International Education Assessments

Cautions, Conundrums, and Common Sense

NATIONAL ACADEMY OF EDUCATION

International Education Assessments

Cautions, Conundrums, and Common Sense

Judith D. Singer, Henry I. Braun, and Naomi Chudowsky, *Editors*

National Academy of Education
Washington, DC

NATIONAL ACADEMY OF EDUCATION 500 Fifth Street, NW Washington, DC 20001

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 the views of the Institute or the U.S. Department of Education.

Additional copies of this publication are available from the National Academy of Education, 500 Fifth Street, NW, Washington, DC 20001; naeducation.org.

Digital Object Identifier: 10.31094/2018/1

Printed in the United States of America

Suggested Citation: Singer, J. D., Braun, H. I., & Chudowsky, N. (2018). *International education assessments: Cautions, conundrums, and common sense*. Washington, DC: National Academy of Education.

NATIONAL ACADEMY *of* EDUCATION

The **National Academy of Education** (NAEd) advances high-quality research to improve education policy and practice. Founded in 1965, the NAEd consists of U.S. members and foreign associates who are elected on the basis of outstanding scholarship related to education. The NAEd undertakes research studies to address pressing issues in education and administers professional development programs to enhance the preparation of the next generation of education scholars.

METHODS AND POLICY USES OF INTERNATIONAL LARGE-SCALE ASSESSMENTS STEERING COMMITTEE

Judith D. Singer (*Chair*), Harvard University
Henry I. Braun, Boston College
Anna Katyn Chmielewski, University of Toronto
Richard Durán, University of California, Santa Barbara
David Kaplan, University of Wisconsin–Madison
Marshall "Mike" Smith, Carnegie Foundation for Advancement of Teaching
Judith Torney-Purta, University of Maryland

Michael J. Feuer (*Principal Investigator*), The George Washington University

Preface

Results from international large-scale assessments (ILSAs) receive considerable attention from academics, policy makers, business leaders, and the media. Reported findings often raise questions—and in some instances alarm—about whether a nation's students are prepared to compete with their counterparts in a globalizing economy. Results also raise concern over how well students are being prepared for citizenship and other adult roles in society.

Although there is widespread recognition that ILSAs can provide useful information—and are invaluable for mobilizing political will to invest in education—there is little consensus among researchers about what types of comparisons are the most meaningful and what could be done to assure more sound interpretation. The central question is simple: What do the results of such assessments really tell us about the strengths and the weaknesses of a nation's education system? Unfortunately, and perhaps not surprisingly, the answer to this question is far more complex.

The challenges of drawing policy conclusions from ILSAs became even more apparent to me during my first trip to Singapore, in October 2017, which happened to coincide with the writing of this report. Singapore is one of the high-scoring East Asian countries, whose stellar performance is often suggested as a source of inspiration for policy makers in other countries seeking to improve the performance of their own students. But comparing Singapore to large countries with decentralized education systems like the United States is challenging. Not only is Singapore's education system completely centralized, there is only one School of Education, which prepares all the nation's teachers, and the city state is

actually so small that there are more school districts in a mid-sized state like Massachusetts than there are schools in Singapore.

This brings us back to the seemingly simple, but actually incredibly complex, question: What do these assessments really tell us? To address this question, the National Academy of Education (NAEd) undertook an initiative to examine future directions for ILSAs from a variety of disciplinary perspectives. The project was made possible through support by the U.S. Department of Education's Institute of Education Sciences' National Center for Education Statistics, and a steering committee was formed to plan and facilitate two workshops, commission papers, and oversee the writing of this summary report.

The first workshop, held on June 17, 2016, focused on methodological issues related to the design, analysis, and reporting of ILSAs. The second workshop, held on September 16, 2016, moved into the less-technical aspects of reporting, interpretation, and policy uses of ILSAs. Both workshops took place in Washington, DC, and all workshop materials, including agendas, videos, and commissioned papers, are available on the project website at naeducation.org. (See Appendix A for the workshop agendas and participants.)

The goal of the workshops was to highlight the strengths, limitations, and complexities of ILSAs, especially as a basis for informing educational policy and practice. The steering committee was not charged with reaching a consensus on a set of conclusions and recommendations. Rather, the purpose of the workshop series was to hear a variety of perspectives to advance our understanding of these issues. In addition, the committee decided to include people who often are missing from education discussions—that is, experts from outside the field of education to offer their perspectives on the value of cross-national comparative research. Collectively, the workshop presentations, commissioned papers, and discussions enabled the committee to write this report, which identifies general areas of agreement and disagreement, as well as what the committee believes are constructive suggestions for moving forward. Readers should view this report as a summary of the arguments presented, not as a consensus document.

There are many individuals whom I acknowledge and thank for their invaluable contributions to this project. First, I was appointed chair of the steering committee by then-president of the National Academy of Education, Michael Feuer, who was instrumental in developing this project as well as providing overall guidance and management as its principal investigator. Dr. Feuer was assisted in this task by NAEd staff, senior program officer Naomi Chudowsky, and executive director Gregory White.

I also thank my fellow editors of this report, Henry Braun and Naomi Chudowsky, who were true partners in this effort, from conceptualization

of the workshops to report writing. Our productive collaboration has been both intellectually stimulating and fun. This report would not have been successfully completed without the time, energy, and insights they contributed.

The success of this project also depended on the commitment and the participation of the steering committee members, who contributed substantial time and expertise in project planning, recruiting participants, making presentations, and developing the report. Steering committee members include Henry Braun, Anna Katyn Chmielewski, Richard Durán, David Kaplan, Marshall "Mike" Smith, and Judith Torney-Purta.

On behalf of the steering committee, I acknowledge and extend our sincere appreciation to the many individuals who authored papers and made presentations at our two workshops. The following list of contributors represents the broad range of experience in assessment research, policy, governmental service, and journalism that was brought to bear in support of this project: Norman Bradburn, Henry Braun, Kevin Carey, Peggy Carr, Anna Katyn Chmielewski, Elizabeth Dhuey, Richard Durán, Michael Feuer, Jan-Eric Gustafsson, John Haaga, Eric Hanushek, Jack Jennings, Nina Jude, David Kaplan, Daniel Koretz, Susanne Kuger, Nicholas Lemann, Hank Levin, Michele McLaughlin, Ina Mullis, Ellen Nolte, Sean Reardon, Leslie Rutkowski, Marshall "Mike" Smith, Judith Torney-Purta, Marc Tucker, Elizabeth Washbrook, and Brad Wible.

Peer review is an essential ingredient to ensuring the quality and the objectivity of reports produced by the NAEd. I thank Judith Warren Little, chair of the NAEd Standing Review Committee, for overseeing the review process for this report, and Jack Jennings and Sean Reardon, who provided a thoughtful review.

Finally, this project was conceptualized in collaboration with the National Academies of Sciences, Engineering, and Medicine's Division of Behavioral and Social Sciences and Education, with the intention that the NAEd and the National Academies will build on the success of these workshops with a continued program of work exploring these issues.

Judith D. Singer
Chair, Steering Committee

Contents

1

International Large-Scale Assessments in Education

International large-scale assessments (ILSAs) have been in existence in one form or another since the mid-1960s. Beginning with the advent of the First International Mathematics Study (FIMS), international assessments in education have since proliferated (see Table 1-1). Of those conducted, the most well-known ILSAs are the ongoing Trends in International Mathematics and Science Study (TIMSS), the Progress in International Reading Literacy Study (PIRLS), and the Programme for International Student Assessment (PISA).

PURPOSES OF INTERNATIONAL LARGE-SCALE ASSESSMENTS

Workshop committee member Henry Braun of Boston College urged participants to "think about the utility of ILSAs," noting that, "presumably there is utility because we keep doing them and we keep spending large amounts of money on them." Over the course of the workshop series, participants generated several purposes of ILSAs, including

1. To *describe* and *compare* student achievement and contextual factors (e.g., policies, student characteristics) across nations.
2. To *track changes over time* in student achievement, contextual factors, and their mutual relationships, within and across nations.
3. To disturb complacency about a nation's education system and *spur educational reforms*.

