National Academy of Education

Workshop Series on Methods and Policy Uses of International Large-Scale Assessment (ILSA)

The Analysis of International Large-Scale Assessments to Address Causal Questions in Education Policy

Anna K. Chmielewski and Elizabeth Dhuey
University of Toronto

December 2016

Contact:
Anna K. Chmielewski, Ontario Institute for Studies in Education, University of Toronto, 252 Bloor Street West, Toronto, ON M5S 1V6, +1 416 978 1174, ak.chmielewski@utoronto.ca

Elizabeth Dhuey, Department of Management, University of Toronto, 1265 Military Trail, Toronto, ON. Canada, M1C 1A4, +1 416-208-2687, dhuey@utsc.utoronto.ca

This paper was prepared for the National Academy of Education's *Workshop Series on Methods and Policy Uses of International Large-Scale Assessment (ILSA)*. The research reported here was supported by the Institute of Education Sciences, U.S. Department of Education, through Grant R305U150003 to the National Academy of Education. The opinions expressed are those of the authors and do not represent views of the Institute or the U.S. Department of Education.

I. Introduction

International large-scale assessments (ILSAs) are examinations of academic achievement or literacy carried out in a large number of countries using the same set of adapted test instruments and methodology.[1] Since their inception more than 50 years ago, ILSAs have been used not only to compare performance, curricula, instructional, and learning strategies across countries but also to try to understand how international variations in education policies—the structure, administration, legal, economic, and political contexts of national and subnational school systems—shape differences in student achievement and other outcomes. In a sense, all policy research is comparative, whether those comparisons are experimental between treatment and control groups or observational between jurisdictions or time periods. Comparing education policies *internationally* has a number of advantages, the most fundamental being the large variation that exists across different countries' education systems. This means that researchers can (1) analyze a range of different kinds of systems and (2) validate whether findings from one country hold in different contexts. The promise of international comparisons was already clear to the founders of the International Association for the Evaluation of Educational Achievement (IEA) in the 1950s, who "viewed the world as a natural educational laboratory" (IEA, 2011).

ILSA data are uniquely suited to making cross-national comparisons of education policies because comparable student achievement scores and background data are collected across all participating countries. This is a significant advantage compared to the alternative of compiling and harmonizing national data sets.[2] However, ILSAs have been relatively underutilized by U.S. education policy researchers. This may be due to some of the challenges and limitations of ILSA data. First, the data available are limited to what can be collected comparably across countries. The second and third limitations both have to do with the difficulty of attempting to make causal inferences with ILSA data. Education policy research is typically interested in asking causal questions about the impacts of policies on student outcomes and experiences. But attempting to draw causal inferences from national characteristics or educational policies to national student results is difficult. Because national characteristics and policies are not randomly assigned to countries but instead are the product of a variety of historical and cultural factors, it is not possible to know whether a difference in a particular national policy is responsible for a difference in ILSA results between two countries. Thus, the second limitation of ILSA data is that researchers must find ways to measure or control for the large cross-national differences in culture, history, and context. The third limitation is that none of the ILSAs includes longitudinal data for students. Many of the quasi-experimental research designs used in education policy research to attempt to make causal inferences with observational data rely on longitudinal data, which is not possible in this context. Despite these limitations, recent improvements in methodology and the accumulating number of different ILSAs in different years have allowed a small but growing number of scholars to design studies that allow them to come closer to making credible causal inferences using ILSA data.

[1] Originally the foci of ILSAs were academic achievement and literacy. In recent years, there has been a shift to assess other domains such as collaborative problem solving, financial literacy, and noncognitive skills.

[2] Harmonizing refers to the attempt of standardizing surveys and assessments with the aim of achieving or improving the comparability of the different surveys or assessments.

When attempting to draw causal inference in education policy research, randomized controlled trials (RCTs) are often considered the "gold standard." However, even if an RCT is executed carefully and accurately, the estimates provided by that single trial often lack much external validity. In addition, even well-randomized trials that use appropriate statistical methodology can often have serious problems with internal validity (Ginsburg & Smith, 2016). In the context of ILSA data, there is no random assignment to policy treatments, so researchers generally need to make do with quasi-experimental methods. This makes causal inference much more difficult and often fraught with assumptions. A detailed discussion of the assumptions needed for particular research designs is beyond the scope of this review. The works of Murnane and Willett (2010) and Shadish, Cook, and Campbell (2002) are both good references that provide an interested reader a detailed review of quasi-experimental research designs which can be used for causal inference and their associated assumptions. Instead, the focus of this review is on the relatively new body of literature using quasi-experimental designs with ILSA data. We aim to showcase the variety of designs that are possible using ILSA data, as well as their attendant limitations and challenges.

Several previous authors have reviewed the uses of ILSA data for causal inference in education research. For example, Robinson (2013) introduces quasi-experimental research designs that support causal inference and illustrates with examples from one study using ILSA data (Bedard & Dhuey, 2006) and several studies using U.S. data. Kaplan (2016) argues for a Bayesian approach to causal inference using large-scale assessments, which allows researchers to incorporate not only uncertainty in parameters but also uncertainty in model choice. Rutkowski and Delandshere (2016) use a validity framework to critique causal claims made by two well-cited educational studies, one using ILSA data (Schmidt et al., 2001) and one using U.S. data only (Mosteller, 1995). Hanushek and Wößmann (2011) review a large body of literature that has used ILSA data to examine questions in the field of economics of education, such as the effects of accountability policies, competition from private schools, and curricular tracking. Many of the included studies use designs to support causal inference, although the main focus of Hanushek and Wößmann's review is on the evidence for substantive economics questions.

In contrast with these previous reviews, the goal of our review is a more comprehensive synthesis of literature across disciplines (economics, sociology, psychology, and education policy) that have used ILSA data to address causal questions in education policy. In Part II, we describe how authors have exploited various aspects of ILSAs to attempt to identify random variation in education policies. In Part III, we then present a variety of methodological challenges confronted by authors using ILSA data, some examples of strategies used in the literature to address these issues, and our own recommendations.

Many ILSAs are referred to by acronyms. Table 1 presents the full names and acronyms, as well as other characteristics of ILSAs analyzed by the studies reviewed in this paper.

Table 1. Names, Acronyms, and Characteristics of Selected International Large-Scale Assessments

Full Name	**Acronym**	**Year(s)**	**Organization**	**Age(s)/Grade(s)**	**Subject(s)**
First International Mathematics Study	FIMS	1964	IEA	Age 13, all students in the grade level where the majority of students were age 13, and pre-university students	Math
First International Science Study	FISS	1970/1971	IEA	Ages 10 and 14, and students in final grade of secondary education	Science
Second International Mathematics Study	SIMS	1980	IEA	Eighth grade and final grade of secondary education	Math
Second International Science Study	SISS	1984	IEA	Fourth, eighth, and final grade of secondary education	Science
Third International Mathematics and Science Study	TIMSS	1995	IEA	Third, fourth, seventh, eighth, and final grade of secondary education	Math, science
Third International Mathematics and Science Study-Repeat	TIMSS-R	1999	IEA	Eighth grade	Math, science
Trends in International Mathematics and Science Study	TIMSS	Every 4 years starting in 2003	IEA	Fourth and eighth grades	Math, science
Progress in International Reading	PIRLS	Every 5 years	IEA	Fourth grade	Reading

Literacy Study		starting in 2001			
Programme for International Student Assessment	PISA	Every 3 years starting in 2000	OECD	Age 15	Reading, math, science
International Adult Literacy Study	IALS	1994	Statistics Canada	Ages 16-65	Literacy, quantitative
Adult Literacy and Life Skills Survey	ALL	2003	Statistics Canada	Ages 16-65	Literacy, numeracy
Programme for the International Assessment of Adult Competencies	PIAAC	Every 10 years starting in 2011	OECD	Ages 16-65	Literacy, numeracy, problem solving

II. Overview of Research Designs Used with ILSA Data

In organizing our review of studies that have used ILSA data to address causal questions in education policy, we classified studies by how the authors exploited ILSA data to find variation in the educational policies, social conditions, or "treatments" that students experience. This resulted in five categories. The first two categories use cross-sectional ILSA data from a single point in time; Category 1 is research that examines policy variation among countries, and Category 2 examines policy variation within countries. The last three categories examine policy variation over time, within countries. Category 3 uses repeated cross-sectional ILSA data to look at variation across birth cohorts or generations of students. Category 4 also uses repeated cross-sectional ILSA data, but in this case to look at variation across age within the same birth cohort within countries. Category 5 encompasses the studies using the rare truly longitudinal ILSA data that follow the same students over time. Each of these study design categories is explained in more detail in the following sections and illustrated with studies from the category.

A. Category 1: Cross-national variation

A1. Correlational cross-national variation

The first category is research that examines cross-national correlations between country contextual or policy characteristics and national performance. As described in the Introduction, it is not possible to know whether these national policy differences *cause* the differences in student performance because policies are not randomly assigned to countries and are confounded with a variety of cultural and historical factors. However, the studies below still represent an improvement over research that makes causal claims about the importance of national policy context using only data from a single country, or a comparison between two or three countries,

because they include a large number of countries. Consistent levels of correlation between national policies and student outcomes across a large number of countries slightly strengthen the case for causal relationships between those policies and ILSA results, as do statistical controls for other national characteristics that might confound the relationship.[3] Still, it should be kept in mind that all of these analyses are still correlational and may be far from true causal evidence.

For example, three studies have examined whether the relative contribution of family background versus school effects to student achievement depends on a country's level of development. This question has been studied using data from FISS, conducted in 1970 (Heyneman & Loxley, 1983), using TIMSS 1995 (Baker, Goesling, & LeTendre, 2002), and using TIMSS 2003 (Chudgar & Luschei, 2009). Note that, although the associations among student achievement, family background, and school effects are calculated using student-level data, the "treatment" of interest is a national-level variable, economic development (measured by gross domestic product per capita). Other studies have examined how ILSA results are related to other country characteristics, such as income inequality (Chmielewski & Reardon, 2016; Chudgar & Luschei, 2009; Marks, 2005), duration of time that a democratic government has been in place (Torney-Purta, Wilkenfeld, & Barber, 2008), gender egalitarianism in society (Wiseman, Baker, Riegle-Crumb, & Ramirez, 2009), curricular differentiation and tracking policies (Buchmann & Park, 2009; Chmielewski, 2014; Chmielewski, Dumont, & Trautwein, 2013; Chmielewski & Reardon, 2016; Marks, 2005; Pfeffer, 2008; Schmidt, Burroughs, Zoido, & Houang, 2015), and the level of socioeconomic segregation between schools (Willms, 2010), among many others.

