Matthias von Davier • Eugenio Gonzalez
Irwin Kirsch • Kentaro Yamamoto
Editors

# The Role of International Large-Scale Assessments: Perspectives from Technology, Economy, and Educational Research

Springer

## Table of Contents

# Chapter 4
# The Role of International Assessments of Cognitive Skills in the Analysis of Growth and Development

**Eric A. Hanushek and Ludger Woessmann**

## Introduction

Economists have found the concept of human capital to be very useful in explaining not only differences in individual earnings but also aggregate variations in the well-being of nations. Because of the importance of human capital, another strand of research has delved into the determinants of relevant skills that fit into human capital. Both lines of inquiry have advanced markedly with development and expansion of international testing of achievement, particularly in math and science.

Economists are now accustomed to looking at issues of skill development from the vantage point of human capital theory. The simplest notion is that individuals make investments in skills that have later payoffs in outcomes that matter. And, in this, it is commonly presumed that formal schooling is one of several important contributors to the skills of an individual and to human capital. It is not the only factor. Parents, individual abilities, and friends undoubtedly contribute. Schools nevertheless have a special place because they are most directly affected by public policies.

The human capital and investment perspective immediately makes it evident that the real issues are ones of long-term outcomes. Future incomes of individuals are related to their past investments. It is not their income while in school or their income in their first job. Instead, it is their income over the course of their working life.

Much of the early and continuing development of empirical work on human capital concentrates on the role of school attainment, that is, the quantity of schooling. The revolution in the United States during the twentieth century was universal

E. A. Hanushek (✉)
Hoover Institution, National Bureau of Economic Research and CESifo, Stanford University, Stanford, CA 94305-6010, USA
e-mail: hanushek@stanford.edu

L. Woessmann
Ifo Institute for Economic Research and CESifo, University of Munich, Poschingerstr. 5, 81679 Munich, Germany
e-mail: woessmann@ifo.de

schooling. This policy goal has spread around the world, encompassing both developed and developing countries. It also has lent itself to regular measurement. Quantity of schooling is easily measured, and data on years attained, both over time and across individuals, are readily available. But quantity of schooling proves to be a poor measure of the skills of individuals both within and across countries.

The growth of standardized measures of achievement has proven extraordinarily valuable in filling out a richer picture of human capital. The research base has expanded significantly through work in the United States and elsewhere that exploits rich school accountability data. The administrative data sets accompanying accountability systems have proven very valuable in understanding the determinants of student achievement.

The research based on the international assessments is perhaps equally important. Importantly, it goes in two different directions. Research designed to understand the underlying determinants of cognitive skills parallels that of the administrative data sets while permitting a range of analyses not possible with the accountability data. Additionally, however, the research based in international data sets has focused on the consequences of skill differences.

By going beyond the use of simple measures of the quantity of schooling, economists have been able to understand better the role of human capital in outcomes and the elements that are important in producing more human capital. International achievement data, developed and refined over the past half century, were not collected to support any specific economic research agenda. But there are a number of research and policy agendas that are uniquely amenable to analysis because of the existence of such data.

This discussion, following the development in Hanushek and Woessmann (2011a), concentrates on the role of achievement as a direct measure of human capital. The international data have distinct advantages over research restricted to single countries or states. The data permit exploitation of variation that only exists across countries. For example, systematic institutional variation between countries—as found with differences in the competitiveness and flexibility of teacher labor markets, forms of accountability systems, the extent of a private school sector, or the structure of student tracking—simply does not exist within most countries. And, even where within-country variation exists, variations across countries in key institutional factors and in characteristics of the schools and population are frequently much larger than those found within any country.

The international achievement data, based on a consistent collection process, provides an opportunity to examine comparable estimates of the determinants and consequences of educational achievement for a diverse set of countries. Such research can thus illuminate whether a result is truly country-specific, applies more generally, or is simply a spurious result from a particular within-country sample. Further, international evidence can identify systematic heterogeneity in effects that differ across countries.

Even where within-country variation exists, for example, in the case of public and private schools operating within the same system, comparisons of student achievement are often subject to severe selection problems. Students who choose to attend a private school may differ along both observable and unobservable dimensions



from students taught in neighborhood public schools. While it is possible to control for some differences in student, family, and school characteristics when estimating the effects of institutional structures, such estimates may still suffer from selection on unobserved characteristics (see Chap. 7). At the country level, it is possible to circumvent these selection problems—in effect measuring the impact of, for example, the share of students in a country attending private schools on student achievement in the country as a whole. Such cross-country evidence will not be biased by standard issues of selection at the individual level. (At the same time, as discussed below, international comparisons present their own analytical challenges).

Importantly, uncovering general equilibrium effects is often impossible in a single country but sometimes feasible across countries. For example, the presence of private schools may influence the behavior of nearby public schools with which they compete for students. As a result, simple comparisons of private and public schools may miss an important part of the effects of greater private involvement in education, while aggregation to the country level can potentially solve the problem. By comparing the average performance of systems with larger and smaller shares of private schools, the cross-country approach captures any systemic effect of competition from private schools.

Research into the consequences of differences in cognitive skills has similar advantages. For example, while the implications of human capital development for macroeconomic outcomes—including, importantly, economic growth—can potentially be investigated with time-series data for individual countries, historical data are effectively limited to school attainment with no information on the cognitive skills that we emphasize here. On the other hand, variations in cognitive skills across different economies can, as we describe below, effectively get at such fundamental questions. Similarly, investigating whether features of the structure of economic activity affect the individual returns to skills is very difficult within a single economy with interlocking labor and product markets.

While international achievement data at times substitute for the collection of national data, the discussion here focuses on the use of international tests for cross-country analyses. These studies have different basic designs. One focuses on within-country variations in achievement or the outcomes of achievement but then considers how these within-country relationships differ across countries. The second emphasizes the cross-country relationships per se.