TABLE 1-1 Major ILSAs in Education

When		Study and Sponsor			Target Population
Year(s)	Planned/ Future	Assessment	Acronym	Sponsor	# Countries
Ongoing Studies					
1995, 1999, 2003, 2007, 2011, 2015	2019	Trends in International Mathematics and Science Study	TIMSS	IEA	29-50
2001, 2006, 2011, 2016	2021	Progress in International Reading Literacy Study	PIRLS	IEA	36-54
2000, 2003, 2006, 2009, 2012, 2015	2018	Programme for International Student Assessment	PISA	OECD	43-75
2012, 2014, 2017	2021	Programme for the International Assessment of Adult Competencies	PIAAC	OECD	40+
Past Studies					
1964		First International Mathematics Study	FIMS	IEA	11
1970		First International Science Study	FISS	IEA	20
1980		Second International Mathematics Study	SIMS	IEA	21

	Domains Assessed						
Ages/ Grade Levels	Math	Numeracy	Quantitative Reasoning	Problem Solving	Science	Reading	Literacy
Grades 4, 8, and 12	✓	✓			✓		
Grade 4						✓	✓
15	✓				✓	✓	
16-65		✓		✓			
13 and final year of secondary school	✓						
10, 14, and final year of secondary school					✓		
13	✓						

continued

TABLE 1-1 Continued

When		Study and Sponsor			Target Population
Year(s)	Planned/ Future	Assessment	Acronym	Sponsor	# Countries
1984		Second International Science Study	SISS	IEA	24
1994		International Adult Literacy Survey	IALS	OECD	20
2003, 2006		Adult Literacy and Lifeskills Survey	ALL	Statistics Canada	4-7

NOTE: IEA = International Association for the Evaluation of Educational Achievement; OECD = Organisation for Economic Co-operation and Development.

4. To *create de facto international benchmarking* by identifying top-performing nations and jurisdictions, or those making unusually large gains, and learning from their practices.
5. To *evaluate* the effectiveness of curricula, instructional strategies, and educational policies.
6. To *explore causal relationships* between contextual factors (e.g., demographic, social, economic, and educational variables) and student achievement.

While many of these purposes may seem relatively straightforward, a great deal of workshop discussion centered around the extent to which ISLAs, as currently designed and administered, can fulfill all of them. If not, how would ILSAs need to change to do so?

COMPARISONS AMONG NATIONS IN EDUCATION

ILSA results are most often presented as a ranking of nations—that is, which countries are at the top in terms of student achievement (as measured by a particular test) and which are at the bottom. This type of ranking would appear to be useful because a nation, for example, can see it is not performing well and needs to improve, and subsequent arguments can be made that more resources should be devoted to education. Yet, Americans often see news stories with a lead paragraph such as this

	Domains Assessed						
Ages/ Grade Levels	Math	Numeracy	Quantitative Reasoning	Problem Solving	Science	Reading	Literacy
10, 14, and final year of secondary school					✓		
16-65			✓				✓
16-65		✓					✓

one from the National Public Radio, under the headline "U.S. Students Slide in Global Ranking on Math, Reading, Science," which stated,

> American 15-year-olds continue to turn in flat results in a test that measures students' proficiency in reading, math, and science worldwide, failing to crack the global top 20. (Chappell, 2013)

The use of the word "slide" in the headline gives the mistaken impression that U.S. scores are declining, although the article goes on to explain that the average scores remain relatively flat. Rather, what is changing is that an increasing number of nations are surpassing the United States in the rankings. Articles like this one certainly capture readers' attention, for better or for worse.

Hanushek and Wößmann (2011) further argue that such rankings "have spurred not only governmental attention but also immense public interest, as is vividly documented by the regular, vigorous news coverage and public debate of the outcomes of the international achievement tests in many of the participating countries" (p. 91). From their perspective, regular reporting of national rankings keeps shortcomings of national performance on the front pages and helps to prevent "we're number one" jingoism.

But not everyone shares their view. Workshop participant Daniel Koretz of Harvard University was far more pessimistic about the promise of international assessments. Because these assessments collect data from

a range of countries that differ in so many ways, he noted, they "have limitations that are simply unavoidable. Some of [the limitations] having to do with enough money or time could be reduced or eliminated, but right now they're not avoidable." From this perspective, ILSAs have shortcomings that need to be recognized before people attach too much importance to international rankings. As discussed in later chapters of this report, careful analyses of ILSA data that go beyond simple rankings are needed to provide important nuance and context.

Of the aforementioned purposes of ILSAs, workshop participants expressed the most concern about the last three: create benchmarking; evaluate effectiveness; and explore causal relationships. Participants did agree that these are worthy goals, but many argued that ILSAs, as currently designed and administered, do not provide the information needed to draw these kinds of conclusions. At the extreme, for example, consider the last purpose—explore causal relationships—which asks researchers to use ILSAs to determine why some nations perform better than others, that is, which polices and practices produce better educational outcomes. Education leaders and researchers have an obvious desire to look at policies and practices in other high-performing nations to see what might be adopted or adapted in their own. According to Chmielewski and Dhuey (2017),

> ILSAs have been used not only to compare performance, curricula, instructional, and learning strategies across countries but also to try to understand how international differences in education policies—the structure, administration, legal, economic, and political contexts of national and subnational school systems—shape student achievement and other outcomes. (p. 1)

The hope is that such comparisons will help establish connections between specific policies and practices and educational achievement. The question of whether causation can be inferred from analyses of ILSA data was a recurring theme of this workshop series, as was the question of what steps, if any, could increase the likelihood of appropriate causal inference. Given the large number of factors affecting student achievement, and the fact that nations differ from one another in terms of demographics, wealth, and beliefs about the value of education, workshop participants disagreed about the extent to which it is currently possible—or would ever be possible—to isolate those specific factors or policies that contribute to improved student achievement in one nation that could be applied elsewhere. Some workshop participants argued that causality cannot be firmly established without randomized controlled experiments or rigorous quasi-experiments, which are not realistic in educational assessment at this scale. Others argued that researchers can

come "reasonably close" to identifying causal relationships using careful and creative research designs and analyses. These issues are explored more fully in Chapters 4 and 5.

INTERNATIONAL COMPARISONS IN OTHER FIELDS

One goal of this workshop series was to learn from international comparative research in fields other than education, since student achievement is but one of many characteristics that can be compared across nations. Indeed, a variety of international organizations compare nations on a wide array of indicators such as economic trends, trade, health, and crime.

The committee invited Ellen Nolte of the London School of Economics, who conducts research on comparative health care delivery for the European Observatory on Health Systems and Policies. Similar to the six purposes of ILSAs outlined above, Nolte presented a more compact set of reasons for conducting cross-national comparisons of health systems:

- **Learning about national systems and policies:** The intent here is to conduct descriptive studies that explore similarities and differences among nations. Descriptive studies may lay the groundwork for more analytical types of studies.
- **Learning why systems and policies have taken the form they do:** The purpose is to identify the factors that have contributed to the policies or the practices taking a particular form. Typically, one would do such a study to generate or test hypotheses, develop typologies, track policy trends over time, or to explain the past. These may be of more limited interest to policy makers.
- **Learning lessons from other countries for application in one's own country:** The intent is to understand a given political event or process by comparing it with similar events or processes elsewhere. First, the focus is on how the particular policy challenge plays out in other countries; and second, on an attempt to identify best practices and the potential for transfer.

The three purposes that Nolte laid out are similar to some of the six purposes for educational comparisons presented at the beginning of this chapter. Nolte, however, is more circumspect about what can be learned from this type of research. This raises the issue of whether the media, researchers, and policy makers are overly ambitious with regard to the strength of the conclusions that can be drawn from ILSAs given the heterogeneity in countries' circumstances and educational systems, as discussed in later chapters of this report.

Nolte argues that cross-national studies are a distinct research method, alongside experimental, observational, and case-study methods, and that researchers who analyze cross-national data face a distinctive set of challenges because of their focus on large macro-social units such as countries, societies, or civilizations. These social units are entities of considerable complexity and there is wide variability within and among them. Whether the focus is health care or education, national social policies are influenced by economic, political, demographic, technological, and cultural environments.