By finding relatively consistent correlations across large numbers of countries (20-60), all of these authors provide important new evidence to debates about policy effects within single countries. Additionally, several of these studies statistically control for other country factors that may confound the relationship between ILSA results and the country characteristic under investigation (Chmielewski & Reardon, 2016; Pfeffer, 2008; Torney-Purta et al., 2008). However, there still may be other unobserved country characteristics that cannot be included in models. Two studies move toward controlling for unobserved country characteristics by examining differences between their results and those for the same countries at earlier points in time.[4] Baker et al. (2002) examine differences in their results for less-developed countries in the 1990s to those of Heyneman and Loxley (1983) in the 1970s. By comparing changing results over time, Baker et al. (2002) avoid comparing countries at vastly different levels of development at a single point in time. However, because the sample of countries changed across the two studies, Baker et al. (2002) are comparing changes within a group of countries (less-developed countries), rather than changes within individual countries, over time. Similarly, Wiseman et al. (2009) compare the size of gender achievement gaps in math across FIMS 1964; SIMS 1980; and TIMSS 1995, 1999, and 2003. However, again because the sample of countries changed across the studies, they were not comparing changes within individual countries over time.

[3] Braun, Wang, Jenkins, and Weinbaum (2006) argue that without a policy history and a score trend it is possible to draw erroneous conclusions in these circumstances.

[4] Another possible technique to try to deal with unobserved factors at the country level would be to perform sensitivity analyses to estimate the potential impact of these unobservables on the outcome of interest (see Rosenbaum & Rubin, 1983).

Although none of the correlational analyses described in this section supports causal inference, they nevertheless give an important descriptive sense of what variation exists around the world, which is helpful for generating hypotheses and suggesting future areas of research.

A2. Random cross-national variation

In contrast with the correlational studies reviewed above, two studies have also examined country-level variation in policies, but used quasi-experimental research designs to attempt to identify random variation in national policies. Both studies use instrumental variables strategies to identify this random country-level variation.[5]

Falck, Heimisch, and Widerhold (2016) are interested in identifying the effect of workers' information and communications technology (ICT) skills on their earnings using data from 19 countries from PIAAC 2011. A simple correlation would confound ICT skills with individuals' general ability and the wealth of their country. Instead, the authors try to identify random variation in individuals' level of exposure to ICT by using an instrumental variable: international differences in types of preexisting telephone network infrastructure, which are related to the timing of broadband Internet introduction but not related to country wealth.[6]

West and Wößmann (2010) are interested in the relationship between the share of private schools in a country and its average level of math, science, and reading achievement in PISA 2003. However, the share of private schools is not randomly assigned and may be confounded with other factors that are related to student achievement. Thus, West and Wößmann also use an instrumental variables strategy to attempt to identify random variation in the expansion of the private school sector. In particular, using historical evidence the authors argue that the share of a country's population that was Catholic in 1900 is strongly related to the size of the contemporary private school sector, and that this variation is related to average country achievement, even after controlling for the size of the contemporary Catholic population.

B. Category 2: Within-country variation

The level of treatment in the previous papers could all be considered at the country level, even though some authors used within-country information to calculate the treatment variable by country. A number of other authors use ILSA data to examine issues where the level of treatment is within a country, such as classroom or month of birth level. Using the within-country variation often allows for more robust estimation strategies that can produce very strong correlational

[5] As with all research designs, there are drawbacks and limitations to instrumental variables strategies. The most difficult limitation to deal with as a researcher is the exclusion restriction because it is untestable. The researcher can only provide evidence that the instrument does not have an independent direct effect. See Murnane and Willett (2010) for a discussion of the limitations and assumptions needed for instrumental variables and other causal designs. In the analyses described in this section, both sets of authors provide evidence that their instruments are valid.

[6] They also examine the variation across German municipalities.

estimates or in even some cases causal estimates. The most popular estimation strategies in this category are fixed effects and/or instrumental variables models.[7]

An example of a study design using within-country variation in treatment can be found in the work of Bedard and Dhuey (2006). This paper examines how a child's age relative to his or her school entry cohort affects their academic performance. Thus, the level of treatment is the child's month of birth. In order to deal with nonrandom retention and school entry of children, the authors instrument the age at which a particular child entered school with the age he or she was assigned to start school.[8] By using ILSA data (pooled TIMSS 1995 and TIMSS 1999), they are also able to confirm that their results are consistent across a variety of countries with different education systems.

A large number of researchers have used variation within countries to attempt to identify the causal impact of classroom-level variables such as peer composition and class size on student achievement. In all of these cases, the treatment is at the classroom level. In one case, the authors argue that the variation across classrooms is random. Ammermueller and Pischke (2009) use PIRLS 2001 to estimate peer effects for fourth graders in six European countries. They argue that, in these six countries, elementary school students are assigned more or less randomly to classrooms—though not to schools—and therefore they use a school fixed-effect model to estimate the impact of the compositions of individual classes within a school. However, in many other countries, students are not randomly assigned to classrooms, so researchers must seek sources of random variation within the ILSA data. Three of the studies of class size effects exploit the unique design of TIMSS 1995, in which two adjacent grades were tested in each school (third and fourth grades, and seventh and eighth grades) (Ammermueller, Heijke, & Wößmann, 2005; Wößmann, 2005; Wößmann & West, 2006). Within each school, students in the two adjacent grades often have different class sizes. These differences are due not only to deliberate administrative and instructional decisions, but also to simple year-to-year fluctuation in the size of cohorts passing through the school. The authors reason that this latter source of variation was random and thus use variability in cohort size as an instrumental variable for class size, along with school fixed effects, to identify the causal impact of class size on student achievement. Altinok and Kingdon (2012) take a different approach with TIMSS 2003 by looking at variation *within students* in the sizes of their eighth grade math and science classrooms. Thus, unlike the previous authors, they use pupil rather than school fixed effects. This allows the authors to control for subject-invariant student and family unobservable characteristics.[9] Again, the ILSA data allow the authors to examine whether their findings hold across a large number of different countries.

[7] Kahn (2007) uses a difference-in-differences strategy to measure the impact of employment protection laws on employment. He uses a selection of countries from IALS94 and finds that, after controlling for skill levels, employment protection laws do not interact with the probability of having a job but they do decrease the probability of having a permanent rather than temporary job for low-skilled workers.

[8] The authors attempt to test some of the instrumental variables assumptions by showing evidence about birth date targeting by different socioeconomic groups and also by controlling for season of birth effects directly in the estimation.

The final type of design that uses within-country variation with cross-sectional data can be explored using the work by Brunello and Checchi (2007). This paper explores whether secondary school tracking strengthens the impact of family socioeconomic background on a variety of adult outcomes (literacy, college enrollment, earnings, etc.). The estimation strategy used is to compare different birth cohorts within each country who were exposed to different tracking policies, using models with country × year fixed effects. Thus, the treatment varies at the level of birth cohorts within countries. For their literacy results, the authors use IALS 1994 (which assessed adults ages 16-65) and split each country sample into 7-year age bands to identify different birth cohorts. The one difficulty of this estimation strategy is that, because IALS is cross sectional, earlier birth cohorts were surveyed at older ages, meaning the authors are unable to distinguish between cohort effects and age effects. The results for other outcomes (e.g., college enrollment and earnings) do not suffer from this limitation, as the authors were able to locate repeated cross-sectional data sets that surveyed different birth cohorts at the same age in different years. The studies in the next category all attempt such a strategy by treating ILSAs as repeated cross-sectional surveys.

C. Category 3: Variation across birth cohorts within countries

When multiple observations of a subject, whether a country or an individual, are available, the ability of the researcher to devise a causal estimation strategy increases. In this situation, the researcher can let the subject be their own control in the model, which allows any time-invariant characteristics of that subject to be omitted. In the ILSA context, researchers can combine multiple years of ILSAs to form a quasi-panel within countries, thus holding constant unobservable country characteristics. Compared to the over-time comparisons discussed in Category 1 (Baker et al., 2002; Wiseman et al., 2009), these designs support stronger causal inference because comparisons are made within countries over time, rather than cross-sectionally at each time point for a changing set of countries. The most common estimation strategies in this category are country fixed-effects models, hierarchical linear models (HLMs), and growth regressions. Thus, in this category, the level of treatment is always the birth cohort. Some studies in this category combine multiple cycles of the same ILSA, while others combine older and more recent tests; we discuss each in turn.

When researchers use multiple cycles of an ILSA, the assessment instruments are designed for measuring trends and therefore levels of performance can be compared more easily across time.[10] Several studies use country fixed-effects models to examine how within-country changes in policies are associated with within-country changes in mean achievement. For example, Hanushek, Link, and Wößmann (2013) use a quasi-panel of PISA administrations to estimate a country fixed-effects model of the effect of school autonomy on student achievement. They use changes in school autonomy over time in particular countries for identification. The country-level fixed effects control for systematic time-invariant cultural and institutional differences at the country level. They also aggregate the data to the country level to ensure that their estimates

[10] Even when using multiple cycles of the same ILSA, it is still an empirical question whether the instruments are perfectly comparable. For example, the length of test booklets was reduced between the 2003 and 2007 cycles of TIMSS and a bridging study undertaken to ensure comparability of TIMSS 2003 and 2007 scales (Olson, Martin, & Mullis, 2008). Researchers are responsible for informing themselves about methods used for scaling trend instruments and how they may affect results from secondary analyses of multiple cycles of ILSAs.

were not affected by within-country selection. Despite this, there is still a possibility that the results of this kind of model are contaminated by other correlated factors or policies. Similarly, Gustafsson (2013) uses two waves of TIMSS (2003 and 2007) to estimate the effect of homework on student achievement by estimating the association between within-country changes in homework time and within-country changes in mean math achievement. Rosen and Gustafsson (2016) model change between the IEA Reading Literacy Study (1991) and PIRLS 2001, and separately model change between PIRLS 2001 and 2006, in order to estimate the effect of home computer use on reading achievement.[11]

Rather than a fixed-effects framework, some authors have used HLMs with country random effects to estimate changes within countries over time. A well-done paper using HLMs descriptively rather than making causal inferences about an education policy is that of Rutkowski, Rutkowski, and Plucker (2012). They construct a quasi-panel of four cycles of TIMSS to describe trends in "excellence gaps" by gender and immigration status—the proportion of students in each group reaching an advanced achievement level on the TIMSS test. Chmielewski and Savage (2015) also use HLMs along with a quasi-panel of multiple ILSA cycles. Specifically, they examine trends in between-school socioeconomic segregation in 10 countries using five cycles of PISA. In this case, Chmielewski and Savage are interested not only in describing trends but also in using changes in policy variables (e.g., private schooling and enrollment rates) to predict changes in segregation. In order to hold constant unobservable country differences in the HLM framework, Chmielewski and Savage country-mean center (i.e., "demean") all time-varying policy variables. Thus, their approach is closely related to a country fixed-effects model.