## International Testing[1]

International consortia were formed in the mid-1960s to develop and implement comparisons of educational achievement across nations. The first major international test was conducted in 1964 when 12 countries participated in the First International Mathematics Study (FIMS). This and a series of subsequent assessments

[1] A more detailed description of historical international testing is found in Hanushek and Woessmann (2011b). This section provides an overview of relevant testing

involved a cooperative venture developed by the International Association for the Evaluation of Educational Achievement (IEA). Since then, the math, science, and reading performance of students in many countries have been tested on multiple occasions using (at each occasion) a common set of test questions in all participating countries. By 2010, three major international large-scale assessment (ILSA) programs were surveying student performance on a regular basis: the Programme for International Student Assessment (PISA), testing math, science, and reading performance of 15-year-olds on a three-year cycle since 2000; the Trends in International Mathematics and Science Study (TIMSS), testing math and science performance (mostly) of fourth and eighth-graders on a four-year cycle since 1995; and the Progress in International Reading Literacy Study (PIRLS), testing primary-school reading performance on a five-year cycle since 2001. In addition, regional testing programs have produced comparable performance information for many countries in Latin America and sub-Saharan Africa, and international adult literacy surveys have produced internationally comparable data on the educational achievement of adults.

These international testing programs have some common elements. They involve a group of voluntarily participating countries that each pay for their participation and administer the same assessment, translated into their own (official) language(s). The set of participating countries has differed across time and even across tested domains of specific testing occasions. Additionally, the different tests differ somewhat in their focus and intended subject matter. For example, the IEA tests, of which the most recent version is TIMSS, are developed by international panels but are related to common elements of primary and secondary school curriculum, while the Organisation for Economic Co-operation and Development's (OECD) PISA tests are designed to measure more applied knowledge and skills.[2] Until recently testing has been almost exclusively cross-sectional in nature, not following individual students' change in achievement.[3]

Along with the assessments of cognitive skills, extensive contextual information and student background data have been provided by related surveys. The motivation for this is using the international databases to address a variety of policy issues relevant to the participating countries.

The IEA and OECD assessments have the broadest coverage and have also adapted regular testing cycles. Table 4.1 provides an account of their major international tests with an indication of age (or grade level) of testing, subject matter, and participating countries. By 2007, there were 15 testing occasions, most of which include subparts based upon subject and grade level.[4]

[2] A separate analysis of coverage and testing can be found in Neidorf et al. (2006).

[3] The Second International Mathematics Study (SIMS) of the IEA did have a one-year follow-up of individual students that permitted some longitudinal, panel information, but this design was not repeated. Recent innovations have permitted development of panel data by individual countries. This comparison over time has been aided by the linking of tests over time—including recent administrations of TIMSS, PIRLS, and PISA.

[4] See Mullis et al. (2007, 2008), and Organisation for Economic Co-operation and Development (2007) for details on the most recent cycle of the three major ongoing international testing cycles. PISA also has conducted a 2009 assessment and both PISA and TIMSS have announced future assessments.



Table 4.1 Tests of educational achievement: IEA and OECD student achievement tests. (Source: Hanushek and Woessmann 2011a)

| Abbreviation | Study | Year | Subject | Age[a,b] | Countries[c] | Organization[d] | Scale[e] |
|---|---|---|---|---|---|---|---|
| FIMS | First international mathematics study | 1964 | Math | 13, FS | 11 | IEA | PC |
| FISS | First international science study | 1970–1971 | Science | 10, 14, FS | 14, 16, 16 | IEA | PC |
| FIRS | First international reading study | 1970–1972 | Reading | 13 | 12 | IEA | PC |
| SIMS | Second international mathematics study | 1980–1982 | Math | 13, FS | 17, 12 | IEA | PC |
| SISS | Second international science study | 1983–1984 | Science | 10, 13, FS | 15, 17, 13 | IEA | PC |
| SIRS | Second international reading study | 1990–1991 | Reading | 9, 13 | 26, 30 | IEA | IRT |
| TIMSS | Third international mathematics and science study | 1994–1995 | Math/Science | 9(3+4), 13(7+8), FS | 25, 39, 21 | IEA | IRT |
| TIMSS-Repeat | TIMSS-repeat | 1999 | Math/Science | 13(8) | 38 | IEA | IRT |
| PISA 2000 2002 | Programme for international student assessment | 2000–2002 | Math/Sci./Read. | 15 | 31+10 | OECD | IRT |
| PIRLS | Progress in international reading literacy study | 2001 | Reading | 9(4) | 34 | IEA | IRT |
| TIMSS 2003 | Trends in Int'l mathematics and science study | 2003 | Math/Science | 9(4), 13(8) | 24, 45 | IEA | IRT |
| PISA 2003 | Programme for international student assessment | 2003 | Math/Sci./Read. | 15 | 40 | OECD | IRT |
| PIRLS 2006 | Progress in international reading literacy study | 2006 | Reading | >9.5(4) | 39 | IEA | IRT |
| PISA 2006 | Programme for international student assessment | 2006 | Math/Sci./Read. | 15 | 57 | OECD | IRT |
| TIMSS 2007 | Trends in international mathematics and science study | 2007 | Math/Science | >9.5(4), >13.5(8) | 35, 48 | IEA | IRT |

*OECD* Organisation for Economic Co-operation and development, *IRT* item-response-theory proficiency scale

[a] Grade in parentheses where grade level was target population

[b] FS final year of secondary education (differs across countries)

[c] Number of participating countries that yielded internationally comparable performance data

[d] Conducting organization: international association for the evaluation of educational achievement (IEA)

[e] Test scale: percent-correct formal (PC)