Despite challenges, Nolte argues that cross-national studies are still of great importance because they provide policy makers with options that they might not otherwise have considered. They allow for mutual learning across borders, cross-fertilization, or even transfer of models and ideas. National policy makers will occasionally use such studies to confirm the positive—for example, that what the country is actually doing is fine—or to see what has failed in other nations and refrain from adopting similar policies.

PRODUCERS AND CONSUMERS

Judith Singer of Harvard University (and steering committee chair) discussed the ecosphere of ILSAs—namely, the producers and the consumers—with the goal of shedding light on the range of professionals who support the continuing conduct of ILSAs.

First are the **producers,** the professionals who design, construct, and validate the assessments and who promulgate the testing and the sampling requirements; these include staff members at international organizations such as the International Association for the Evaluation of Educational Assessment (IEA), headquartered in Amsterdam and in Hamburg, or the Organisation for Economic Co-operation and Development (OECD), headquartered in Paris. These organizations rely on international committees of researchers who design assessment frameworks and background questionnaires, test-item specifications, and analysis plans. There are also numerous contractors who work on the assessments, many of whom are located in the United States (e.g., the Educational Testing Service and the International Study Center at Boston College). The organization that oversees the conduct of these assessments on behalf of the United States is the U.S. Department of Education, which regularly seeks advice from researchers about the methodology and the content of a given assessment.

On the **consumer** side, there is a broad array of professional groups:

- **Policy makers:** These include everyone from officials in the federal and the state governments, including members of the U.S.

Congress and state legislatures and their staff, to leaders of local school districts. They often see ILSA results as a way to both obtain an external perspective on educational performance and to spur educational reform.

- **Education advocacy organizations:** These include organizations such as the Knowledge Alliance or the Center on Education Policy, which are both concerned with the quality of U.S. education and use test score data to inform and bolster their advocacy efforts. Workshop participant Michele McLaughlin, president of the Knowledge Alliance, explained that her organization and others use ILSA results to pressure the U.S. Congress to make greater use of education research to inform policy.
- **The media:** When reporting to the public, the press in the United States and other countries trumpets headlines about the relative standing of its education system, especially when the news is negative. Unfortunately, not enough attention is paid to nuance—that is, the inherent heterogeneity of educational systems—especially in the United States, or to the shortcomings of the ILSAs themselves, let alone other complex research matters. Moreover, as workshop participant Nicholas Lemann, former dean of the Columbia School of Journalism, explained, the media is currently in the throes of massive changes, which have resulted in a substantial reduction in the time and the attention devoted to education coverage in general, as discussed further in Chapter 2.
- **Education researchers:** These are primarily individuals in academia; however, the category also includes people in governmental agencies, research organizations, nongovernmental organizations, and international organizations including but not limited to those who develop ILSAs. Generally, researchers are interested in using data to address questions of interest to them or to sponsors (referred to as "secondary analysis"); this is encouraged by organizations, especially IEA, which holds an international conference every 2 years, during which results of secondary analyses are presented. One issue with ILSAs, noted Judith Singer, is that in empirical research, ideally "you start with the questions and then you get the data." To some extent, that order appears to be reversed with ILSAs, in that "you start with the data and then you go fishing for questions." To be fair, the ILSA design process does begin with broad core conceptual frameworks (at least starting with the Programme for International Student Assessment [PISA] 2009). The core conceptual frameworks delineate the types of questions that the assessments are designed to address, but researchers often want to delve into areas that the data may not

necessarily support. As discussed at length in Chapters 4 and 5, there is often a mismatch between what is collected and what academic researchers want.

In short, there is great heterogeneity among ILSA consumers with regard to what they expect or want from these assessments. In addition to "speaking different languages," the various groups have differing needs, varying levels of technical expertise, and varying amounts of time to devote to understanding how ILSAs work and how the results can be interpreted. Part of the challenge for ILSA producers is to understand how, and what types of, information can be presented to policy makers and the media. One of the main themes of the workshop series was how to help the media do a better job of reporting results, and part of the answer to that question may be that education researchers and ILSA producers must do a better job of presenting and interpreting information for the public and the press.

ORGANIZATION OF THIS REPORT

This report covers the proceedings of two workshops held in Washington, DC, in June 2016 and September 2016. It also includes papers commissioned for this project by the National Academy of Education. One of the workshops focused on policy issues, the other on design issues. To weave these topics into a coherent narrative while remaining true to what was presented at the workshops, the main content of this report is organized into four chapters, each covering a different aspect of ILSAs and their uses.

Chapter 2, on reporting and interpretation, discusses the way the media presents ILSA results, and why this reporting is often shallow, without the background information need to help readers correctly interpret the results. The decline in education coverage in the news media is discussed as well. Prominent researchers also weigh in on the complexities and shortcomings of ILSA results that tend not to be mentioned in media reports.

Chapter 3, on policy uses of ILSAs, addresses how ILSAs have affected the educational policy landscape in the United States. Advocacy organizations use ILSA results in their lobbying efforts and as a tool for reform. To see how other fields use cross-national studies to support policy making, the committee drew from examples of research on aging and child development that also use international comparative perspectives.

Chapter 4, on design issues, traces the purposes of ILSAs from the 1960s forward and discusses how they were designed to serve those purposes. Education researchers and policy makers continuously demand

more from ILSA data, however, and the chapter discusses current limitations of ILSAs with respect to these ends, specifically cross-cultural comparability, data quality, measurement error, and the credibility of causal inferences. The implications of computer-based testing for ILSAs are also discussed.

Chapter 5, on statistical analysis, discusses the ways in which some education researchers have used creative strategies to analyze ILSA data, hoping to mitigate or even circumvent their limitations. In particular, the debate concerning the use of existing data to make causal inferences is summarized, and shortcomings of some research methodologies are discussed, as is the desire to move toward some sort of longitudinal data-collection system (i.e., one that would follow the same students as they progress through school).

Chapter 6 wraps up the report with a synthesis of areas of agreement and disagreement over ILSA methods and uses, along with suggestions for moving forward.

2

Interpretation and Reporting

International assessment results are definitely headline generators. When the Programme for International Assessment (PISA) results were released in late 2016, all the major U.S. news media outlets reported on the results. Americans are obviously interested in how the United States ranks in comparison to other nations, and, at first blush, the news does not appear to be good. In addition, there is often great attention devoted to the highest-ranking countries, such as Singapore and Finland. In the early 2000s, the latter nation was trending in the news because it was a top scorer on PISA. As a result, Finland was referred to in *The Atlantic Monthly* as "an education superpower," and the United Kingdom's *Guardian* liked the fact that Finnish children play a lot and do not start school until the age of seven (Butler, 2016; Partanen, 2011).

Should the United States emulate Finland? Several scholars have taken issue with these kinds of studies and news articles, arguing that cross-sectional comparisons of the results of average scores on a single test cannot tell the whole story. There are numerous differences in educational policies and practices that often go unmentioned in reporting international large-scale assessment (ILSA) scores. Furthermore, nations differ in terms of demographics, wealth, culture, beliefs about the value of education, and the status of teaching as a profession, among numerous other factors. Understanding these contextual factors in other nations should, of course, influence how their scores are interpreted. There is also the issue that different ILSAs test different domains and any single test necessarily provides an incomplete picture of student achievement. As a

cautionary tale, note that Finland's rankings started slipping in 2012 and have continued to decline (Heim, 2016). In this chapter, we review issues related to how ILSA scores are—or should be—interpreted, as well as background on media reporting of ILSAs.

At the workshop, Judith Singer of Harvard University pointed out a headline from *The New York Times*: "Top Test Scores from Shanghai Stun Educators" (Dillon, 2010). The city of Shanghai was ranked at the top of PISA scores for 15-year-olds in science, reading, and math, with the United States ranked far below. The headline, however, was misleading. The sample of students tested in Shanghai was not representative, even of the relevant age-group population of that city. When the PISA data were collected, China had an internal passport system whereby a person from the countryside could not move to the city of Shanghai and gain access to social services, including education. Thus, there were many migrant teens living in Shanghai who did not qualify to go to school and were not tested (Loveless, 2013). (Of note, this system is currently under reform in China.)