All of the above referenced studies needed to make assumptions that the instruments associated with each wave of the ILSA are comparable over time and, therefore, levels of performance can be compared across time. The next papers also use a quasi-panel of ILSA data, but these cases combine different older and more recent tests where the instruments associated with each wave were not designed to measure trends. Authors use various methods to attempt to equate the different ILSA scales, all of which rely on different sets of assumptions. We discuss these methods in detail in the Recommendations section.

One of the first studies to combine older and more recent tests is that of Hanushek and Kimko (2000), who drew on ILSAs over a number of decades (FIMS, FISS, SIMS, SISS, IEAP-I, and IEAP-II).[12] The authors estimated countries' average performance on these ILSAs in order to model the effect of cognitive skills on macroeconomic growth.[13] Thus, in this study, the authors did not set up the older and more recent ILSAs as a quasi-panel or estimate within-country changes in skills, but rather countries' average level of skills across the whole time period. However, in the large economic literature following Hanushek and Kimko, there are several studies that do construct quasi-panels in order to rule out unobserved country-level effects by

[11] Both Gustafsson (2013) and Rosen and Gustafsson (2016) refer to their models as difference-in-differences models, though the most common name for such models in econometrics would be first-difference models. In this case with two time points, the first difference estimator is equivalent to fixed effects.
[12] Lee and Barro (2001) extend this data set to investigate the cross-country determinants of educational quality.
[13] Altinok and Murseli (2007) use a different methodology to adjust the test scores over time. In particular, they choose not to adjust the survey variance to allow for differences in variance between surveys but find comparable results.

focusing on within-country changes.[14] For example, Hanushek and Wößmann (2012a) extend Hanushek and Kimko's set of ILSAs by including TIMSS (1995-2003) and PISA (2000-2003) and use growth regressions to estimate the association between within-country changes in test scores and within-country changes in economic growth.

Trying to estimate causal effects is difficult in these data situations. However, using within-country variation over time by constructing a quasi-panel of a number of different ILSA data sets, Gundlach, Wößmann, and Gmelin (2001) also use growth regressions along with a quasi-panel of different ILSAs (FIMS, FISS, SIMS, SISS, and TIMSS 1995). They are interested in estimating change in "schooling productivity" internationally, that is, whether within-country increases in real expenditures correspond to increasing student performance over the same time period.[15]

Finally, Falch and Fischer (2012) construct a quasi-panel of six different older and more recent ILSAs in order to examine the effect of decentralization of public-sector spending on school performance.[16] They use a country fixed-effects model to investigate whether within-country changes in government spending decentralization are associated with changes in student achievement.[17]

D. Category 4: Variation in age (synthetic cohorts)

Like the third category, the fourth category also consists of study designs that use repeated cross-sectional data to observe change within countries over time. However, in this category, authors are interested not in educational change across different cohorts or generations but instead in how educational variables evolve as students age and move through the education system. Compared to the third category, this means that studies in the fourth category are not only allowing each country to act as its own control but allowing each *country cohort* to act as its own control and account for both unobserved country *and cohort* characteristics. Ideally, longitudinal data would be used to follow the same students over time in each country. However, because there are no longitudinal ILSAs, these authors instead match different cross-sectional ILSAs by the birth year of the test population to form what are sometimes called "synthetic cohorts." Because each ILSA tests a nationally representative sample of the same birth cohort, such a design can theoretically provide approximate estimates of how skills changed in the birth cohort in the interval between the two tests.[18] This design is useful in education policy research because measuring changing outcomes between the two time points can help to identify the causal impact of a policy that the cohort experienced during the interval.

---

[14] See Barro (2001); Gundlach, Rudman, and Wößmann (2002); Bosworth and Collins (2003); Ramirez, Luo, Schofer, and Meyer (2006); Jamison, Jamison, and Hanushek (2007); Altinok (2007); Hanushek and Wößmann (2008); Ciccone and Papaioannuou (2009); Appleton, Atherton, and Bleaney (2008); and Hanushek and Wößmann (2012b).

[15] Gundlach and Wößmann (2001) do a similar exercise using Asian countries and find similar results.

[16] They used data from the SIMS 1980, SISS 1984, International Assessment of Education Progress (IAEP) 1990/1991, TIMSS 1994/1995, TIMSS 1998/1999, and PISA 2000.

[17] Hanushek and Wößmann (2011) suggest that the identification coming from short-term variations in government measures that has immediate effects on student achievement warrants more investigation.

[18] Some authors have argued that synthetic cohort data are superior to longitudinal data because they are not subject to attrition bias or retest effects (Salthouse, 2009, 2010).

One well-known study in this category is by Hanushek and Wößmann (2006). The authors are interested in the effects of curricular tracking on country average performance and inequality. Cross-sectional correlations investigating these relationships with data from a single ILSA may conflate country tracking policies with other educational policies or characteristics that also influence performance and/or inequality. Comparisons across different generations within a country that changed its tracking policy (as in Category 3) may conflate the tracking policy change with other policy or social changes experienced by the different birth cohorts. Therefore, Hanushek and Wößmann match assessments of fourth grade students (TIMSS fourth grade and PIRLS) with assessments of lower secondary school students (TIMSS eighth grade and PISA). This allows them to use a difference-in-differences design to compare the changes in math, reading, and science performance and inequality in countries that begin curricular tracking between fourth grade and lower secondary school and the countries that do not begin curricular tracking in this interval. Rather than comparing the absolute level of performance or inequality across countries or cohorts, they compare changes *within* country cohorts.

The synthetic cohorts design requires the two matched data sets to be sampled from an identical population, including the same birth year. Hanushek and Wößmann match some contemporaneous ILSAs, meaning students were born in different years (TIMSS 1995 fourth and eighth grades; TIMSS 2003 fourth and eighth grades; and PIRLS 2001 and PISA 2000), as well as some ILSAs from different years where the students were born in the same or nearly the same year (PIRLS 2001, born in 1991; PISA 2003, born in 1988; and TIMSS 1995 fourth grade and TIMSS 1999 eighth grade, both born in 1985). Several other studies have used similar difference-in-differences designs with synthetic cohort data to investigate changes in inequality in achievement between primary and secondary school in early and late-tracking countries. Ammermueller (2013) matches only 2 contemporaneous ILSAs (PIRLS 2001 and PISA 2000), Jakubowski (2010) matches the same 8 pairs of tests as Hanushek and Wößmann, and Ruhose and Schwerdt (2016) match 18 different pairs of contemporaneous or birth year–matched tests. While Hanushek and Wößmann argue that contemporaneous test pairs should be preferred over birth year–matched pairs to avoid contamination by possible changes in other school policies between the test years, Ruhose and Schwerdt contend that contemporaneous pairs could be biased by cohort effects while birth year–matched pairs could be biased by calendar time effects, both of which are potential problems. Aside from ensuring that cohorts have the same birth year, a number of other issues arise in the comparability of different test populations, which are discussed below in the section on methodological recommendations.

A number of other authors have matched cross-sectional ILSAs to form synthetic cohorts in order to describe changes in skills as a cohort ages, without focusing on identifying the causal impact of a particular policy like the studies of curricular tracking above. For example, Schmidt et al.'s well-known 2001 book *Why Schools Matter: A Cross-National Comparison of Curriculum and Learning* exploits the fact that TIMSS 1995 administered the same assessments to two pairs of adjacent grades (third and fourth, and seventh and eighth grades) to attempt to estimate how much mathematics and science students learn in one year in different countries. The authors assume that the difference between being born in 1986 versus 1985 (third or fourth grade) and the difference between being born in 1982 versus 1981 (seventh or eighth grade)—cohort effects—are less consequential than the effect of having an extra year of schooling. Merry

(2013) matches national early childhood assessments from two countries, the United States and Canada,[19] to PISA 2000, 2003, and 2009 data for both countries in order to compare changes in the mean and distribution of performance from school entry to age 15 across the two countries. Because there are many educational, social policy, and social context differences between the United States and Canada, Merry does not attempt to identify the causal impact of a single policy but rather to compare the collective impact of the two systems.

Some authors have matched ILSAs by birth year to surveys of adult skills in order to observe how literacy and numeracy skills change from childhood to adulthood. For example, in an Organisation for Economic Co-operation and Development (OECD) brief, Thorn and Montt (OECD, 2014b) match PISA 2000 data with 26- to 28-year-olds in PIAAC 2011 and PISA 2003 data with 23- to 25-year-olds in PIAAC 2011 in order to compare skills growth between these ages across countries. Examining a much longer passage of time, Skirbekk et al. (2014) match data for 13-year-olds from FIMS 1964 with adult skills for 52- to 55-year-olds from the 2004 Survey of Health, Aging, and Retirement in Europe in five countries for the birth cohort 1949-1952. Again, neither set of authors is focused on identifying the causal impact of a single education policy, but instead describing the strength of the correlation between country results in childhood versus adult assessments.

Studies that match data for only one or two successive birth cohorts encounter a similar issue to Brunello and Checchi's (2007) study (described in Category 2) that compares multiple birth cohorts at a single cross-sectional point in time. Likewise, a study that follows only one cohort over time cannot distinguish between cohort effects and age effects. Several studies have explicitly attempted to disentangle cohort effects from age effects by matching data to form a large number of different synthetic cohorts born in different years. Green and Riddell (2013) match IALS 1994 and ALL 2003 data for five different birth cohorts across three countries (Canada, Norway, and the United States). Green and Riddell's birth cohorts are defined by 10-year bands due to limitations of available age data in some countries. They find that cross-sectional data underestimate age-related decline in cognitive skills, while analyses with synthetic cohort data reveal larger declines with age in all three countries. Paccagnella (2016) uses data from IALS 1994, ALL 2003, and PIAAC 2011 to match four birth cohorts (also defined by 10-year bands) across a large number of countries. He similarly finds substantial declines in skills by age in most countries. Gustafsson (2016) estimates cross-cohort country trends in achievement across five administrations of PISA (2000, 2003, 2006, 2009, and 2012) and compares them to country trends in skills across two cohorts of PIAAC 2011 (16-19 years and 25-29 years). He finds a positive correlation between countries' changes in achievement across the five PISA birth cohorts and changes in adult skills across the two PIAAC cohorts.

E. Category 5: Variation in age (longitudinal)

The last section includes the small number of ILSA studies that are truly longitudinal in nature, following the same sample of individuals over time. No ILSA has ever incorporated a longitudinal component, with the exception of the SIMS (1980), which tested eighth grade

[19] Merry uses early childhood data from the National Longitudinal Study of Youth 1979—Children and Youth (NLSY79) for the United States and the National Longitudinal Study of Children and Youth (NLSCY) for Canada, both of which administer the Peabody Picture Vocabulary Test–Revised (PPVT-R) to 4- to 5-year-old children.

students after conducting a pretest one year earlier in a limited number of countries. Twenty years later, these pre/posttest data for five countries were used by Zimmer and Toma (2000) to examine the effect of having high-achieving classmates using value-added models. Simply examining the cross-sectional correlation between students' own achievement and that of their classmates would overestimate peer effects because eighth grade students are not randomly assigned to classrooms in many countries. Controlling for prior achievement removes a major source of bias, but there are still many other student, classroom, and school characteristics that could confound the relationship between peer achievement and students' value added, many of which the authors control for in their models; they also include country and school-sector (public versus private) fixed effects.