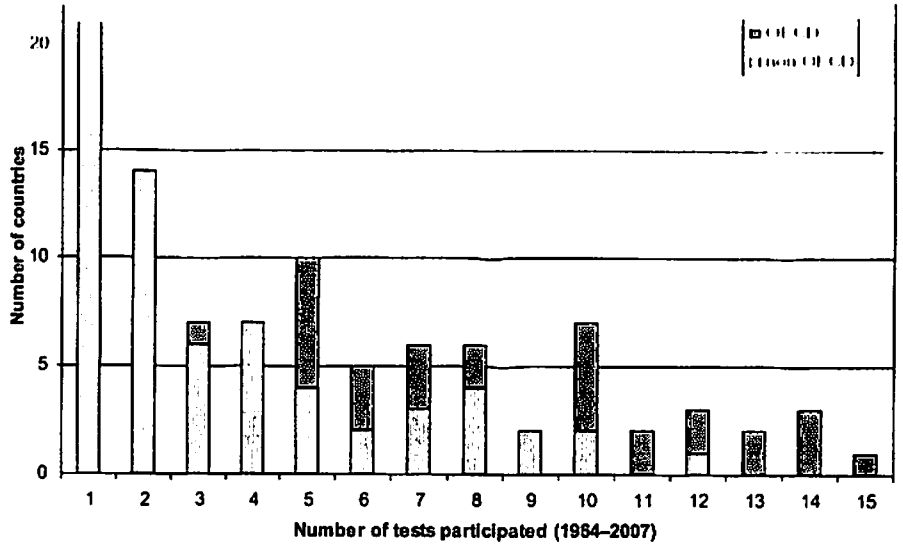


**Fig. 4.1** Participation in international student achievement tests of IEA and OECD through 2007. (Source: Hanushek and Woessmann 2011a)

The major IEA and OECD testing programs have expanded dramatically in terms of participating countries. While only 29 countries participated in these large-scale assessments through 1990, a total of 96 countries have participated by 2007. Three additional countries participated in 2009, and another three planned to participate in 2011, raising the total number of countries ever participating in one of these international tests to 102. Only the United States participated in all 15 testing occasions, but an additional 17 countries participated in 10 or more different assessments. Figure 4.1[5], from Hanushek and Woessmann (2011a), shows the histogram of participation on the IEA or OECD tests between 1964–2007, divided by OECD and other countries. From this figure, it is clear that the depth of coverage is much greater for developed than for developing countries. Further, much of the participation in one or two different test administrations occurs after 2000.

At the same time, a number of more idiosyncratic tests, some on a regional basis, have also been developed. These tests have been more varied in their focus, development, and quality, and they have in general been used much less frequently in analytical work. Of the ten additional testing occasions, six are regional tests for Latin America (ECIEL, LLECE, SERCE) or Africa (SACMEQ I and II, PASEC); *see* Hanushek and Woessmann (2011a). One difficulty with these regional tests has been the lack of linkage to the other international tests, implying that any cross-country analyses must rely exclusively on the within-region variance in institutions, populations, and achievement.

The remaining international assessments and surveys cover a broader set of countries but are somewhat different in focus. The International Assessment of Educational Progress (IAEP) I and II are tests constructed to mirror the National Assessment of Educational Progress (NAEP) that has been used in the United States since 1970 and that aligns to the US school curriculum. The International Adult Literacy Survey (IALS) and the Adult Literacy and Life Skills (ALL) survey have a very different structure involving sampling of adults in the workforce.[6] The IALS survey data in particular have been used in a variety of studies about the consequences of education and cognitive skills.

Interestingly, the TIMSS tests, with their curricular focus, and the PISA tests, with their real-world application focus, are highly correlated at the country level. For example, the correlation coefficients at the country level between the TIMSS 2003 tests of eighth graders and the PISA 2003 tests of 15-year-olds across the 19 countries participating in both are 0.87 in math and 0.97 in science, and they are 0.86 in both math and science across the 21 countries participating both in the TIMSS 1999 tests and the PISA 2000/02 tests. There is also a high correlation at the country level between the curriculum-based student tests of TIMSS and the practical literacy adult examinations of IALS (Hanushek and Zhang 2009). Tests with very different foci and perspectives tend to be highly related at the country level, suggesting that they are measuring a common dimension of skills (see also Brown et al. 2007).

## The Explosion of Studies

Economists largely ignored the existence or potential of these international assessments until fairly recently. They made little use of the possibility of comparative studies across countries. But the last decade has seen a tremendous upsurge in research activity on cross-country issues.

As noted, economists have pursued two separate lines of inquiry, each related to notions of human capital. The first subsection considers studies that take the cognitive skills measures from the international tests as a direct measure of human capital and focuses on the determinants of varying levels of human capital. This work, commonly referred to as analyses of education production functions, investigates how various inputs to education affect outcomes. The traditional investigations of how families and school resources influence achievement have been supplemented by a range of studies into economic institutions—accountability, choice, etc.

[5] Number of tests in which a country has participated in the following 15 IEA and OECD tests: FIMS, FISS, FIRS, SIMS, SISS, SIRS, TIMSS, TIMSS-Repeat, PISA 2000/02, PIRLS, TIMSS 2003, PISA 2003, PIRLS 2006, PISA 2006, TIMSS 2007. Total number of participating countries: 96.

[6] The OECD has currently also embarked on a new endeavor, the Programme for the International Assessment of Adult Competencies (PIAAC), which will update and expand the adult testing, in terms of both the scope of the test and the number of participating countries. This assessment began being administered in 2011.