Among ILSA experts, there is a general sense that media reporting of ILSA results has a somewhat superficial character and, in many cases, may be misleading. Thus, a sizable proportion of one workshop was devoted to ways of improving the reporting of ILSA results. A panel was held on this topic that provided a rather sobering view of what can be expected of education journalism.

THE STATE OF THE NEWS MEDIA

Workshop participant Nicholas Lemann, a well-known writer on education-related topics and a former dean of the journalism school at Columbia University, described the news media's declining capacity to devote time and space to education stories. Most notably:

- **Declining newspaper advertising revenue:** Since reaching a high of about $65 billion in 2000, revenue that newspapers receive for advertising plummeted to just less than $20 billion in 2012. The decline was never recovered by an increase in advertising for the Web versions of newspapers, as revenues from that source have been rather stagnant and less than $5 billion annually. Instead, advertising revenue transitioned from print media to digital media (e.g., Google and Facebook).
- **Declining newsroom workforce:** In 1990, there were more than 55,000 newsroom staff across the country. By 2015, there were about 33,000, and Lemann argues that this is directly attributable to the revenue decline. "I can't think of a white-collar industry sector that's declining that rapidly," said Lemann, noting with

> some irony that soon there will likely be more members of the American Educational Research Association (AERA) than reporters in the United States.

Why does this matter insofar as ILSA reporting is concerned? In the past, newspapers "used to do high social value/low demonstrable economic value" parts of journalism, such as investigative reporting, international reporting, and of course education reporting. However, the decline in advertising revenue caused the "disappearance of the traditional newspaper education desk," even at flagship papers such as *The New York Times*.

To address this problem, education reporting is increasingly being outsourced to, or is being replaced by, bloggers, freelancers, and people who work on education issues at think-tanks. Lemann estimates that about 20 percent of the content in *The New York Times* is provided by the nonprofit sector. Often, these reports are offered to major media outlets for free. While it may be the case that more in-depth reporting is the result, the problem is that often this reporting is supported by foundations and nongovernmental organizations. Many of these groups have political or policy agendas and may be more likely to influence the content of the reporting. This is a change from traditional newsroom economics, when the local Chevrolet dealer, who purchased advertisements that funded the local newspaper, cared little about what reporters wrote. "So our idea that corporations are bad and foundations are good doesn't really apply in this area," said Lemann.

On the issue of giving the public a fuller and more nuanced view of ILSA results, workshop participant Brad Wible, editor of *Science* magazine, stated, "Generally it's no mystery that communication of science and public understanding of science [leaves] a lot to be desired." In general—not just when it comes to ILSAs—scientists and policy makers have a hard time communicating with one another. With the media, the problem is the tendency for headline writers to exaggerate the findings of scientific studies, including but hardly limited to ILSAs. *Science* magazine covers scientific topics for a broadly educated academic audience, and, as a result, its reporting on ILSA results goes beyond the usual coverage of how the United States ranks. In fact, *Science* has covered education topics such as how the United States' performance on ILSAs relates to science standards, teacher recruitment and retention (focusing on Finland and Europe, in general), treatment of women and gender gaps in performance on science tests, computerized assessment, and the association between ILSA performance and gross domestic product. He stated that it was difficult for much of the mainstream media to "look under the hood" of the test numbers in the same way *Science* does.

Because of time constraints, the committee did not delve into suggesting specific types of messaged or visual displays that would improve public communication of ILSA results. However, this is certainly a topic for further consideration. Interested readers are referred to a recent report on science communication by the National Academies of Sciences, Engineering, and Medicine (2017), as well as another by Singer and Braun (2018).

ILSA RESULTS AS A CATALYST FOR CHANGE

Kevin Carey introduced himself as precisely the type of new journalist that Lemann described. He works at the New America Foundation but is also a freelancer on education issues, including for *The New York Times.*

Carey did note that there is great interest in stories about U.S. rankings in education, and it has long been this way. A pattern has emerged whereby whatever nation presents a geopolitical challenge to the United States at the time is also the country that is "beating our socks off in the classroom," as illustrated by the attention focused on U.S. math and science education after the U.S.S.R. launched the Sputnik satellite. This continued with Japan in the 1990s, and now China, with its (seemingly) stellar Shanghai results. These "Sputnik moments" can be valuable to the education community because they create a kind of external shock to the national dialogue that has the effect of elevating education above other national priorities.

Carey recalled how in 2014, he himself created a small Sputnik moment for American higher education using data from the 2011 Program for the International Assessment of Adult Competencies (PIAAC). He noted how many in the education community console themselves about the state of U.S. K-12 education by pointing out that the U.S. higher education system is still the world's finest. Unfortunately, that is not entirely the case. It is true that based on global rankings of *top* universities, most of the best universities in the world are in the United States. But it is not the case that our universities are the best in terms of *average* student literacy and numeracy skills. "American college graduates are about middle in literacy and below average in math, in other words exactly the same broadly speaking as when we look at 15-year-olds, which shouldn't surprise us because they're the same people, just [a few years] later." He noted that *The New York Times* article he wrote about this topic (Carey, 2014) was among his most popular pieces.

WHAT SHOULD BE DONE?

There was some discussion at the workshop about how those who conduct and interpret ILSA results could better interact with the media.

One point of agreement was that it is impossible to expect strong reporting from the fast-paced cable news media, which runs short story cycles. Suggestions seemed to gravitate toward deeper, advanced training and better preparation for specific news releases. Longer-term training of reporters in advance of the release of results would be aimed at making reporters more social sciences-literate. The work of the Education Writers Association (EWA) was mentioned, because at several of its meetings, issues of international test reporting have been raised. Lemann encouraged joint efforts with EWA. He also suggested that efforts can start with a small number of journalists, perhaps a few hundred, who produce most education stories. They can be trained to look deeper into the data. Because of the "echo chamber" of the Web, their stories will be shared with a larger audience. He further suggested that foundations and other research funders should require that grantees be asked to take steps to increase "research literacy" about comparative and international studies among journalists as well as individuals involved in policy.

More specifically, to help reporters and others interpret ILSA results as they are released, workshop participant Norman Bradburn of The University of Chicago proposed creating an impartial, national board charged with providing guidance on ILSA design, analysis, reporting, and interpretation. Such a board could provide useful information to the Institute of Education Sciences' (IES's) National Center for Education Statistics (NCES). Bradburn recounted a board at the National Research Council on international comparative studies in education (namely, BICSE) that he chaired almost 30 years ago (the board was disbanded in the early 2000s). Marshall Smith of the Carnegie Foundation for the Advancement of Teaching and a workshop steering committee member also chaired BICSE for several years and described it as a valuable effort. It was supported primarily by NCES, the National Science Foundation (NSF), and the U.S. Department of Defense. BICSE gave advice about planning and reporting on ILSAs. The Board evaluated the quality of proposed studies and advocated to funders for those studies that met its criteria. Bradburn believes that it is time to form a similar impartial body to deal with recurring issues of ILSA design, analysis, interpretation, and reporting.

On a similar note, workshop participant Jack Jennings, retired president and CEO of the Center on Education Policy, proposed that an independent, respected group prepare and release a national report each year, which summarizes what can be concluded about American education from recently released ILSA reports and analyses. First, such a report would explain the differences among the various ILSAs and interpret the results as a whole. This would address the problem of the press reporting "one day that American kids succeed, another day American kids fail." Jennings proposed that a second part of the report would explain what

ILSAs can and cannot tell us about how factors in society and in schools affect student achievement.

PERSPECTIVES FROM RESEARCHERS

Workshop participant Sean Reardon of Stanford University offered his thoughts on how education researchers might do their part to improve media reporting of ILSA results, or at least offer caveats and background on ILSAs that would encourage people to think more deeply about the results. Below are some examples of caveats that Reardon believes could be communicated to the public when ILSA results are reported, but tend to go unmentioned.