Since 1980, no ILSA has included a cross-national follow-up. However, several countries have conducted their own national longitudinal studies following PISA students after age 15. For example, Canada's Youth in Transition Survey (YITS), Denmark's PISA-Longitudinal (PISA-L) survey, and Switzerland's Transitions from Education to Employment survey followed the PISA 2000 sample, while the Longitudinal Survey of Australian Youth (LSAY) followed the PISA 2003 sample. All longitudinal PISA follow-ups survey participants on educational and employment experiences. In addition, two countries assessed literacy skills in young adulthood: Canada's YITS conducted a Reading Skills Reassessment in 2009, which administered selected PISA questions to the sample as 24-year-olds (OECD, 2012), and Denmark's PISA-PIAAC study administered the PIAAC skills survey to the sample in 2012 when they were 27 years old (The Ministry of Education, 2014). Although these longitudinal PISA-related studies present a unique opportunity for cross-national analysis of education policy effects, they have not been extensively used for cross-national comparisons or causal inference.

Jakubowski (2013) used the Australian LSAY data to analyze the predictive power of PISA 2003 test items for students' future educational attainment in 2007 and 2010. He found that, after adjusting for students' overall score, mathematics items related to proportional reasoning and statistics were most predictive of higher educational qualifications at ages 20 and 23. Jaeger (2009) used the 2004 follow-up of the Danish PISA-L study to model the effect of parental cultural capital on students' choice at age 16 of the upper secondary academic or vocational track or leaving school. He estimates joint statistical models predicting the effects of students' reading ability and cultural capital on their secondary school choices and predicting the effects of parental investment on students' reading ability and cultural capital, while including random effects for three latent classes of families to account for unobserved family characteristics.

Finally, Jerrim and colleagues appear to be the only authors who have used multiple PISA longitudinal follow-ups for cross-national comparison. These authors combine data from four anglophone countries: LSAY in Australia and YITS in Canada (in one study), as well as the Longitudinal Study of Young People in England (LSYPE), using a mapping of LSYPE math and reading scores onto the PISA scale by Micklewright and Schnepf (2006), and the U.S. Educational Longitudinal Study of 2002, which administered selected PISA items, allowing estimation of PISA scores. Across the selected countries, Jerrim and colleagues compare low–socioeconomic status students' likelihood of entering bachelor's programs (Jerrim & Vignoles, 2015) and entering selective universities (Jerrim, Chmielewski, & Parker, 2015), as well as the

advantage of private secondary school attendance on educational and occupational attainment (Jerrim, Parker, Chmielewski, & Anders, 2016), all while controlling for PISA scores at age 15.

III. Methodological Recommendations

In this section, we discuss several methodological challenges confronting analyses of ILSA data. We focus on issues particularly relevant to designs addressing causal questions in education policy using ILSA data. We do not discuss general challenges that face researchers attempting causal inference but only focus on issues that are unique to using ILSA data. For each issue, we provide examples of different strategies authors have used, as well as our own recommendations. We first discuss two issues relevant to all ILSA analyses: (A) cross-cultural equivalence and (B) weighting and variance estimation adjustments. We then discuss three issues specific to many of the designs above that combine data from more than one ILSA: (C) combining different test instruments, (D) differences in target populations, and (E) differences in background questionnaire wording.

A. Cross-cultural equivalence

As in any cross-national policy research, the validity of causal research using ILSA data relies on the assumptions that (1) the causal question and research design is equally valid in all countries and (2) all measures have equivalent meaning across countries. Formulating causal questions and designs that are equally valid across units is an issue in cross-state, cross-city, or cross-school comparisons, but it becomes more challenging in cross-national comparisons. Authors of other reviews of causal research using ILSAs have described this challenge well. Kaplan (2016) writes that the counterfactual must be possible in all countries, the process of selection into the treatment must be the same in all countries (or if not, the researcher must be able to model the differences), and the treatment itself must be the same in all countries. Similarly, Rutkowski and Delandshere (2016) caution that causal mechanisms may operate differently depending on national context. Thus, the researcher must become well informed about the policy context of each included country. For example, in their study of age effects, Bedard and Dhuey (2006) conducted extensive background research on school entry cutoff dates and excluded any countries with ambiguous cutoff dates from their analysis to ensure that the treatment was comparable across countries.

Regarding the cross-cultural equivalence of ILSA measures, there is a large literature on the cross-cultural comparability of ILSAs and cross-national data more generally, including construct and measurement invariance of assessments and background questionnaire items and constructed scales. ILSA data undergo extensive validation by international experts during framework development, item development, translation, national adaptation, piloting, and field testing (Mullis, Martin, Ruddock, O'Sullivan, & Preuschoff, 2009; OECD, 2013). Nevertheless, it is important for secondary analysts without extensive psychometric expertise to familiarize themselves with the technical background, strengths, and limitations of scales available in ILSA data. For example, research on response patterns on Likert-type scales shows cultural differences in the propensity to give extreme versus midpoint responses (Chen, Lee, & Stevenson, 1995). Thus, researchers should be cautious about interpreting cross-national differences in self-ratings of personal characteristics such as self-efficacy as true differences.

Even the more objective-seeming categories such as educational levels may have different meaning across countries. The International Standard Classification of Education (ISCED) scheme is intended to enable comparisons of education programs across countries, and ISCED is extensively used in ILSAs for comparisons between subjects' current educational programs, their parents' educational attainment, and their own educational aspirations. However, the ISCED scheme itself is the subject of ongoing research into its cross-national validity and comparability (Schneider, 2008). The ISCED scheme is discussed in further detail in Section E below.

B. Weighting and variance estimation adjustments

All of the ILSAs discussed above were designed to provide valid, nationally representative measures of achievement of a target population. However, assessing the entire target population on all test items is generally too expensive and too time consuming. Therefore, some samples of individuals are assessed on only some of the items. Within this sample, it is often the case that not all students have equal probabilities of selection, which means the sample must be weighted to produce nationally representative results. Additionally, the sample is often drawn in two steps to form a two-stage stratified cluster sample design where the first stage is the selection of the schools and the second stage is the selection of students or intact classes within schools, which introduces uncertainty into the results. Finally, many ILSAs use rotated booklet designs in which not all students take all test items, which introduces further uncertainty. All ILSAs have technical reports that outline the specific sampling procedures employed. Researchers need to carefully consider weighting and variance estimation techniques when performing analysis using the data. We discuss each kind of adjustment in turn. We focus first on using proper weighting, which allows for the correct estimation of *point estimates,* and then on using proper variance estimation techniques, which allows for the correct estimation of the associated *standard errors*.

In order to adjust for selection probabilities and sampling outcomes, most ILSA databases will provide at least two levels of weights: school and student (and classroom, if applicable). At each of these levels, the weights are generally the inverse of the probability of selection at that particular level together with an adjustment for nonparticipation. Using the proper weights is vital to obtain unbiased descriptive statistics and point estimates from predictive models. Meinck (2015) shows why sampling weights for large-scale assessments are important by demonstrating the potential consequences of not using weights from assessments with complex sampling schemes.

Although it is relatively straightforward to use proper school, class, and student weights, researchers conducting pooled analyses of multiple countries must make more difficult decisions about country weights. There are at least three possible approaches to country weights: (1) weight each country in proportion to the size of its target population, (2) weight all countries equally ("senate weighting"), or (3) weight each country in proportion to the size of its test sample ("house weighting"). The total student weights provided in most ILSA databases (e.g., PISA, TIMSS, and PIRLS) are multiplied by a different constant in each country such that they sum to the total size of its target population (e.g., the total population of 15-year-old or fourth grade students). Thus, if a researcher pools all countries in an analysis and applies only the total student weight, more populous countries will contribute more to the results. This is usually not

desirable. For researchers who instead wish for each country to contribute equally in the analysis, most ILSAs include a "senate weight" or "weight factor" that can be used to transform the weights so that the sum of each country equals a constant. It is important to note that the constant varies by ILSA. For instance, the "senate weight" in PIRLS adds up to 500 within a country whereas the "weight factor" for PISA equals 1000. Therefore, if the researcher is combining tests, he or she will need to adjust these weights so that the constant is equivalent across countries and assessments.[20] Finally, for analyses which are sensitive to sample sizes (e.g., chi-square tests), "house" or "normalized" weights should be used to make the sum of the weights within each country equal to its sample size.[21] The choice between the three types of country weights should depend on the research question and the analyses used. However, it can be useful for authors to test the sensitivity of their results to choice of country weight. For example, using PISA 2003, Chmielewski et al. (2013) run two versions of their analyses, one weighting countries proportional to their 15-year-old population and one weighting all countries equally, and confirm that their results hold in both cases.

Using correct weights is further complicated when using more complex models. For example, when using a hierarchical linear model, researchers should generally apply the school weight at the school level but should not apply the total weight at the student level because this weight includes the school weight, and thus would double-weight each observation by its school weight. Instead, the conditional within-school weight should be used at the student level. However, no such weight is currently provided in ILSA databases. Rabe-Hesketh and Skrondal (2006) recommend that researchers divide the total weight by the school weight to create conditional within-school student weights. This method has been used in some published research (e.g., Chmielewski, 2014; Chmielewski et al., 2013). However, many authors using hierarchical models with ILSA data do not clearly explain which weights are applied at each level of the model. Some researchers may be able to skip this confusion by avoiding hierarchical models altogether. Huang (2014) shows in simulations that, in analyses using only student-level variables, other alternatives to hierarchical modeling (such as Taylor-series linearization, the design effect standard errors approach, and fixed-effects modeling) produce very similar standard errors. This conclusion likely also extends to the variance estimation techniques discussed later in this section, such as balanced repeated replication and jackknife repeated replication. However, in practice, most policy-focused research will not use only student-level variables but will also be interested in classroom-, school-, or country-level treatment variables. Thus, proper weighting in hierarchical models remains an important issue.

Finally, weighting is particularly difficult when combining multiple tests. Jakubowski and Pokropek (2015) highlight this difficulty in their careful description of analyses combining data from PISA and PIRLS, which required significant (author) reweighting. In particular, they attempt to correct for differences in the target populations of PISA and PIRLS by reweighting the PIRLS data to make student background characteristics more similar to those in the PISA sample (differences in target populations are further discussed in a later section of this paper). In performing this reweighting, Jakubowski and Pokropek follow Tarozzi (2007) and multiply the original survey weights by conditional probabilities estimated using logit regression with a

[20] See Gonzalez (2012) for a useful discussion of rescaling sampling weights and selecting minisamples from ILSA data.
[21] See Rutkowki et al. (2010) for more details regarding "house" weights and other weighting schemes.

dependent variable equal to 1 for PISA and 0 for PIRLS and a set of student characteristics as control variables. They argue that this type of reweighting produces data sets that are balanced in terms of background characteristics, which allows for better comparisons of outcomes.