**Table 4.2** Economic studies of the determinants of human capital using international achievement tests (Source: Hanushek and Woessmann 2011a)

| Data source | Determinants of student achievement | | | | Achievement equity | Total | Unique studies |
|---|---|---|---|---|---|---|---|
| | Family background plus school inputs | | Institutions | | | | |
| | Within country | Cross-country | Within country | Cross-country | | | |
| IEA | 15 | 2 | 1 | 2 | 1 | 21 | 20 |
| OECD | 6 | 4 | 3 | 7 | 2 | 22 | 20 |
| Other | | 2 | 2 | 1 | | 5 | 4 |
| Combined | 3 | 3 | | 4 | 6 | 16 | 16 |
| *Total* | *24* | *11* | *6* | *14* | *9* | *64* | *60* |

The second major line of inquiry has turned to cross-country investigations of the outcomes of human capital and is discussed in the second subsection. The traditional labor market studies of the determination of earnings across individuals have been placed in an international context, permitting some investigation of how different economies reward human capital. Additionally, studies of outcomes have looked at the distribution of earnings within countries and at differences in economic growth across countries.[7]

## *Studies of the Determinants of Achievement*

Table 4.2 summarizes the economic studies found in the review in Hanushek and Woessmann (2011a).[8] A total of 60 unique studies have considered the determinants of cognitive skills across countries. Interestingly, only four of these studies were published before 2000.[9] The recentness of the analysis partially reflects recent expansion in the scope of international testing, but it also derives from more recent appreciation of the kinds of analyses that are possible with the international data.

For the determinants of achievement, a prime distinction from an analytical viewpoint is whether the study uses the between-country variation in performance in the basic estimation. Studies that are labeled "within country" estimate a series of models based on samples stratified by country. The results are then compared across countries. The studies labeled "cross country" use the variations in outcomes among countries in the basic estimation. The within-country analyses always rely

[7] Studies of outcome differences related to cognitive skills are reviewed and evaluated in Hanushek and Woessmann (2008).

[8] The primary requirement for inclusion in the review is that the studies are comparative in nature, relying on the comparisons across countries. Some studies relying on the international data sets along with a large number of studies employing single country data sources have maintained a focus simply on the determinants of achievement within an individual country and are not included here.

[9] Heyneman and Loxley (1983), Bishop (1989a), Bishop (1997), and Toma (1996).



on the microdata sets from the various international studies, while the cross-country studies include a mixture of those relying on microdata and those using country aggregate data from the international data sets.

The studies of determinants are further subdivided into those primarily considering the role of families and school resources and those that highlight institutional factors. Quite naturally, studies of families and resources tend to rely most on within-country variation, while institutional studies rely more on cross-country variation. Institutions that set the general rules for school operations structure much of what goes on in the schools of every country—but they cause analytical difficulties because they often apply to all schools in a country. Thus, it is difficult to observe any variations of what occurs with different institutions, and it is difficult to understand fully the impact on achievement of both the institutions and other features of the educational system. With educational system level variables such as reliance on accountability systems or reliance on private schools, there is generally limited variation within countries, and the variation that exists is often contaminated by selection factors that make the identification of effects difficult. Therefore, it is necessary to look across countries where the institutional variation exists.

A total of 51 studies investigate differences across countries in the production of achievement.[10] Another nine studies look at the variation in achievement—or equality of achievement—across countries and what factors influence that.[11]

The second element of the table is tracing the data that lies behind each of the studies. The studies to date have been dominated by the various IEA and OECD data collections. Here the importance of the IEA and OECD is clear, with relatively few using other sources. Moreover, the majority of the combined studies employ the various IEA and OECD assessments only.

The international investigations of the determinants of educational achievement have followed a voluminous literature based on data for individual countries.[12] Indeed the data available within individual countries is often superior to that from the international surveys. Specifically, more recent studies tend to rely heavily on panel data sets that follow the achievement of individual students and that can link this achievement growth to characteristics of families, schools, and teachers. With these extensive data sets, identification of separate causal determinants of achievement is frequently much clearer than in the simple cross-sections of data supplied by the international assessments.

What makes the international data valuable in these studies is the chance to observe influences that cannot be readily analyzed within a single country. The most straightforward example is the application of test-based accountability. Since these frequently apply to entire countries, there is no variation within countries that can

[10] Three studies appear in more than one column of studies of determinants because they focus equally on institutional factors and on families and schools.

[11] One of these studies also appeared in the tabulation for the four preceding columns, making a total of 60 unique studies of various aspects of the determinants of achievement.

[12] See, for example, the review in Hanushek (2004) and the international perspective in Woessmann (2004).

be used (except perhaps the information about before and after introduction of a system).[13] But systems vary across countries, allowing variation that can be exploited to understand the impacts of accountability. Similarly the impacts of broad based preschool programs or general choice of schools is subject to selection problems if just program take-up is considered, and any general equilibrium effects (improvements to all schools) are difficult to detect within individual systems.

The clearest and most unique evidence provided by this international work is that the overall set of educational institutions has a significant impact on student achievement. In particular, countries with test-based accountability systems, with more school choice, and with more local decision making or more local autonomy tend to do better (Hanushek and Woessmann (2011a)). Moreover, the work has provided some important, policy-relevant details. For example, having local decision making over teacher salaries only appears to make sense if the country has other supportive institutions such as an accountability system that will focus attention on the appropriate set of outcomes.[14]

The analytical tradeoff, of course, with the international surveys is that it is often difficult to be sure that cultural factors and other systematic differences across countries are satisfactorily dealt with in the analysis. In simplest terms it is generally difficult to be sure that international results are not driven by unmeasured culture, institutions, and the like. Therefore, these international assessments are not a substitute for national data systems but instead are a complement that permits alternative kinds of studies. Moreover, as mentioned, a number of studies cannot be done within the confines of a single country.

## *The Studies of Outcomes*

The outcome studies are quite different. They look at the economic implications of varying achievement. Table 4.3 summarizes the existing studies reviewed in Hanushek and Woessmann (2011a), all but one of which has been published from 2000 onwards.

The studies that have been conducted have each addressed issues that cannot be studied with data on an individual country. They specifically rely on the cross-country variation in measured skills.