Small *N* Problem

Only 50 or 60 countries routinely administer ILSAs. That is a small number, and from a statistical point of view, it is difficult to discern patterns from that small of a group. Rankings are also influenced by variations across assessments and by which countries choose to participate in a particular ILSA administration. People can look at the list of 50 countries and their average test scores and find support for whatever education theory they prefer. "There are far more hypotheses and far more folk theories about how education systems work than there are countries in the world," noted Reardon. People tend to have incomplete information about other countries, which leads to a lot of "generalizing, hand waving, and anecdotal information" being used as evidence.

Unit of Analysis

Reardon also noted that most countries can be divided into smaller subunits, whether those subunits are locally responsible for education (as they are in the United States) or not. For example, the Shanghai test results mentioned at the beginning of this chapter describe a single city (if they even do that!), but they certainly do not describe all of China. To overcome the small *N* problem and to get a richer picture of what is going on in various countries, Reardon argued against viewing nations as monolithic entities and was in favor of breaking down the data into smaller units of analysis. The United States is the extreme case with 50 states, 14,000 school districts, and more than 100,000 schools. Where possible, scholars should go into detailed within-country studies of characteristics of national and subnational units of education systems, rather than make cross-national comparisons.

Do Tests Measure the Quality of Education?

Another issue, which Reardon calls a "huge problem," is that when people see a test score, the first assumption is that the test score is the measure of the quality of the school or education system. Yet, it has been well established that student achievement is to a great extent a product of out-of-school factors. "Test scores are the product of the full set of experiences and opportunities a kid had to learn in [his or her] entire life, some of which happens in school, but a lot of which happens outside of school, in the home, after school, in preschool." Therefore, we may be misattributing differences in test scores solely to differences in school quality when, in reality, they are in large measure the product of differences in other societal factors. Reardon does not believe that it is obviously the case that ILSA scores are mainly reflective of the quality of education systems.

Student-Level Changes Over Time

If the purpose of ILSAs is to make comparisons among education systems, then a crucial missing piece is illustrating where students start out in terms of achievement when they enter the school system, and how much they improve as they progress through the grades. Currently, ILSAs do not enable that type of analysis because the tests administered at different grades are not aligned (i.e., vertically scaled) in such a way as to enable measurement of growth in achievement as individual students progress through school.

To illustrate this point, Reardon used an example from his domestic research to show why measuring change over the grades is important. He presented Figure 2-1, which shows that Chicago student scores grow more than those in Baltimore. Reardon and colleagues (2016) were able to examine this growth comparatively by mapping both jurisdictions' state test scores onto the National Assessment of Educational Progress (NAEP), which is vertically scaled so that growth can be measured from grade to grade. Chicago students started out in third grade in 2001 about one grade level below Baltimore students, yet by eighth grade in 2013, Chicago students were outperforming Baltimore students by nearly two grade levels, and were close to the national average. Reardon noted,

> Chicago is making rapid gains in those school years. Kids aren't getting great opportunities before third grade, maybe in early childhood, maybe in preschool, maybe in early elementary school. But something is happening in third through eighth grade that is making them catch up. This is not an artifact of our data. You see this in the NAEP data, as well. Chicago makes big increases from fourth to eighth grade.

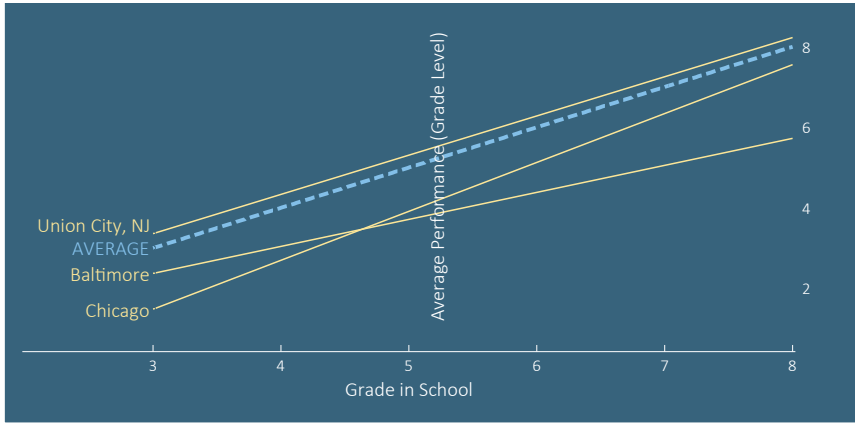


FIGURE 2-1 Average academic performance growth, grades 3 through 8, by school district.
SOURCE: Sean F. Reardon, Stanford University.

By linking assessments, researchers can more directly investigate the differences between Chicago and Baltimore to identify what may account for differential growth. If we apply this example to thinking about comparing performance internationally, ILSAs as currently administered cannot help us account for such differences. They are very thin snapshots that do not capture, for example, relative rates of performance growth that should be key to understanding the strengths and the weaknesses of national education systems.

Socioeconomic Status Matters

Reardon presented information on U.S. average test scores by school district, mapped onto the NAEP scale (Reardon, Kalogrides, & Ho, 2016; see Figure 2-2).

The map bolsters Reardon's point that although any country—including the United States—obviously does have a mean assessment score, the *variation* in student achievement across the thousands of U.S. school districts is enormous. The difference between the lowest and the highest district's mean test scores, expressed in grade equivalents, is estimated to be five grade levels. Not surprisingly, his research team found a high correlation between socioeconomic status (e.g., income levels, levels of parent education) and these district-level scores. When the United States is treated as a monolith and nations are ranked on the basis of a

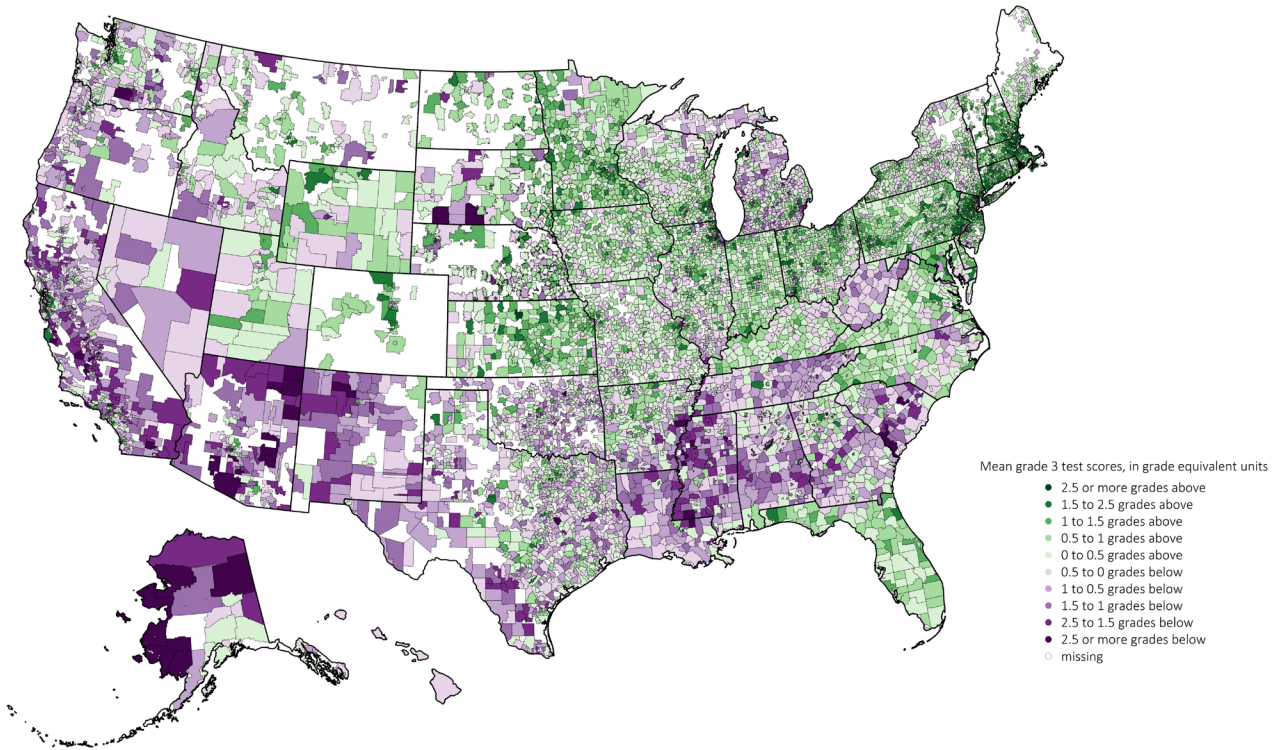


FIGURE 2-2 Average third grade test scores by U.S. public school districts, 2009–2015.
SOURCE: Sean F. Reardon, Stanford University: cepa.stanford.edu/seda/maps.

single national average, these differences are masked. Such differences in other nations are similarly hidden. If Finland, Singapore, or South Korea were states, they would rank at about 1.5 years above the U.S. national average, medium green on the map (see Figure 2-2). However, that still means there are thousands of school districts across the United States where test scores are higher than the average of those three countries. Thus, there are very high-performing school districts in the United States; though, as Reardon noted, they tend to be high supplemental educational services districts, but that is not always the case. For example, there are districts, in Massachusetts and Kentucky in particular, that perform far better than their characteristics would predict. They are relatively poorer districts that perform like wealthier ones. These places should be of great interest to researchers. They are "where the action is," Reardon said.