When authors need to transform weights in any of the situations described above, generally, the information needed to create new weights is included in the data and technical user guides. However, this is not always the case. Jakubowski (2010) provides an example of a situation where sampling weights needed to be corrected but could not be because the needed information on sampling, nonresponse, and weight adjustment was not provided in the documentation. However, in this situation, the author performs a sensitivity analysis: He replicates the analysis with the incorrect weights and with no weights and finds relatively similar results. Jakubowski therefore is able to provide some evidence that the conclusions would be similar regardless of the weighting system. Overall, it is clear that correctly using sampling weights is vitally important to estimating unbiased estimates but it is often less clear how to use weights in more complex analysis situations.[22]

All of the weighting issues discussed in the first part of this section are necessary to obtain correct, nationally representative *point estimates*. Now we turn to variance estimation techniques for obtaining correct *standard errors*. In ILSA studies, not all students in a population are tested and the ones who are tested are not administered all the items. Therefore, it is important not to calculate standard errors using the standard variance formula, as the standard errors of estimates from these studies have two additional components. The first is sampling error resulting from sampling students from a population in a stratified multistage sampling design. The second is measurement error (specifically, imputation error) resulting from sampling assessment items from a universe of items. To correct for the first component, there are a variety of approaches, the most common of which are replication-based variation estimation procedures such as jackknife repeated replication (JRR) and balanced repeated replication (BRR). All of these methods work by drawing subsamples from the full data set, computing the statistic of interest for each subsample, and then using the variance among subsamples to estimate sampling variation. Replicate weights for use in these procedures are provided in ILSA databases. Generally, IEA studies (e.g., TIMSS and PIRLS) provide JRR weights, PISA provides BRR weights, and PIAAC provides weights that can be used for BRR or various jackknifing procedures. However, in theory, any of these variance estimation procedures could be used with any database, and research shows that no particular replication-based procedure performs better than the others (E. S. Lee, Forthofer, & Lorimor, 1989). Thus, it may be preferable for the researcher to use the same variance estimation procedure across all databases so that standard errors will be comparable. However, the full information needed for alternate variance estimation procedures may not be provided in every database. We could find no examples of published research using the same variance estimation procedure with multiple ILSAs.

The second error component is measurement error in estimating achievement. Large-scale assessments try to limit the burden on individual students while still enabling accurate population estimates by limiting the number of assessment questions a particular student has to take.

[22] See Rutkowski et al. (2010) for a more in-depth treatment of weights along with examples of how to calculate and which weights are appropriate for many multilevel model situations.

Each participating student receives only a subset of items through a matrix-sampling or balanced incomplete block test design. As a result, precise estimates of achievement can be generated for national populations and subpopulations but not for individual students. Instead, since 1995, all ILSA databases provide a number of plausible values of achievement for each student.[23] The plausible values are generated through multiple imputation methods (Rubin, 1987; Schafer, 2003) by using student responses to the subset of items they received as well as background characteristics to estimate a distribution of proficiency for each student and then make a number of random draws from that distribution.[24] Researchers using ILSA data should then use the variability among the provided plausible values to estimate imputation variance. In the literature reviewed in this paper, we found several examples of authors who report using all provided plausible values to estimate standard errors (Ammermueller, 2013; Ammermueller & Pischke, 2009; Chmielewski, 2014; Chmielewski et al., 2013; Chudgar & Luschei, 2009; Gustafsson, 2016; Rosén & Gustafsson, 2016; D. Rutkowski et al., 2012; Schmidt et al., 2015). A small number of authors report averaging the plausible values, which produces biased estimates. Several authors report using only one of the plausible values, which yields unbiased point estimates but underestimates standard errors and, thus, is considered appropriate only for exploratory analyses. A small number of authors appear to have the misconception that plausible values need only be used when achievement is the outcome; however, plausible values also need to be used when achievement is an independent variable.[25] The majority of the studies reviewed in this paper make no mention of plausible values, so we cannot tell whether the authors make proper use of this method. Correctly using plausible values is generally not difficult. However, it is more complex when combining different tests. Older ILSAs (e.g., FIMS, SIMS) do not come with plausible values, as they did not use rotated booklet designs. Combining ILSAs using a different number of plausible values is also complicated. For example, Gustafsson (2016) matches birth cohorts from several cycles of PISA (2000-2012), which have five plausible values, to PIAAC 2011, which has ten plausible values. He uses all plausible values from each ILSA, but this is only possible because he does not pool observations from different ILSAs but instead computes separate estimates of trends in skills across the years of each ILSA, meaning he can use different methods to compute the standard errors for each ILSA.

Among the previously reviewed articles, it appears that nearly all authors apply weights. However, despite the importance of variance estimation techniques for correcting standard errors, a large number the articles either do not employ variance estimation methods correctly or do not provide details on their procedures.[26] We believe that this may be largely due to the difficulty of using such methods in the context of complex research designs without built-in commands in the usual statistical software. To facilitate proper ILSA analyses, the IEA Data Processing and Research Center has released a program for the SPSS software package called IDB Analyzer. The program incorporates weights, replication-based variation estimation procedures, and plausible values, for both IEA and OECD studies. However, the program runs only very simple analyses such as means, frequencies, and ordinary least squares regression and

[23] Balanced incomplete block test design and plausible values methodology are also used in the National Assessment of Educational Progress (NAEP) and other large-scale assessments.
[24] TIMSS, PIRLS, and early years of PISA provide five plausible values of achievement. PISA 2015 and PIAAC 2011 provide ten plausible values of achievement.
[25] See Rutkowski et al. (2010) for a nice example of problems that occur by not following the correct methods for using plausible values.
[26] Most of the economics research that is reviewed uses clustering of standard errors as a correction technique.

cannot be used for many of the more complex designs described in this paper (hierarchical models or any analysis combining more than one ILSA). Thus, researchers must attempt to program these methods in their own preferred software. In the case of plausible values, it will often be possible to use built-in commands designed for multiply imputed data. If not, it is relatively straightforward simply to run each model five (or ten) times, average the estimates, and compute standard errors using standard formulas for imputation variance. Replication-based variance estimation procedures can be implemented using built-in survey data commands (e.g., Stata's "svy") (Kreuter & Valliant, 2007), though such commands are not always easily paired with complex quasi-experimental models. Recently, the OECD has released the "repest" command for Stata (Avvisati & Keslair, 2014), which computes replicate weights and plausible values and is far more flexible than previously released macros in allowing a wide range of complex models. However, "repest" can be used only with OECD databases, not IEA. More work should be done in this area so that researchers have the tools needed to run the analyses using ILSA data. In the meantime, when standard error corrections have not been made, it is important for the readers to understand that the reported standard errors may be severely biased.

C. Combining different assessments

The most obvious challenge in many of the designs that combine data from different ILSAs is that test scores are derived from different instruments designed to measure different competencies, even though they have the same or similar labels.[27] That both PISA and the two largest IEA assessments (TIMSS and PIRLS) have international mean scores of 500 and standard deviations of 100 does not indicate that 1 point has equivalent meaning in all three tests. One crucial distinction in the assessment frameworks is that the TIMSS math and science tests are curriculum based, while PISA and PIRLS measure students' *literacy*, both in reading and—in the case of PISA—in math and science as well (Mullis et al., 2009; OECD, 2013). This means that, compared to the TIMSS math and science assessments, the PISA assessments in these subjects depend less on formal knowledge of laws and formulas and more on application of competencies to real-world situations. Additionally, item formats differ; PISA questions tend to require students to read more text, as well as to write more, as PISA contains more open-ended and fewer multiple choice items than TIMSS.

Despite these differences, the similar subject matter and target populations of PISA and TIMSS eighth grade have led some to question whether the two tests are similar enough as to be redundant (Hutchison & Schagen, 2007). Indeed, Wu (2009) shows that for 22 jurisdictions that participated in both PISA 2003 and TIMSS 2003, 93% of the discrepancy in countries' performance on the math assessments can be explained by only two factors: content balance and the additional years of schooling completed between eighth grade and age 15. Content balance refers not to orientation toward curriculum versus literacy but simply to the fact that TIMSS contains more algebra, geometry, and measurement questions and PISA more data and number questions. The importance of years of schooling is due to differences in the PISA and TIMSS target population definitions, which are discussed in more detail in the next section below. Wu's (2009) results suggest that one may be able to obtain at least approximately correct estimates from analyses that combine PISA and TIMSS scores, as long as one carefully models the time

[27] Braun and Mislevy (2005) do a good job discussing testing and the dangers associated with ignoring the principles and methods of educational measurement.

lag between the tests and accounts for content balance, such as by reweighting scores by content domain or analyzing domains separately. However, most authors reviewed above do not account for these two factors. Jakubowski (2010) does account for years of schooling, which is discussed further below in the section Changes in Target Population Across Assessments. We did not find any examples of authors who took into account differences in content balance. The best method for doing so is unclear, as a simple reweighting of PISA scores does not replicate the item response theory (IRT) model used to generate TIMSS scores (Wu, 2009). Still, exploratory weighting methods may be useful as a sensitivity analysis.

The discussion above focuses on PISA and TIMSS eighth grade assessments, which are most similar in subject matter and target populations. However, many of the studies reviewed here combine other tests, such as fourth grade and secondary school assessments, or historical ILSAs such as FIMS and SIMS with contemporary ILSAs such as PISA and TIMSS. Such analyses require a more stringent set of assumptions. For example, several authors use historical ILSAs to attempt to estimate long-term trends in countries' performance. Two main approaches are taken, one focusing on relative and the other on absolute performance. In the first approach, Falch and Fischer (2012) assume that the pooled mean performance of a "core" group of the most frequently participating countries is constant over time, set at zero. They then estimate the association between changes in their independent variable of interest, public-sector decentralization, and changes in countries' performance relative to the international mean. The stability of these estimates depends strongly on which countries are included in the "core" group of frequently participating countries, and it is not possible to draw any conclusions about absolute performance levels, as the standardization procedure cancels out any secular trends of globally increasing or declining performance across all or most participating countries. In the second approach, authors use an external anchor or benchmark to estimate absolute changes in performance over time. All authors in this group use as their external anchor the United States' NAEP. Hanushek and colleagues standardize scores in each ILSA to equate the U.S. mean with its NAEP performance (Hanushek & Kimko, 2000; Hanushek & Wößmann, 2012a). Gundlach et al. (2001) analyze ILSAs from FISS 1970 to TIMSS 1995, during which period they observe that U.S. NAEP performance changed very little. Thus, they assume U.S. performance was constant during this period and standardize all ILSAs accordingly. Hanushek and colleagues' method assumes that NAEP and ILSAs test identical skills. Gundlach et al.'s method assumes that the small fluctuations in U.S. mean NAEP scores between the 1970s and 1990s are not meaningful, and that U.S. students' math and science skills were truly constant during this period.[28] However, Gundlach et al. (2001) also check for robustness of their results by using different assumptions regarding mean and standard deviation of the test results.