Because the benefits of investment in human capital necessarily come over time, the standard international data collection at a given school age does not provide direct information on the value of cognitive skills for individuals in the labor market. Indeed this has been a general problem in looking at wage determination even within countries, because general census data and other surveys do not follow individuals over time. As a result, studies of individual earnings never use the IEA or PISA data but instead have relied on the IALS because that survey collects information on individuals of varying ages along with their earnings.[15]

One of the most interesting results from the international studies of that different economies appear to value cognitive skills to quite different degrees. Hanushek and Zhang (2009) trace the returns to higher cognitive skills across 11 countries participating in IALS.[16] Figure 4.2, which plots of proportional increase in earnings associated with a one standard deviation increase in achievement, shows that the US economy appears to reward skills more than any of the other countries observed. Some countries, like Poland and Sweden, however, provide little labor market

[13] The United States with varying state accountability systems prior to No Child Left Behind has similarities to the international differences—where there is no institutional variation within states but there is variation between states. See Carnoy and Loeb (2002) and Hanushek and Raymond (2005).

[14] Hanushek et al. (2011) Combine all of the PISA data into a country-level panel. With this, they investigate how school autonomy in various areas affects achievement. They find that developed countries, particularly those with high performing school systems and with text-based accountability tend to perform better with local decision making. However, less developed countries appear to do worse when there is more autonomy in decision making.

Table 4.3 Studies of the economic consequences of human capital using international achievement tests (Source: Hanushek and Woessmann 2011a)

| Data source | Economic consequences | | | Total |
|---|---|---|---|---|
| | Individual earnings | Equity | Aggregate outcomes | |
| IEA | | 1 | 1 | 2 |
| OECD | | | | |
| Other | 6 | 3 | 1 | 10 |
| Combined | | | 13 | 13 |
| *Total* | *6* | *4* | *15* | *25* |

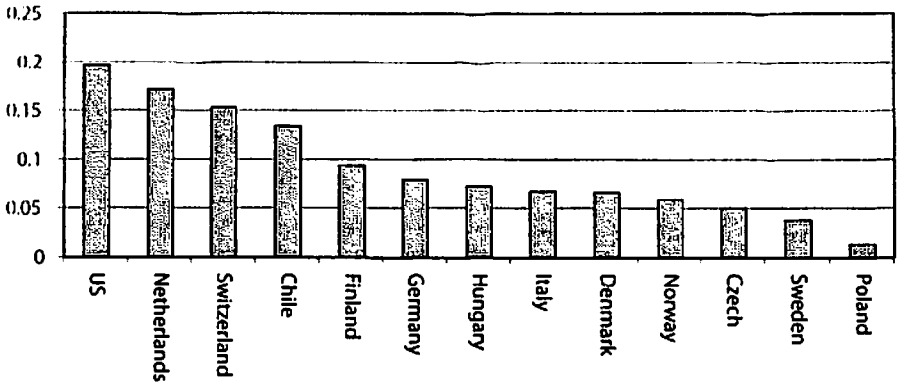


Fig. 4.2 Returns to cognitive skills, international adult literacy survey. (Source: Hanushek and Zhang 2009)

[15] The only exception to use of IALS data is Bedard and Ferrall (2003), which combines observations of Gini coefficients with early IEA data.

[16] The analysis follows what is commonly referred to as a Mincer earnings function in which differences in individual earnings are related to school attainment and labor market experience. These estimates simply add the IALS measure of cognitive skills to such a relationship.

reward to higher skills. (The explanation of the causes of these differences awaits further research).

The most unique use of the international tests—and in many ways the most important—has related to aggregate economic performance of nations. Economists have spent considerable effort over the past two decades trying to understand why some countries grow faster than others. This is an extraordinarily important question because it is economic growth that determines the long run well-being of societies.

Much of the initial work by economists recognized that the economic performance of a nation had to relate to the human capital of the nation, but it was hampered by measurement issues. In particular, the only readily available information on skills was school attainment. But use of school attainment for nations requires an assumption that learning in a year of schooling is the same across countries—an almost ludicrous assumption.

The international achievement measures provide a much more defensible way to measure skill differences. This approach was first pursued in Hanushek and Kimko (2000)) and has subsequently been reproduced and extended elsewhere (see the review in Hanushek and Woessmann (2008)). The underlying idea is to combine test from the various existing international assessments. These ILSA programs have included a varying group of participating countries, and the tests are (until recently) not linked to each other. But, we develop a comparable scale for them by noting that the US has participated in all of the assessments *and* the US has a linked national assessment in the National Assessment of Educational Progress (NAEP). By using the scores on NAEP to adjust the US scores on comparable international exams (by age and subject), it is possible to create a time-consistent series of performance of the US We then develop an estimate of the appropriate variance for each international test by using the variance within a set of comparison countries from those with well-developed schooling systems at the time of the earlier tests. This variance estimate allows us to put all countries who ever participate in an international assessment onto a common scale. For most purposes, then, we take the simple average of all observed scores for a country as a measure of the achievement that is relevant for the labor force. (For details on the construction of the comparable test data over time, see Hanushek and Woessmann (2009)).

The power of these measures is easy to see. Figure 4.3[17] shows the relationship between achievement and average annual economic growth in GDP per capita from 1960–2000 for 50 countries with the necessary data.[18] The strength of the relationship between skills and growth is apparent from this figure. Behind this figure is a

[17] Added-variable plot of a regression of the average annual rate of growth (in percent) of real GDP per capita in 1960–2000 on the initial level of real GDP per capita in 1960, average years of schooling in 1960, and average test scores on international student achievement tests.