Yet, ILSA results, as currently administered and reported, tell us very little because we only have average test scores for an entire country.[1] We lose the variation within a country as states and regions have varying social and economic characteristics. Country-level ILSA results cannot point us in the right direction in terms of why achievement is improving in some parts of the country and not in others, or how students at different socioeconomic levels are performing. This point is also made by Carnoy and Rothstein (2013).

Additional Complexities

Norman Bradburn further explained some of the difficulties in interpreting ILSA results. When making judgments about what can be learned from ILSA results or policies, there are four main sources of what he referred to as "analytic complexity":

- **Variation among countries:** There are differences among nations in the way that educational systems are organized. There is social heterogeneity stemming from ethnic and language differences, differences in curricula, and differences in policies, such as student tracking, that affect the entire system.
- **School-level variation:** There are variations that come from the particular school a student attends. These include differences in levels of teacher and student autonomy, teacher qualifications and

[1] Committee member Anna Katyn Chmielewski noted that it is possible to measure variance within countries in all ILSAs (at the student and the school level in PISA; at the student, the school, and the classroom level in IEA studies). But the United States does not collect state-level data in any ILSA (unlike other countries, such as Canada, Germany, and Mexico).

experience, school safety, whether a school has entrance requirements, or whether it is public, private, or religiously affiliated.
- **Classroom variation:** This can include class size and time spent on particular subjects, as well as the qualifications, experience, and behavior of teachers who are actually teaching the students taking the test.
- **Student-level variation:** This includes such factors as family background, individual interest in the subject matter, motivation, and health, among others.

Bradburn asserted, "All of these sources of variation need to be accounted for in any analysis that attempts to explain differences in assessment outcomes." Ideally, studies involving international comparison must attempt to get data from each of these levels of analysis; currently these data tend to come from background questionnaires completed by the student, class and school questionnaires completed by teachers and principals, and administrative data collected by the organizations that administer the tests. Yet, as we explore in Chapter 4, different ILSAs measure different background variables, or the same variables in different ways. According to Bradburn, "The result is a large array of studies that have little cohesion. We have lots of studies of trees, but little understanding of the forest."

Building on Reardon's presentation, Bradburn made the point that because of the wide variability among and within nations, it is difficult to isolate the critical factors in other nations in order to inform social and educational policy in the U.S. context, as an analytical study might try to do. For this reason, ILSAs may be useful for descriptive purposes, but they are less useful for policy-relevant analytical purposes. Bradburn reiterated the view that it might be more useful to do comparative studies within the United States with its 50 states and 14,000 school districts. "Explanatory studies based on state differences or district differences would be more likely to turn up policy ideas that can be more fully understood in the American context and are more likely to be palatable to citizens and politicians in the several states," he explained.

Workshop participant Henry Levin of the Teachers College at Columbia University has been involved with PISA since 2006. He raised several issues related to the complexities of interpreting test results because of different practices in different nations. Specifically:

- **Temptation of causal inference:** Levin feels that even PISA itself, when issuing its reports, slips into the language of causal inference, especially when scores of higher-performing nations are discussed. Levin stated, "these studies can only be correlational.

They are not causal studies. They are not longitudinal [at the student level]. We need to keep that in mind because everyone here knows that. But when you go to the PISA results and the league tables, they are interpreted as being causal inferences drawn from certain countries." (This point is further discussed in Chapter 5.)

- **Use of test results to make judgments about the quality of schools:** ILSA test results are measurements of student performance in certain limited domains. But Levin believes that there are other educational goals schools pursue that are not covered by ILSAs. Social and emotional attributes of students are getting more attention in the United States, such as interpersonal skills, behavior, and all that the education system does to produce responsible citizens and productive members of society.
- **Variance in the population tested:** There are differences among nations in the pool of students tested. As mentioned at the beginning of this chapter, PISA results for the city of Shanghai excluded children whose families did not have residency permits, basically low-income children coming from rural areas. "But the results had already gone out all over the world about the Shanghai system as if the sampling had been done in the way that would normally be expected. There needs to be concern about that."
- **Shadow education:** Private tutoring is popular in many countries in Asia, as well as in Africa and in Latin America. South Korea is the best-known example, with private tutoring focused on maximizing test scores because these scores are the sole criterion for college admission. The results are high test scores, but what does that say about the quality of South Korea's public education system? The nation allocates approximately 4.5 percent of its gross domestic product to government schools. But private tutoring accounts for the equivalent of an additional 3 percent of gross domestic product—that is another 75 percent added to what the government spends. South Korea, therefore, does well on testing, but one interpretation may be that, in fact, South Korean schools are not very good, as parents find it necessary to contribute their own money—and quite a sizable amount—to ensure their children get a good education and matriculate at a better college.

Levin stated that these differences really matter and "the challenges are not well understood, I think, by a lot of researchers. They are certainly not well understood by those who use the results of PISA in order to recommend study of other countries and policy."

The bottom line is that there is enormous complexity underlying assessment scores, not just internationally but within the United States

itself. This complexity is not appreciated by those outside the research world or even by some researchers. Bridging this gap will require creative and sustained effort on the part of ILSA administrators, researchers, and the media. As many workshop participants suggested, this may well be the time to establish a committee modeled on the Board on International and Comparative Studies in Education (BICSE) that existed 30 years ago when the IEA's Trends in International Mathematics and Science Study (TIMSS) was being planned.

3

Policy Uses and Limitations

International large-scale assessments (ILSAs) are certainly of great interest to people in the education research world; as discussed later in this report, the body of research literature based on ILSA data is voluminous. But what is the utility of ILSAs for policy makers? What effects do they have on policy? What effects should (or could) they have on policy? Some take the positive view that ILSAs matter in the policy world because the results are used to spur reform and to avoid complacency by highlighting differences in achievement among nations. Others believe the results are all too often misinterpreted and can lead to a misallocation of resources.

In the paper that Leslie Rutkowski of the University of Oslo prepared for the committee (Rutkowski, 2017), she notes that over the past 20 years or so, international assessments have come of age and have assumed a prominent place in educational policy and research discussions. A standout in this regard is the Organisation for Economic Co-operation and Development's (OECD's) Programme for International Student Assessment (PISA), the results of which have stimulated considerable changes in many participating nations. An example is the education reform effort that took place in Germany as a result of that nation's disappointing showing in the PISA 2000, the so-called "PISA shock" (Ertl, 2006). Similar effects were experienced in several European countries, including Denmark (Egelund, 2008), Finland (Dobbins & Martens, 2012), and others (Grek, 2009). In the United States, PISA results have been likened to

Sputnik (Finn, 2010); they have also formed the basis for calls to improve U.S. educational standards (Duncan, 2013).[1]

ILSAs IN U.S. EDUCATIONAL POLICY MAKING

At the workshop, we heard from several workshop participants who have a great deal of experience in the trenches of educational policy making at the U.S. federal level. One such participant was Jack Jennings, who spent much of his career as a senior staff member on the U.S. House of Representatives Committee on Education and Labor, and founded the Center on Education Policy. Also present were Mark Tucker of the National Center on Education and the Economy and Michelle McLaughlin of the Knowledge Alliance.

