early educational tracking increase migrant-native achievement gaps? Differences-in-differences evidence across countries. Econ. Educ. Rev. 52, 134–154.
Singer, Judith D., Braun, Henry I., Chudowsky, Naomi (Eds.), 2018. International Education Assessments: Cautions, Conundrums, and Common Sense. National Academy of Education, Washiington, DC.
Spence, A.Michael., 1973. Job market signalling. Q. J. Econ. 87 (3 (August)), 355–374.
Welch, Finis., 1970. Education in production. J. Polit. Econ. 78 (1 (January/February)), 35–59.
Woessmann, Ludger, Luedemann, Elke, Schuetz, Gabriela, West, Martin R., 2009. School Accountability, Autonomy, and Choice Around the World. Edward Elgar, Cheltenham, UK.
World Bank, 2018. World Development Report 2018: Learning to Realize Education's Promise. World Bank, Washington, DC.