[18] This plot is an added-variable plot where the other estimated underlying regression model also includes initial level of gross domestic product (GDP) per capita. In simplest terms, it is easier for a low-income country to grow faster because it only needs to imitate the technologies in more advanced countries while advanced countries must develop innovations in order to grow



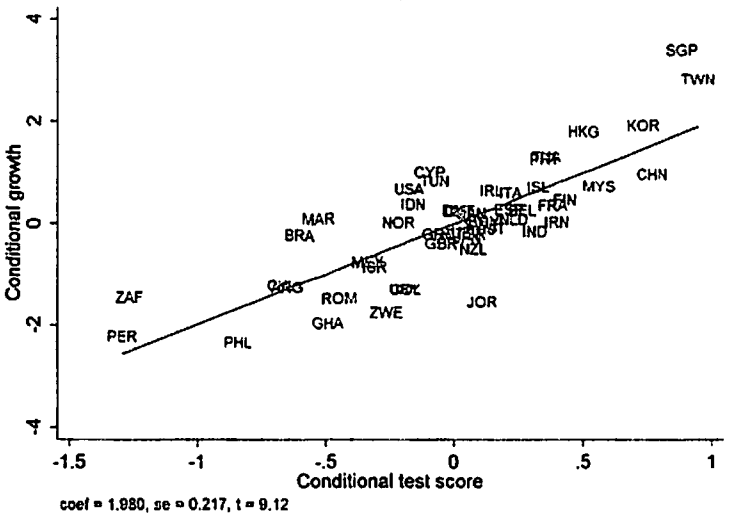


Fig. 4.3 Cognitive skills and economic growth. (Source: Hanushek and Woessmann 2008)

simple statistical relationship that relates annual growth rates to GDP per capita in 1960 and our calculation of achievement for each country.

It is also possible in a parallel manner to show the traditional story based on school attainment. Figure 4.4[19] describes the simple relationship of school attainment and growth (taking into account initial income levels). As the top panel shows, attainment is correlated with growth—but much less closely than we saw for cognitive skills. But, once cognitive skills are included, there is no relationship between school attainment and growth (bottom panel). In other words, only school attainment that translates into learning and achievement has an impact.

There are of course many caveats and qualifications to this. Perhaps the most important is worry about whether the relationship can be assumed to represent a causal relationship and not merely an association in this particular sample. Hanushek and Woessmann (2009) provide a variety of tests that support a causal interpretation, although it remains difficult in a small cross-sectional of countries to obtain conclusive evidence.[20]

If we use the underlying estimates of the growth relationship, we can vividly see the importance of achievement. Hanushek and Woessmann (2011b) simulate the impact of the US economy (and other OECD economies) for a series of scenarios

[19] Added variable plot of a regression of the average annual rate of growth (in percent) of real GDP per capita in 1960–2000 on the initial level of real GDP per capita in 1960 and average years of schooling in 1960. The bottom panel additionally controls for average test scores on international student achievement tests, whereas the top panel does not.

[20] That study also discusses in detail the construction of the underlying data series along with a variety of interpretive issues.

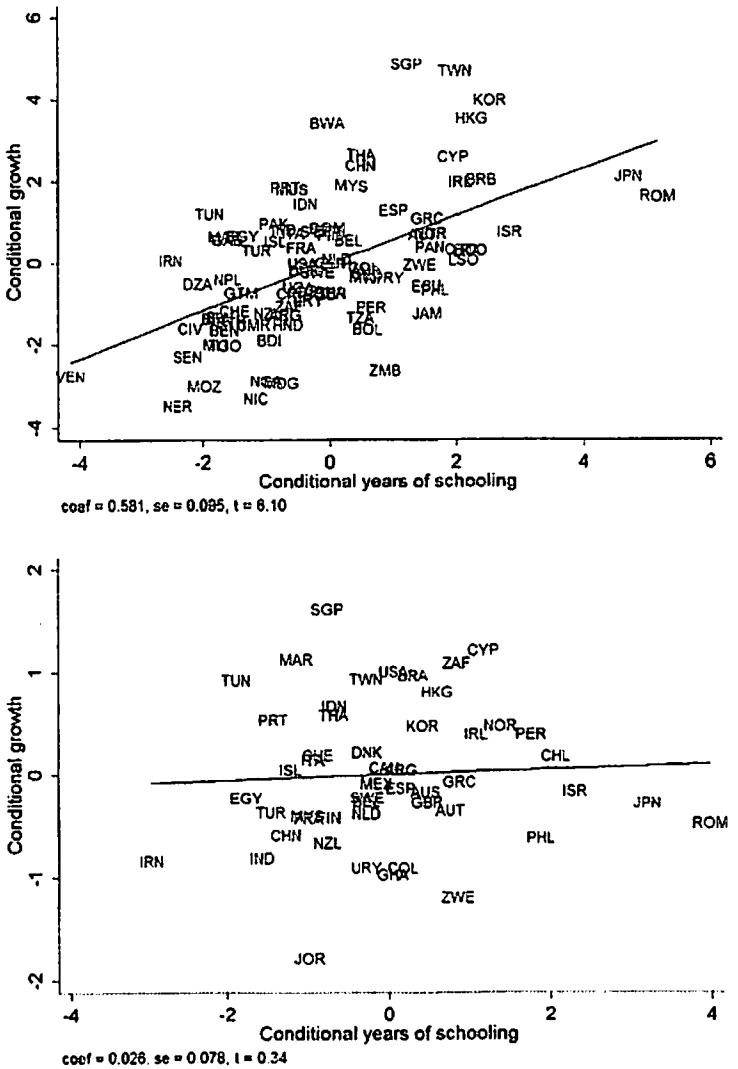


Fig. 4.4 Years of schooling and economic growth without and with test-score controls. (Source: Based on Hanushek and Woessmann 2008)

representing different school improvement programs. In each, it is assumed that the United States takes 20 years to reach new achievement levels. The three scenarios are as follows: (1) a gain of 25 points (1/4 S.D.) on the PISA tests; (2) a movement up to the level of Finland, the world leader on PISA; and, (3) movement of all students scoring below 400 (one standard deviation below the OECD mean, or generally Level 1). The simulations presume that the cognitive skills–growth relationship



Table 4.4 Estimated long run impact of improvement in achievement. (Source: Hanushek and Woessmann 2011b)

| | Scenario I: Increase avg. performance by 1/4 S.D. | Scenario II: Bring each country to Finnish level of 546 points on PISA | Scenario III: Bring all to minimum of 400 points on PISA |
|---|---|---|---|
| | (1) | (2) | (3) |
| OECD Aggregate Improvement in trillion US$ | 123.1 | 275.4 | 226.3 |
| United States Improvement in trillion US$ | 43.8 | 111.9 | 86.1 |

Discounted value of future increases in OECD GDP until 2090, expressed in trillion US$ (PPP)

observed across the past half-century hold into the future, and this permits estimating how much higher gross domestic product (GDP) would be with added achievement compared to the current levels.