Jennings traced the policy effects of international assessments in the United States. During the 1960s, the United States was a relatively inward-looking nation. ILSAs, in addition to Sputnik, contributed to pulling the United States out of that stance with regard to education policy. The fact that high-performing nations had higher academic standards helped bolster the argument for the adoption of similar standards in the United States. Many states adopted more rigorous standards in the 1980s and 1990s; and this cause was taken up at the federal level as well. Various incarnations of the Elementary and Secondary Education Act, most notably No Child Left Behind, required states to adopt rigorous standards and to test students against these standards. The United States has gone from giving tests little attention to placing extraordinary emphasis on them. Jennings cautioned that this heavy focus on assessment may give the impression that increasing student test scores is the principal purpose of education.

Jennings noted that, in 2016, the National Council of State Legislators (NCSL) announced the No Time to Lose effort, which drew attention to U.S. performance on PISA in order to build support for a variety of reforms, not simply raising test scores. NCSL also urged policy makers to look at what could be adopted in the state context from high-ranking nations such as Finland and Singapore. The report asserts that the United States is unprepared for the "twenty-first century economy" and that its

[1] On the other hand, Marshall Smith, who has held various high-level posts in the U.S. Department of Education (ED) under the Carter, Clinton, and Obama administrations, noted later that there actually has been little talk about PISA in ED. There was more talk in the 1990s, but an argument was almost never made internally that the United States needed to be more like other countries. Sometimes researchers gave congressional presentations about PISA, but the great variation in test scores within the United States was far more compelling to education leaders.

workforce is falling behind in terms of knowledge and skills. However, it adds:

> The good news is, by studying these other high-performing systems, we are discovering what seems to work. Common elements are present in nearly every world-class education system, including a strong early education system, a reimagined and professionalized teacher workforce, robust career and technical education programs, and a comprehensive, aligned system of education. These elements are not found in the U.S. in a consistent, well-designed manner as they are found in high performers. (NCSL, 2016, p. 3)

NCSL's project aims to identify policies that are in place in high-performing nations and adopt them in the United States.

Tucker stated, "the single most important research question facing the United States is how to identify the factors that contribute to superior education system performance" and "how to match the performance of a group of more than 25 countries that are outpacing the United States in reading, mathematics, science, and problem solving." In fact, these are primary concerns of his organization. Tucker does not argue for blindly copying policies, but rather to look for "common principles" that are in operation in successful countries but are not in operation here. "You are looking at the principles that differentiate the successful from the unsuccessful countries." It is assumed that adoption of such principles will change things for the better. In other words, a causal inference is being made; whether it should be is explored in more detail in subsequent chapters.

McLaughlin stated that education advocacy organizations often use ILSA results to pressure policy makers, particularly members and staff at the U.S. Congress. It is a fairly difficult task, as explained in Box 3-1. One way her organization has been successful is to present ILSA results in a way that connects them to shortcomings in human capital development, specifically jobs in a congressperson's district.

INTERNATIONAL COMPARISONS AND POLICY IMPLICATIONS IN OTHER FIELDS

Also presenting at the workshop were researchers who have conducted policy relevant cross-national research on child development and on aging. The purpose was to explore how such research succeeds, or not, in changing policy in those fields.

BOX 3-1
Michelle McLaughlin's Reflections on Lobbying the U.S. Congress

As president of Knowledge Alliance, an advocacy organization focused on federal investment in education research, Michelle McLaughlin spends a lot of time lobbying the U.S. Congress. She explained that this can be challenging, because many in the U.S. Congress are distracted with other issues, suspicious of international organizations, and have staff members with little time for presentations of data and statistical analyses.

McLaughlin observed:

> What's on their plate on any given day is so enormous, particularly I would say for senators it's even greater, I distinctly remember briefing [U.S.] Senator Harkin about something about how we're changing away from adequate yearly progress to the next thing in our bill when we were working on ESEA, and he was very into it and we were talking very intense, and the buzzer goes off to let them know it's time to vote, there's a vote on the floor.
>
> He's been in [the U.S.] Congress for a long time, so of course he ignores it for the first 15 minutes, and then eventually his scheduler opens the door, and the legislative director steps in, which is like you really do have to go now; he said it's a vote on whether we should have a no-fly zone over Syria.
>
> So he's going from this really intense conversation about education policy to whether we should have a no fly zone over Syria. This is what's on the plate of a U.S. Senator every day, so I think that's important to keep in mind. Staff really are key. Committee staff have a lot more time. People who actually work for the chairman or the ranking member, they really have the most time to focus on this stuff.

Child Development and Poverty

Elizabeth Washbrook, a professor of quantitative methods at the Graduate School of Education of the University of Bristol in the United Kingdom, provided insight into the conduct and the complexity of cross-national research, based in part on her experience participating in a major study on social mobility in four English-speaking countries: Australia, Canada, the United Kingdom, and the United States (Bradbury et al., 2015). The point of the work was to see what was replicable in the United States, as far as social policy was concerned, to reduce inequality:

> Our focus is on the gaps in achievement [among] children of different family backgrounds, and how those gaps in the United States compare to those in our three other countries. We do not argue that societies should try to compensate for all the different sources of un-

> equal opportunity. . . . Our comparative cases illustrate that there are other countries similar in many respects to the United States, where [achievement] gaps [among] families of different socioeconomic backgrounds are significantly smaller. (Bradbury et al., 2015, p. 8)

Instead of using ILSAs as outlined in Table 1-1, the research team used truly longitudinal data gathered on 30,000 students across the four nations as they progressed from 5 years of age to 14 years of age. One of the main goals of the study was to make judgments about social mobility and the success of each nation in overcoming the effects of poverty. They investigated two main research areas:

1. The degree of inequality in the skills children brought with them when they started school at age 5; and
2. The extent to which those inequalities widened or narrowed during the school years.

Bradbury and colleagues aimed to disentangle these two factors to assess how effective a school system is in addressing inequality. They found that of the four nations studied, the United States was least successful in overcoming the effects of early childhood poverty. It was possible to draw these inferences because the researchers used a truly longitudinal design that tracked individual students over time, which yields more credible evidence regarding growth and the possible factors that contribute to that growth. One question discussed at the workshop is whether this important observation from outside the education field—which echoes calls by methodologists who study education—suggests that it is time for some ILSAs to be designed with a longitudinal component (as discussed in Chapters 4 and 5).

AGING STUDIES

John Haaga, then acting director (and now director) of the Behavioral and Social Research Unit at the National Institute on Aging at the U.S. Department of Health and Human Services, expressed his surprise at how much attention is paid to the way the United States ranks in a number of areas, including sports, specifically baseball. "We spend a lot of money on the players, and we belong at the top." He made the interesting point that education rankings get quite a lot of attention as well, but for some reason, international comparisons of the health of U.S. citizens rarely get reported in the media: "In health, we're utterly impervious to these sorts of comparisons. . . . It's remarkable how well we can absorb bad news and simply ignore it."

Haaga provided examples of interesting cross-national aging studies:

- A longitudinal study on the relationship between education/supplemental educational services (SES) and mortality in both the United States and Costa Rica showed that despite lower levels of education and overall wealth, the latter has quite good health and mortality outcomes (Rosero-Bixby & Dow, 2016). This study followed individuals from age 50 until death.
- A cross-sectional study showed the effect of retirement on cognitive functioning. In some European countries, tax and pension policies create a strong incentive for people to retire earlier than in the United States. The study found that in nations with public policies rewarding early retirement, cognitive functioning for older men was worse than it was in the United States. Was that causal? The study set off a wave of discussion about what could be done to improve people's situations in retirement (Rohwedder & Willis, 2010).
- There is also considerable research being done on the relationship between education levels and dementia. Data show that dementia is on the decline globally, and the best, actually only, currently available explanatory variable is the level of educational attainment early in life. Why early-life educational experiences would affect instances of dementia 40 or 50 years later is still a mystery.

With regard to the data needed to conduct this type of research, Haaga, like others, emphasized that most of the payoff in this type of research would come from an analysis of truly longitudinal data collected at the individual level, such as the Costa Rica study described above.