In addition to estimating trends in countries' performance *levels*, many authors wish to compare the *dispersion* of scores over time. The most common approach is to assume that the pooled standard deviation of a "core" group of the most frequently participating countries is constant over time. This method has been used by several authors comparing long-term trends in performance (Falch & Fischer, 2012; Gundlach et al., 2001; Hanushek & Wößmann, 2012a), as well as several authors matching synthetic cohorts across fourth grade and secondary school

[28] There is evidence that there was a very small upward trend in mathematics scores for 9- and 13-year-olds from 1970 to 2000. In particular, the average math scale score out of 500 was 266 (219) in 1973 but 276 (232) in 1999 for 13- (9-)year olds (Kena et al., 2016).

using TIMSS, PIRLS, and PISA (Ammermueller, 2013; Jakubowski, 2010; Ruhose & Schwerdt, 2016). Results produced using this method should be interpreted only in terms of performance relative to this core group of countries, as it may be inaccurate to assume that true score dispersion does not change across birth cohorts or across age groups. In contrast with this more common strategy, Hanushek and Wößmann (2006) compute the standard deviation of each country's scores in the original scale of each test and then standardize these measures to the average national standard deviation of each test. This method also requires the authors to use one constant "core" group of countries that does not change over time, and it also requires the assumption that, within this "core" group, the standard deviations among these countries is constant over time.

Finally, when authors are examining neither absolute performance nor dispersion of scores but disparities in performance between different groups, another approach is to treat combining scores from different ILSAs as a meta-analysis. Authors standardize scores within each country for each ILSA and then compute effect sizes for the score gaps between groups. For example, Chmielewski and Reardon (2016) use this method to compute achievement gaps between students from the 90th and 10th percentiles of household income in PIRLS and PISA. Andon et al. (2014) use a similar method to compute achievement gaps between native-born and immigrant students across numerous waves of PISA, TIMSS, and PIRLS. Using this method, one must assume only that the different tests rank students equivalently, not that scales are equivalent; and in fact one must not even assume that scales are equivalent for different countries participating in one wave of the same ILSA. However, authors must take care to interpret results in relative rather than absolute terms, as the countries with the largest achievement gaps in effect sizes may not be those with the largest absolute gaps in skills.

Each of the methods described above for combining scores from different ILSAs requires a set of assumptions that are unlikely to be fully met, as well as careful interpretation of results in light of the standardization procedures used. All of these problems can be avoided by conducting analyses using multiple cycles of the same ILSA, which have been specifically designed to measure trends (e.g., Gustafsson, 2013, 2016; Hanushek et al., 2013; D. Rutkowski et al., 2012). However, such analyses can only be conducted for a relatively recent time period (starting in 2000 for PISA, 1995 for TIMSS, and 2001 for PIRLS). Additionally, such analyses can only examine changes across different birth cohorts (Category 3) rather than across ages for the same birth cohort using a synthetic cohorts design (Category 4). Given that researchers wish to examine long-term trends and synthetic cohorts, more research is needed to understand the consequences of the standardization and benchmarking methods used above for the validity of conclusions. Comparisons between results obtained by combining ILSAs and those obtained using multiple waves of the same ILSA may prove useful in this regard.

D. Changes in target population across assessments

Another issue that arises when conducting analyses combining different ILSAs is that different ILSAs usually do not define their target populations in identical ways. Although there are many commonalities between the target population definitions in the OECD's PISA and the IEA's assessments (e.g., students must be enrolled in formal school, students may be excluded due to disability or low proficiency in the test language, schools may be excluded for geographical

inaccessibility or offering radically nonstandard programs, and total exclusions may not exceed 5% of a country's target population) (Martin & Mullis, 2013; OECD, 2014a), there is still at least one critical distinction. While OECD assessments such as PISA and PIAAC use age-based samples, contemporary IEA assessments such as TIMSS and PIRLS have grade-based samples (although some older IEA assessments, such as FISS, use age-based samples). This creates challenges whenever researchers combine different cross-sectional ILSAs to form quasi-panels, whether those panels consist of different birth cohorts or generations of students (Category 3 above) or synthetic birth cohorts of students born in the same year and tested in different years (Category 4 above). Age-based and grade-based samples represent virtually identical populations in countries like Japan that have strictly enforced school age cutoffs and very little grade retention, but researchers cannot assume samples are comparable in countries with more complex retention and promotion policies, where many students are outside of the age-appropriate grade. One solution used by Jakubowski (2010) is to restrict the PISA data to students enrolled in the modal grade for their country, as well as excluding any students who report ever repeating a grade. However, it should be noted that these sample restrictions still do not turn PISA from an age-based into a grade-based sample and it necessarily changes the external validity of the results. A grade-based sample should include students who are enrolled in the grade in question but are not the typical age due to previously repeating a grade or other factors, but the grade-restricted PISA sample excludes students who are enrolled in the modal grade but are not 15 years old. Rather than employing extreme sample restrictions, researchers may consider including indicator variables for students' age or grade when combining ILSAs with age-based and grade-based samples.

Synthetic cohort analyses (Category 4) require a particularly strict standard of comparability, because the repeated cross sections must be sampled from exactly the same population. For example, when comparing two grade-based samples (e.g., TIMSS fourth grade and eighth grade), any grade repetition that occurs between fourth and eighth grades makes the eighth grade sample not fully representative of the population of students who were in fourth grade 4 years previously. Additionally, true synthetic cohort data should theoretically include students who exit the target population between fourth and eighth grade, such as by enrolling in a special school for the disabled or dropping out of formal schooling. Finally, students who immigrate to the country in question between fourth and eighth grade should be excluded. Because it is difficult to know the exact timing of immigration, many synthetic cohort analyses exclude foreign-born individuals altogether (Green & Riddell, 2013; Gustafsson, 2016; Jakubowski, 2010; Paccagnella, 2016; Ruhose & Schwerdt, 2016). Although this exclusion creates more comparable synthetic cohort data, the tradeoff is that the analysis becomes less generalizable to the full national population, which does include foreign-born individuals who were fully or partially educated in the country in question.

The three international surveys of adult skills (IALS, ALL, and PIAAC) are age-based samples, and they are selected from households rather than schools. Because of this difference, synthetic cohorts created from adult surveys are subject to many of the comparability issues above (e.g., immigration). In addition, they introduce a wider range of types of sample attrition, including institutionalization, incarceration, and literal mortality. None of these issues can be directly measured in the available data which makes it difficult for researchers estimate the size of the

bias.[29] The household-based adult surveys can include individuals with no or very little formal education, which is an advantage for population representativeness but a disadvantage when matching with school-based childhood assessments, because adults who were not enrolled in school in fourth grade, in eighth grade, or at age 15 should be excluded. However, researchers are unable to make this kind of correction. Another difficulty arises because the surveys of adult skills sample a wide range of ages (16-65), which causes the sample of individuals born in a given year to be very small. Researchers often group respondents into 3-, 5-, or 10-year birth cohort bands (Brunello & Checchi, 2007; Green & Riddell, 2013; OECD, 2014b; Paccagnella, 2016), which means that individuals born as many as 9 years apart may be considered part of the same birth cohort. When matching these samples with childhood assessments to create synthetic cohorts, birth cohorts in the childhood data constitute much narrower bands, meaning that samples do not represent exactly the same population and also making it difficult to cleanly identify differences in education policies experienced by different birth cohorts. Also, when matching synthetic cohorts across different adult surveys, even if birth year bands are defined identically, the distribution of birth years may change within bands due to disproportionate mortality from the older years of the range.

The majority of the issues discussed above cannot be remedied by restricting samples according to measurable characteristics. Instead, careful researchers should conduct sensitivity analyses to attempt to estimate the size of the bias created (using techniques in the spirit of Horowiz and Manski [2000] and Lee [2002]) to estimate whether it is large enough to cancel out estimated effects. Finally, when combining age-based and grade-based ILSAs, researchers can incorporate the length of time elapsed between the tests into their models. Wu (2009) identifies international variation in the year(s) of formal education experienced between eighth grade and age 15 as a major source of discrepancies in country rankings in PISA 2003 and TIMSS 2003. Relatedly, Jakubowski (2010) advocates modeling the length of time elapsed between fourth grade assessments and PISA, rather than conducting a simple difference-in-differences analysis that treats all primary-secondary pairs as separated by the same length of time in every country.

E. Changes in background questionnaire wording

Another important issue to consider when working with ILSA data is that wording of the questionnaire often changes over time or is slightly different across waves or even different across countries. This makes it incredibly important for researchers to carefully examine the background questions used in successive questionnaires to make sure that they are comparable across countries, across waves, and across surveys.

A good example of a change in wording in a background question can be seen in the U.S. adaptation of PISA in 2003 and 2009. The general question that was used is the following: “How old were you when you started [ISCED 1]?” ISCED 1 indicates primary education. In the U.S. national adaptation of PISA in 2003, the questionnaire substituted “elementary school” for ISCED 1. In 2009, the questionnaire instead substituted the words “first grade.” In 2003, the average response in the United States was 5.38 but increased with the changed wording to 5.90 in 2009. The difference in average response indicates that these two questions may have been interpreted differently across years by individuals answering the background questionnaire. It is

[29] The United States recently conducted a PIAAC survey of incarcerated adults.

also important to note that even something as "standardized" as an ISCED level has actually changed over the years. This does not affect the previous example, but in 2011 the ISCED levels were updated so comparisons pre- and post-2011 would need to be carefully considered.

National adaptations are regularly made in an ILSA administration. They often include modifications to questions to suit the national context and, in some cases, questions are not administered in particular countries. Therefore, it is important that the researcher understands and carefully interprets any adaptations in the background questions used from the questionnaires. An example to highlight this issue comes from TIMSS 2011 eighth grade sample. The background question states, "How far in your education do you expect to go?" In the U.S. adaptation, the selection of category 5 indicated either a 2-year or a 4-year postsecondary degree. In this survey, 42.6% of responders choose category 5 as their answer. However, a geographically close neighbor to the United States is the province of Ontario, Canada. Ontario's adaptation for this question only included a 4-year bachelor's degree in category 5, and other postsecondary degrees were included in a different category. In Ontario, only 22.9% of students answered that particular background question as category 5 and 19.1% answered in the lower category 4. Making a naive comparison across the United States and Ontario, Canada, indicates that the students in the United States have higher aspirations when in fact if one adds both category 4 and 5 in Ontario, the level of aspirations is almost identical (42.6%).