The implications for the economy of these differences are truly astounding. Economic growth is projected over an 80-year period (the expected life of somebody born today), and then the present value of the gains is calculated.[21] Table 4.4 summarizes estimates of the three scenarios for all of the OECD countries and for the United States by itself. A 25-point improvement (something obtained by a number of other countries in the world) would have a present value of US$ 44 trillion for the United States (and US$ 123 trillion for the entire OECD). Reaching the performance levels of Finland would add US$ 112 trillion in present value to the US economy. Just bringing everybody up to basic skills (400 points on PISA)—something akin to achieving No Child Left Behind—would, however, yield a striking US$ 86 trillion.

From a policy point of view, these calculations underscore the need for aggressive (and successful) policies aimed at improving achievement and skills. From a research point of view, the ability to uncover such fundamental relationships highlights the enormous value of the underlying large scale international surveys.

## Some Things to be Addressed

The existing literature has produced a number of interesting and useful results. But it also has faced a number of continuing problems and challenges. Here we simply list some of the biggest issues.

[21] The present value weights economic gains closer to today more heavily than those in the future. It is easiest to interpret as the amount of money that, invested at an assumed return of 3 % per year, could produce the projected GDP pattern over time

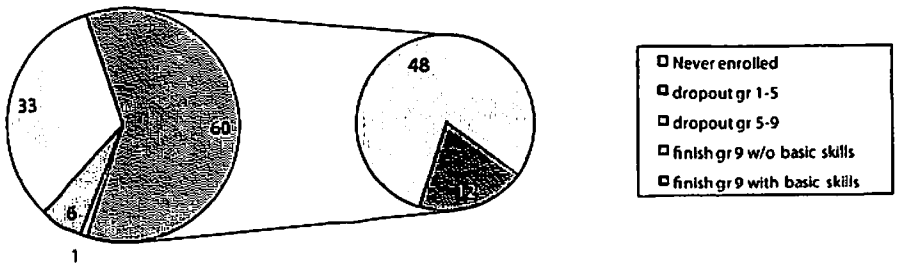


Fig. 4.5. School performance in Peru. (Source: Based on Hanushek and Woessmann 2008)

## *Some Measurement Issues*

The international assessments meet a variety of purposes for the individual countries and for development organizations. One important purpose is to provide individual countries with a benchmark both of what is possible and of where the country stands.

These issues are important for all countries, but they are especially important for developing countries. And here the story is not very pretty. Look for example at schooling in Peru (Fig. 4.5). Peru has a high level of school attainment—but few of its students appear to be learning much when in school. Only one-fifth of the students are achieving at the basic 400-point level on PISA. From a measurement viewpoint, one has to wonder if the PISA test is even giving meaningful information about the skills of students in Peru and other countries similarly situated in terms of performance. An obvious direction in the testing evolution is developing tests that provide meaningful information within and across developing countries while also providing linking information to show relative standings in the world. This could be accomplished, for example, by continuing regional tests that were aimed at specific populations while including meaningful linking items to the PISA and TIMSS tests.

A second issue is the ability to link assessments to earlier experiences or to ones that were originally conducted in parallel, such as TIMSS and PISA. The ideal approach involves including linking items on all tests and, for parallel tests, going to large-scale studies administering both assessments to equivalent samples of students. Statistical adjustments such as the one described by Hanushek and Woessmann (2009) may be used, but rely on strong assumptions. All of the repeated international assessments have recently made progress on linkages of assessment cycles over time. Further work, including linkages between PISA and TIMSS, would have substantial pay-offs. These issues are relevant both for studies of educational production functions and for studies of the economic outcomes.



## *Issues of Causation*

A prime difficulty in the existing analyses is being confident about the identification of causal effects. Almost all of these studies are concerned with policy issues—either improving achievement or using achievement to obtain improved economic outcomes. It is obviously difficult to produce randomized experiments in a number of these areas. Pushing forward on causal issues is frequently quite difficult.

One of the key issues, particularly when looking at the determinants of individual achievement, is to follow the growth trajectories of students over time. The importance of collecting panel data on student performance is that it facilitates isolating the impact of specific interventions on achievement. Of course, as discussed above and elsewhere, other approaches such as exploiting natural experiments for exploring causal influences should also be pursued. The use of panel data simply provides a broadly applicable way to going deeper into the policy questions that are important. This conclusion, for example, comes out of the extensive work on administrative data bases within individual countries. With the exception of the IEA in the Second International Mathematics Study (SIMS), this has not been pursued in the main assessments. Interestingly, however, several countries have developed their own national follow-on studies, beginning with the sampled students for the PISA assessment. This kind of activity should clearly be encouraged.

## *Understanding Individual Economic Outcomes*

As noted, there has already been preliminary work done on adult assessments and surveys that permit investigation of labor market outcomes. These surveys have been very important for research into the determinants of earnings. Expansion of these would permit research into the deeper question of what aspects of an economy drive the demands for human capital and skills. While there have been a few attempts to get at these issues, the work to date is quite rudimentary.[22]

Looking in the opposite direction, validating the importance of measured tests for economic outcomes could provide valuable information about the tests themselves. A variety of people have criticized current testing systems because of potential problems such as teaching to the test or outright cheating.[23] If on the other hand the scores on these achievement assessments prove to be closely related to economic outcomes that we care about, we would have less concern about focusing on such test performance.

[22] See, for example, the innovative attempt to understand supply and demand for skills in Leuven et al. (2004).

[23] See, for example, Hout and Elliott (2011). Although, the evidence behind these critiques has been extraordinarily limited and weak, indicating that other approaches to validating the tests are necessary (Hanushek 2011).

## Conclusions

The development of international testing and assessments has been quite extraordinary. From humble beginnings, when the question was more, "Can it be done?", assessments have become embedded in the international world.

Much of the development of these assessments has been driven by a general notion that having comparisons across countries is a good idea—without much explicit consideration of how these assessments might be used in a larger research and policy context.

The burgeoning literature that considers both what factors contribute to score differences and what impacts scores have on economic outcomes shows the larger value of these assessments. It is perhaps time to consider how these large-scale international assessments could be made even more useful through direct linkage to the larger research activities.

## References


Bedard, Kelly, and Christopher Ferrall. 2003. Wage and test score dispersion: some international evidence. *Economics of Education Review* 22 (1):31–43.

Bishop, John H. 1995. The impact of curriculum-based external examinations on school priorities and student learning. *International Journal Of Educational Research* 23 (8):653–752.

Bishop, John H. 1997. The effect of national standards and curriculum-based examinations on achievement. *American Economic Review* 87 (2):260–264.

Brown, Giorgina, John, Micklewright, Sylke V., Schnepf and Waldmann, Robert. 2007. International surveys of educational achievement: How robust are the findings? *Journal of the Royal Statistical society A* 170 (3):623–646.

Carnoy, Martin, and Loeb, Susanna. 2002. Does external accountability affect student outcomes? A cross-state analysis. *Educational Evaluation and Policy Analysis* 24 (Winter 4):305–331.

Hanushek, Eric A. 2003. The failure of input-based schooling policies. *Economic Journal* 113 (February 485):64-98.

Hanushek, Eric A. 2012. Grinding the anti-testing ax: The national research council's accountability report. *Education Next* 12 (Spring 1):68–73.

Hanushek, Eric A., and Kimko, Dennis D. 2000. Schooling, labor force quality, and the growth of nations. *American Economic Review* 90 (December 5):1184–1208.

Hanushek, Eric A., Link, Susanne and Ludger, Woessmann. 2011. Does school autonomy make sense everywhere? panel estimates from PISA. *NBER Working Paper*17591 (November). Cambridge: National Bureau of Economic Research.

Hanushek, Eric A., and Raymond, Margaret E. 2005. Does school accountability lead to improved student performance? *Journal of Policy Analysis and Management* 24. (Spring 2):297–327.

Hanushek, Eric A., and Woessmann, Ludger. 2008. The role of cognitive skills in economic development. *Journal of Economic Literature* 46. (September 3):607–668.

Hanushek, Eric A., and Woessmann, Ludger. 2009. Do better schools lead to more growth? cognitive skills, economic outcomes, and causation. NBER Working Paper 14633. (January). Cambridge: National Bureau of Economic Research.

Hanushek, Eric A., and Woessmann, Ludger. 2011a. The economics of international differences in educational achievement. In *Handbook of the economics of education, Vol. 3*, eds. Eric A. Hanushek, Machin, Stephen, and Woessmann, Ludger, 89–200. Amsterdam: North Holland.





Hanushek, Eric A., and Woessmann, Ludger. 2011b. How much do educational outcomes matter in OECD countries? *Economic policy* 26 (67):427–491.

Hanushek, Eric A., and Zhang, Lei. 2009. Quality-consistent estimates of international schooling and skill gradients. *Journal of Human Capital* 3 (Summer 2):107–143.

Heyneman, Stephen P., and Loxley, William. 1983. The effect of primary school quality on academic achievement across twenty-nine high and low income countries. *American Journal Of Sociology* 88 (May 6):1162–1194.

Hout, Michael, and Elliott, Stuart W. eds. 2011. *Incentives and test-based accountability in education*. Washington, DC: National academies press.

Leuven, Edwin, Hessel, Oosterbeek and van Ophem, Hans. 2004. Explaining international differences in male skill wage differentials by differences in demand and supply of skills. *Economic Journal* 114 (April 495):466–486.

Mullis, Ina V.S., Michael. O. Martin, and Foy, Pierce. 2008. *TIMSS 2007 International mathematics report: Findings from IEA's trends in international mathematics and science study at the fourth and eighth grades*. Chestnut Hill: TIMSS & PIRLS International Study Center, Lynch School of Education, Boston College.

Mullis, Ina V. S., Michael. O, Martin, Ann M., Kennedy and Foy, Pierce. 2007. *PIRLS 2006 international report: IEA's progress in international reading literacy study in primary schools in 40 countries*. Chestnut Hill: TIMSS & PIRLS International Study Center, Lynch School of Education, Boston College.

Neidorf, Teresa S., Marilyn, Binkley, Kim, Gattis and Nohara, David. 2006. *Comparing mathematics content in the national assessment of educational progress (NAEP), Trends in international mathematics and science study (TIMSS), and program for international student assessment (PISA) 2003 assessments* (May). Washington: National Center for Education Statistics.

Organisation for Economic Co-operation and Development. 2007. *PISA 2006: Science competencies for tomorrow's world* Vol. 1 Analysis Paris: OECD.

Toma, Eugenia F. 1996. Public funding and private schooling across countries. *Journal of Law and Economics* 39 (1):121–148.

Woessmann, Ludger. 2003. Schooling resources, educational institutions, and student performance: the international evidence. *Oxford Bulletin of Economics and Statistics* 65 (2):117–170.