Careful researchers must be sure to compare questionnaires across years, surveys, and national adaptations across countries to avoid making inappropriate comparisons. The IEA publishes all national adaptations in the documentation for each cycle of TIMSS and PIRLS, and the OECD publishes most—but not all—national adaptations in the documentation for each cycle of PISA. The remaining PISA national adaptations must be obtained from national project offices. For example, we obtained the U.S. PISA national adaptations above from information released by the U.S. National Center for Education Statistics, rather than the OECD.

Gustafasson (2013) provides a useful example that highlights the difficulty in working with quasi-panels constructed with ILSA data when different background questions are used in different waves. In particular, he is interested in understanding the effect of homework on achievement. However, while conducting this research he finds that different methods were used to measure the amount of time spent on homework in different TIMSS waves. In particular, in TIMSS 1995 and TIMSS Advanced 2008, students were asked to state the number of minutes per week they spend on homework in mathematics. In TIMSS 2003 and 2008, they were asked to separately indicate the frequency of homework per week and the amount of time typically spent on each assignment. Using an algorithm that combined the different definitions, Gustafsson (2013) computed an estimate of the total number of minutes spent on math homework per week across the waves. This is a good example of a careful researcher understanding the changes in background questions and trying to construct a common variable that can be used for comparisons across waves.

IV. Conclusions

The literature reviewed reveals a strikingly diverse set of approaches to the design and the methodological challenges inherent in ILSA analyses. We reviewed a number of different

designs and several of the studies have strong designs that approximate something close to random assignment and help to advance our understanding of education policy effects. In classifying designs by highlighting the variation in educational policies or "treatments" that students experience, we hope to inspire new and innovative approaches to using ILSA data to attempt to explore causal questions about education policy.

We also summarized a variety of important methodological issues that arise in the analysis of ILSA data and described the many different approaches taken by authors to address these issues. Although none of the approaches is perfect, some rely on stronger assumptions than others. Researchers should carefully consider how methodological decisions around issues such as weighting or combining ILSAs that use different test instruments or target populations may affect the interpretation as well as external validity of their results. As this literature grows, we will gain a better body of knowledge regarding education policy in the international context. In the meantime, more methodological research is needed to weigh the relative advantages of the different approaches. There are important roles to be played both by academic researchers as well as testing organizations such as the IEA and OECD and their contractors and by national governments in this research arena. Importantly, the IEA and OECD should enhance their roles in disseminating technical user guides and software programs and macros that support methodological best practices for the increasingly complex analyses carried out by education policy researchers using ILSA data.

V. References

Altinok, N. (2007). *Human Capital Quality and Economic Growth*. IREDU working paper.

Altinok, N., & Murseli, H. (2007). International database on Human Capital Quality. *Economics Letters, 96*, 237-244.

Altinok, N., & Kingdon, G. (2012). New evidence on class size effects: A pupil fixed effects approach. *Oxford Bulletin of Economics and Statistics, 74*(2), 203-234.

Ammermueller, A. (2013). Institutional Features of Schooling Systems and Educational Inequality: Cross-Country Evidence From PIRLS and PISA. *German Economic Review, 14*(2), 190-213.

Ammermueller, A., Heijke, H., & Wößmann, L. (2005). Schooling Quality in Eastern Europe: Educational Production during Transition. *Economics of Education Review, 24*(579-599).

Ammermueller, A., & Pischke, J.-S. (2009). Peer Effects in European Primary Schools: Evidence from the Progress in International Reading Literacy Study. *Journal of Labor Economics, 27*(3), 315-348.

Andon, A., Thompson, C. G., & Becker, B. J. (2014). A quantitative synthesis of the immigrant achievement gap across OECD countries. *Large-scale Assessments in Education, 2*(1), 1.

Appleton, S., Atherton, P., & Bleaney, M. (2008). *International School Test Scores and Economic Growth*. Retrieved from CREDIT Research Paper No. 08/04.

Avvisati, F., & Keslair, F. (2014). REPEST: Stata module to run estimations with weighted replicate samples and plausible values. Accessed 19/05/2015 from https://ideas.repec.org/c/boc/bocode/s457918.html.

Baker, D. P., Goesling, B., & LeTendre, G. K. (2002). Socioeconomic status, school quality, and national economic development: a cross-national analysis of the "Heyneman-Loxley

Effect" on Mathematics and Science achievement. *Comparative Education Review, 46*(3), 291-312.
Barro, R. J. (2001). Human Capital and Growth. *American Economic Review, 91*(2), 12-17.
Bedard, K., & Dhuey, E. (2006). The persistence of early childhood maturity: International evidence of long-run age effects. *The Quarterly Journal of Economics*, 1437-1472.
Bosworth, B., & Collins, S. (2003). The Empirics of Growth: An Update. *Brookings Papers on Economic Activity, 2*(113-206).
Braun, H. I., & Mislevy, R. (2005). Intuitive test theory. *Phi Delta Kappan, 86*(7), 488-497.
Braun, H. I., Wang, A., Jenkins, F., & Weinbaum, E. (2006). The Black-White Achievement Gap: Do State Policies Matter? *Education Policy Analysis Archives, 14*(8), 1-110.
Brunello, G., & Checchi, D. (2007). Does school tracking affect equality of opportunity? New international evidence. *Economic Policy, 22*(52), 781-861.
Buchmann, C., & Park, H. (2009). Stratification and the formation of expectations in highly differentiated educational systems. *Research in Social Stratification and Mobility, 27*(4), 245-267.
Chen, C., Lee, S.-y., & Stevenson, H. W. (1995). Response style and cross-cultural comparisons of rating scales among East Asian and North American students. *Psychological Science*, 170-175.
Chmielewski, A. K. (2014). An International Comparison of Achievement Inequality in Within- and Between-School Tracking Systems. *American Journal of Education, 120*(3), 293-324.
Chmielewski, A. K., Dumont, H., & Trautwein, U. (2013). Tracking Effects Depend on Tracking Type: An International Comparison of Students' Mathematics Self-Concept. *American Educational Research Journal, 50*(5), 925-957.
Chmielewski, A. K., & Reardon, S. F. (2016). Patterns of Cross-National Variation in the Association Between Income and Academic Achievement. *AERA Open, 2*(3), 1-27.
Chmielewski, A. K., & Savage, C. (2015). *Socioeconomic Segregation Between Schools in the United States and Latin America, 1970-2012.* Paper presented at the Land and the City: Proceedings of the 2014 Land Policy Conference, Cambridge, MA.
Chudgar, A., & Luschei, T. F. (2009). National Income, Income Inequality, and the Importance of Schools: A Hierarchical Cross-National Comparison. *American Educational Research Journal, 46*(3), 626.
Ciccone, A., & Papaioannou, E. (2009). Human Capital, the Structure of Production, and Growth. *The Review of Economics and Statistics, 91*(1), 66-82.
Falch, T., & Fischer, J. A. V. (2012). Public sector decentralization and school performance: International evidence. *Economics Letters, 114*, 276-279.
Falck, O., Heimisch, A., & Wiederhold, S. (2016). Returns to ICT skills. CESifo Working Paper No. 5720.
Ginsburg, A., & Smith, M. S. (2016). Do Randomized Controlled Trials Meet the "Gold Standard"? *American Enterprise Institute.* Retrieved March, 18, 2016.
Gonzalez, E. J. (2012). *Rescaling sampling weights and selecting mini-samples from large-scale assessment databases*. Retrieved from IERI Monograph Series: Issues and Methodologies in Large-Scale Assessments, Volume 5.
Green, D. A., & Riddell, W. C. (2013). Ageing and literacy skills: Evidence from Canada, Norway and the United States. *Labour Economics, 22*, 16-29.

Gundlach, E., Rudman, D., & Wößmann, L. (2002). Second Thoughts on Development Accounting. *Applied Economics, 34*, 1359-1369.
Gundlach, E., & Wößmann, L. (2001). The Fading Productivity of schooling in East Asia. *Jounral of Asian Economics, 12*(3), 401-417.
Gundlach, E., Wößmann, L., & Gmelin, J. (2001). The Decline of Schooling Productivity in OECD Countries. *The Economic Journal, 111*, C135-C147.
Gustafsson, J.-E. (2013). Causal inference in educational effectiveness research: a comparison of three methods to investigate effects of homework on student achievement 1. *School Effectiveness and School Improvement, 24*(3), 275-295.
Gustafsson, J.-E. (2016). Lasting effects of quality of schooling: Evidence from PISA and PIAAC. *Intelligence, 57*, 66-72.
Hanushek, E. A., & Kimko, D. D. (2000). Schooling, labor-force quality, and the growth of nations. *American Economic Review*, 1184-1208.
Hanushek, E. A., Link, S., & Wößmann, L. (2013). Does school autonomy make sense everywhere? Panel estimates from PISA. *Journal of Development Economics, 104*, 212-232.
Hanushek, E. A., & Wößmann, L. (2006). Does Educational Tracking Affect Performance and Inequality? Differences-in-Differences Evidence Across Countries. *The Economic Journal, 116*(March), 63-76.
Hanushek, E. A., & Wößmann, L. (2008). The Role of Cognitive Skills in Economic Development. *Journal of Economic Literature, 46*(3), 607-668.
Hanushek, E. A., & Wößmann, L. (2011). The economics of international differences in educational achievement. In E. A. Hanushek, S. Machine, & L. Woessmann (Eds.), *Handbooks in Economics Vol. 3* (pp. 89-200). The Netherlands: North-Holland.
Hanushek, E. A., & Wößmann, L. (2012a). Do better schools lead to more growth? Cognitive skills, economic outcomes, and causation. *Journal of Economic Growth, 17*, 267-321.
Hanushek, E. A., & Wößmann, L. (2012b). Schooling, cognitive skills, and the Latin American growth puzzle. *Journal of Development Economics, 99*, 497-512.
Heyneman, S. P., & Loxley, W. A. (1983). The Effect of Primary-School Quality on Academic Achievement across Twenty-nine High- and Low-Income Countries. *American Journal of Sociology, 88*(6), 1162-1194.
Horowitz, J. L., & Manski, C. F. (2000). Nonparametric Analysis of Randomized Experiments with Missing Covariate and Outcome Data. *Journal of the American Statistical Association, 95*, 77-84.
Huang, F. L. (2014). Alternatives to Multilevel Modeling for the Analysis of Clustered Data. *The Journal of Experimental Education, 84*(1), 175-196.
Hutchison, D., & Schagen, I. (2007). Comparisons between PISA and TIMSS: Are we the man with two watches. *Lessons learned: What international assessments tell us about math achievement*, 227-261.
IEA. (2011). Brief History of the IEA: Fifty-five years of educational research. Retrieved from http://www.iea.nl/brief-history-iea.
Jaeger, M. M. (2009). Equal Access but Unequal Outcomes. *Social Forces*.
Jakubowski, M. (2010). Institutional Tracking and Achievement Growth: Exploring Difference-in-Differences Approach to PIRLS, TIMSS, and PISA Data. In J. Dronkers (Ed.), *Quality and Inequality of Education* (pp. 41-81). Dordrecht: Springer Science+Business Media B.V.

Jakubowski, M. (2013). Analysis of the predictive power of pisa test items.
Jakubowski, M., & Pokropek, A. (2015). Reading achievement progress across countries. *International Journal of Educational Development, 45*, 77-88.
Jamison, E. A., Jamison, D. T., & Hanushek, E. A. (2007). The Effects of Education Quality on Income Growth and Mortality Decline. *Economics of Education Review, 26*(6), 771-788.
Jerrim, J., Chmielewski, A. K., & Parker, P. (2015). Socioeconomic inequality in access to high-status colleges: A cross-country comparison. *Research in Social Stratification and Mobility, 42*, 20-32.
Jerrim, J., Parker, P. D., Chmielewski, A. K., & Anders, J. (2016). Private Schooling, Educational Transitions, and Early Labour Market Outcomes: Evidence from Three Anglophone Countries. *European Sociological Review, 32*(2), 280-294.
Jerrim, J., & Vignoles, A. (2015). University access for disadvantaged children: a comparison across countries. *Higher Education, 70*(6), 903-921.
Kahn, L. M. (2007). The impact of employment protection mandates on demographic temporary employment patterns: International microeconomic evidence. *The Economic Journal, 117*(521), F333-F356.
Kaplan, D. (2016). Causal inference with large-scale assessments in education from a Bayesian perspective: a review and synthesis. *Large-scale Assessments in Education, 4*(1), 1-24.
Kena, G., Hussar W., McFarland J., de Brey C., Musu-Gillette, L., Wang, X., Zhang, J., Rathbun, A., Wilkinson-Flicker, S., Diliberti M., Barmer, A., Bullock Mann, F., & Dunlop Velez, E. (2016). The Condition of Education 2016 (NCES 2016-144). U.S. Department of Education, National Center for Education Statistics. Washington, DC.
Kreuter, F., & Valliant, R. (2007). A survey on survey statistics: What is done and can be done in Stata. *The Stata Journal, 7*(1), 1-21.
Lee, E. S., Forthofer, R. N., & Lorimor, R. J. (1989). Analyzing Complex Survey Data, Volume 71, Quantitative Applications in the Social Sciences: Newbury Park: Sage Publications.
Lee, J.-W., & Barro, R. J. (2001). Schooling Quality in a Cross-Section of Countries. *Economica, 68*(271), 465-488.
Lee, D.S. (2002). Trimming for Bounds on Treatment Effects with Missing Outcomes. NBER Technical Working Paper 277.
Marks, G. N. (2005). Cross-National Differences and Accounting for Social Class Inequalities in Education. *International Sociology, 20*(4), 483-505.
Martin, M. O., & Mullis, I. V. S. (2013). *Methods and Procedures in TIMSS and PIRLS 2011*. Retrieved from Boston.
Meinck, S. (2015). *Computing Sampling Weights in Large-Scale Assessments in Education*. Retrieved from Survey Insights: Methods from the Field, Weighting: Practical Issues and 'How to' Approach.
Merry, J. J. (2013). Tracing the U.S. Deficit in PISA Reading Skills to Early Childhood Evidence from the United States and Canada. *Sociology of Education*, 0038040712472913.
Micklewright, J., & Schnepf, S. V. (2006). *Response bias in England in PISA 2000 and 2003, Department for Education and Skills Research Report 771*. Retrieved from www.education.gov.uk/publications/eOrderingDownload/RR771.pdf.
Mosteller. (1995). The Tennessee Study of Class Size in the Early School Grades. *The Future of Children: Financing Schools, 5*(Summer/Fall), 113-127.

Mullis, I. V. S., Martin, M. O., Ruddock, G. J., O'Sullivan, C. Y., & Preuschoff, C. (2009). *TIMSS 2011 Assessment Frameworks*. Retrieved from Boston: TIMSS & PIRLS International Study Center, Lynch School of Education, Boston College and IEA.

Murnane, R. J., & Willett, J. B. (2010). *Methods matter: Improving causal inference in educational and social science research*. New York: Oxford University Press.

OECD. (2012). *Learning beyond Fifteen: Ten Years after PISA*. OECD Publishing.

OECD. (2013). *PISA 2012 Assessment and Analytical Framework: Mathematics, Reading, Science, Problem Solving and Financial Literacy*. OECD Publishing.

OECD. (2014a). *PISA 2012 Technical Report*. OECD Publishing.

OECD. (2014b). Do countries with high mean performance in PISA maintain their lead as students age? *PISA in Focus, 45*.

Olson, J. F., Martin, M. O., & Mullis, I. V. S. (2008). *TIMSS 2007 Technical Report*. Retrieved from TIMSS & PIRLS International Study Center, Lynch School of Education, Boston College.

Paccagnella, M. (2016). *Literacy and Numeracy Proficiency in IALS, ALL and PIAAC*. OECD Education Working Papers No. 142.

Pfeffer, F. T. (2008). Persistent inequality in educational attainment and its institutional context. *European Sociological Review, 24*(5), 543.

Rabe-Hesketh, S., & Skrondal, A. (2006). Multilevel Modeling of Complex Survey Data. *Journal of the Royal Statistical Society, 169*(4), 805-827.

Ramirez, F. O., Luo, X., Schofer, E., & Meyer, J. W. (2006). Student Achievement and National Economic Growth. *American Journal of Education, 113*(1), 1-29.

Robinson, J. P. (2013). Causal Inference and Comparative Analysis with Large-Scale Assessment Data. In L. Rutkowski, M. von Davier, & D. Rutkowski (Eds.), *Handbook of international large-scale assessment: Background, technical issues, and methods of data analysis*. Boca Raton, FL: CRC Press.

Rosén, M., & Gustafsson, J.-E. (2016). Is computer availability at home causally related to reading achievement in grade 4? A longitudinal difference in differences approach to IEA data from 1991 to 2006. *Large-scale Assessments in Education, 4*(1), 1.

Rosenbaum, P. R., & Rubin, D. B. (1983). Assessing sensitivity to an unobserved binary covariate in an observational study with binary outcome. *Journal of the Royal Statistical Society. Series B (Methodological)*, 212-218.

Rubin, D. (1987). *Multiple imputation for nonresponse in sample surveys*. New York: John Wiley & Sons.

Ruhose, J., & Schwerdt, G. (2016). Does early educational tracking increase migrant-native achievement gaps? Differences-in-differences evidence across countries. *Economics of Education Review, 52*, 134-154.

Rutkowski, D., & Delandshere, G. (2016). Causal inferences with large scale assessment data: using a validity framework. *Large-scale Assessments in Education, 4*(1), 1.

Rutkowski, D., Rutkowski, L., & Plucker, J. A. (2012). Trends in education excellence gaps: A 12-year international perspective via the multilevel model for change. *High Ability Studies, 23*(2), 143-166.

Rutkowski, L., Gonzalez, E., Joncas, M., & von Davier, M. (2010). International large-scale assessment data issues in secondary analysis and reporting. *Educational Researcher, 39*(2), 142-151.

Salthouse, T. A. (2009). When Does Age-Related Cognitive Decline Begin? *Neurobiology of Ageing, 30*(4), 507-514.
Salthouse, T. A. (2010). Influence of Age on Practice Effects in Longitudinal Neurocognitive Change. *Neuropsychology, 24*(5), 563-572.
Schafer, J. L. (2003). Multiple imputation in multivariate problems when the imputation and analysis models differ. *Statistica Neerlandica, 57*, 19-35.
Schmidt, W. H., Burroughs, N. A., Zoido, P., & Houang, R. T. (2015). The Role of Schooling in Perpetuating Educational Inequality An International Perspective. *Educational Researcher, 44*(7), 371-386.
Schmidt, W. H., McKnight, C. C., Houang, R. T., Wang, H., Wiley, D. E., Cogan, L. S., & Wolfe, R. G. (2001). *Why Schools Matter: A Cross-National Comparison of Curriculum and Learning*. New York: John Wiley & Sons.
Schneider, S. (2008). *The International Standard Classification of Education (ISCED-97). An Evaluation of Content and Criterion Validity for 15 European Countries*. Retrieved from Mannheim: Mannheimer Zentrum für Europäische Sozialforschung (MZES).
Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Boston: Houghton Mifflin.
Skirbekk, V., Bordone, V., & Weber, D. (2014). A cross-country comparison of math achievement at teen age and cognitive performance 40 years later. *Demographic Research, 31*, 105.
Tarozzi, A. (2007). Calculating comparable statistics from incomparable surveys, with an application to poverty in India. *Journal of Business & Economics Statistics, 25*(3), 314-336.
The Ministry of Education. (2014). *Summary of the Danish PISA-PIAAC survey*. Copenhagen: Undervisnings Ministeriet.
Torney-Purta, J., Wilkenfeld, B., & Barber, C. (2008). How Adolescents in 27 Countries Understand, Support, and Practice Human Rights. *Journal of Social Issues, 64*(4), 857-880.
West, M. R., & Wößmann, L. (2010). 'Every Catholic Child in a Catholic School': Historical Resistance to State Schooling, Contemporary Private Competition and Student Achievement across Countries. *The Economic Journal, 120*(546), F229-F255.
Willms, J. D. (2010). School Composition and Contextual Effects on Student Outcomes. *The Teachers College Record, 112*(4), 3-4.
Wiseman, A. W., Baker, D. P., Riegle-Crumb, C., & Ramirez, F. O. (2009). Shifting Gender Effects: Opportunity Structures, Institutionalized Mass Schooling, and Cross-National Achievement in Mathematics. In D. P. Baker & A. W. Wiseman (Eds.), *Gender, Equality and Education from International and Comparative Perspectives* (Vol. 10, pp. 395-422). Bingley, U.K.: Emerald Group Publishing Limited.
Wößmann, L. (2005). Educational production in Europe. *Economic Policy, 20*(43), 446-504.
Wößmann, L., & West, M. R. (2006). Class-size effects in school systems around the world: Evidence from between-grade variation in TIMSS. *European Economic Review, 50*, 695-736.
Wu, M. (2009). A comparison of PISA and TIMSS 2003 achievement results in mathematics. *Prospects, 39*(1), 33-46.
Zimmer, R. W., & Toma, E. F. (2000). Peer effects in private and public schools across countries. *Journal of Policy Analysis and Management*, 75-92.