In the field of public health there is movement afoot to attain more "harmonization" of data across countries. "Harmonization" refers to using common definitions and measures across nations for psychosocial attributes like dementia or depression. These attributes may mean different things or be measured in different ways in different countries; for example, researchers in the United States and in Europe cannot agree on the definition of depression. Even medical or technical terms may have different meanings. Information on the policy and the institutional context for each nation is also helpful.

Haaga emphasized that in the social epidemiology field of aging studies, there is growing interest in the role of educational attainment in improving health and quality of life during retirement years. More of these connections are being made, which is why Haaga, as a researcher on aging, is concerned with stagnation in U.S. educational attainment (i.e., the highest degree a person has attained, such as a high school diploma

or a doctorate degree). There is the possibility that this stagnation affects health later in life. The percentage of adults with high school and college degrees, as well as the percentage of people with less than a high school degree, is roughly the same for American adults aged 65–69 years as it is for adults aged 25–29 years. The big increase in educational attainment is over, which Haaga fears may cause future rates of dementia to stabilize rather than to fall further. "We are not expecting any more improvement in educational attainment like the one that we think caused this improvement in dementia rates and a lot of other outcomes in the past decades. We've done it. Unless something big changes or unless we figure out what the magic ingredient is and start doing it to midlife adults, we've had our improvement."

These presentations ultimately illustrated that education researchers can learn a great deal from the research methods used by social scientists in other fields. Both Washbrook and Haaga stressed that much of this remarkable cross-national research would not be possible without truly longitudinal data collected at the individual level.

POLICY-CENTERED DESIGN

How might ILSAs be made more useful for informing education policy? Henry Braun recommends the adoption of a "policy-centered design" approach, where both the cognitive assessment and the background questionnaire components are designed to gather information relevant to key policy issues, which is not effectively happening now. One positive step is that more attention is being paid to the background questionnaires (which take students about 30 minutes to complete) in international studies. Braun does not see quite the same level of rigorous thinking about background questionnaire design as that which goes into the cognitive domains. He hopes that improvements on that front will provide a stronger evidence base for the kinds of policy questions that are of greatest interest to researchers and policy makers. One approach to achieving this goal would be to develop a small set of policy relevant questions to be addressed by secondary analysts and to have the advisory teams for both the cognitive and the background questionnaires commit to providing the data necessary to carry out the appropriate analyses. We discuss this in more detail in Chapter 4.

4

Design Issues

A recurring theme at both workshops was the mismatch between the types of questions researchers want to ask and the way that international large-scale assessments (ILSAs) have been, and continue to be, designed. As Marshall Smith of the Carnegie Foundation for the Advancement of Teaching noted, "design relies on intent." In an ideal world, the design of any study derives from its purpose, but because ILSAs have numerous possible purposes, as outlined in Chapter 1, strong linkages between all possible purposes and a single design is not possible to achieve. Current ILSA designs largely follow that of earlier decades, when the research questions were primarily descriptive, while current research questions—which increasingly focus on why we observe the patterns we do—require different kinds of designs.

The central issue—and the one that generated the most disagreement at both workshops—is whether ILSA data could ever support causal inferences, and if so, what design changes would be required. Workshop chair Judith Singer of Harvard University stated that there is legitimate disagreement in the field about whether causal inferences can be drawn from any designs other than randomized controlled trials (RCTs) or strong quasi-experiments (e.g., regression discontinuity designs). Even these designs—which, all agree, provide more compelling evidence of causality than the aggregate cross-sectional observational data typical of ILSAs—usually provide effect sizes at only a single point in time, on only a single measure, and often just at a single location (Ginsberg & Smith, 2016). Scholars have different proclivities, based on their disciplinary

BOX 4-1
Two Important Design Features of Modern ILSAs

When students participate in a modern ILSA, they do two things: they take one or more assessments of subject-matter achievement, and they fill out a background questionnaire. Here we explain how each is designed and administered (we discuss them in reverse order, because as described below, the background questionnaires are also used in estimating student achievement).

Background questionnaires. Participating students (and in some ILSAs, school administrators, teachers, and parents) each fill out a background questionnaire. The student background questionnaire, which typically takes about 30 minutes to complete, asks questions about topics such as the parents' level of education, the number of books in the household, television watching habits, computer use, etc. Workshop participant Leslie Rutkowski said that these background questionnaire data are important not only for "contextualizing achievement" but also "increasingly as outcomes in their own right beyond achievement, such as [indicators of] affective, behavioral, experiential constructs." Background variables are also vital to the process by which aggregate test score distributions are produced (as we describe next).

Achievement items are administered using a matrix sampling approach. ILSA designers face a challenging data-collection problem: there are many more items that they would like each student to take than there is testing time. For example, if each student was required to answer all the items developed for the Trends in Mathematics and Science Study (TIMSS) 2011, more than 10 hours of testing time would be needed (Mullis et al., 2009). To minimize the amount of time each student has to participate, each student now answers only a subset of the entire pool of test items using what is known as a matrix sampling design. The exact subset of items given to each student is selected using a sophisticated test booklet design that rotates the entire set of items across booklets. In TIMSS 2011, for example, test developers divided the total test content into 14 non-overlapping mathematics blocks and 14 non-overlapping science blocks. These blocks were subsequently arranged into 14 student test booklets, each containing two science and two mathematics blocks; and each student randomly receives one of the 14 booklets. This design ensures linking across all blocks because each block (and therefore each item) appears in two different booklets paired with different blocks. Because each student takes only a fraction of the item pool, psychometricians use a "plausible values" methodology to quantify each student's "proficiency distribution" in the tested domain. Without delving into technical details, this distribution is estimated using the student's answers to items in the background questionnaire, as well as the student's responses to the achievement items administered. The matrix sampling design prioritizes content coverage across the population of students over precision of individual student scores. As we discuss in Chapter 5, the matrix sampling design has major consequences for statistical analysis because individual student scores cannot be reported. Although the matrix sampling approach may sound unusual, it is actually standard in other large-scale assessment programs, including the U.S. National Assessment of Education Progress (NAEP).

backgrounds and their views about what data and designs are required to draw causal inferences. "These differences in opinion are not specific to ILSAs, but came up acutely at the workshop because ILSAs are not designed to answer [most of] the questions we're trying to ask," she said.

In this chapter, we begin by discussing the original purposes of ILSAs and how they influenced the tests' initial design. We then explain how ILSAs are currently designed and administered, and the extent to which they meet the needs of the research community, with results that, in turn, inform the policy community and the public. What was the original purpose of ILSAs that shaped their design? Can improvements be made? Because understanding these arguments requires some technical background, we begin with Box 4-1 that explains two important design features of modern ILSAs.

ORIGINS OF ILSA DESIGNS

Ina Mullis, professor at Boston College's Lynch School of Education and executive director of the International Study Center (ISC) for TIMSS and the Progress in International Reading Literacy Study (PIRLS) since the 1990s, reminded the workshop audience that TIMSS and PIRLS were designed originally by researchers for school improvement, not for national accountability.

> We are devoted to assessing what countries expect students to learn, and this is accompanied by an extensive array of background questionnaire data. Some people forget that IEA (the International Association for the Evaluation of Educational Achievement) started FIMS (the First International Mathematics Study) to study background issues, not [only] to measure achievement, so our studies have their roots most definitely in the questionnaire data.

The aims were to provide descriptive information to participating nations and to be as helpful as possible to nations that did not have robust testing systems of their own. Unfortunately, according to Mullis, the ranking aspect (i.e., the way the achievement results are portrayed in the media with lists of top performers) can have a negative effect in terms of political support for expanding and improving ILSAs (although we note that the number of participating countries has increased over time).

One way ILSA designers have tried to make the tests more useful is to customize the assessments and the background questionnaires for some countries. For example, the International Study Center (ISC), which administers TIMSS and PIRLS, started to see "floor effects" on TIMSS and PIRLS for some nations that had less-rigorous curricula. (A "floor effect" occurs when too many students are unable to answer any questions cor-

rectly; a "ceiling effect," discussed later in this chapter, occurs when too many students are able to answer all questions correctly. Both can be problematic because they make it more difficult to distinguish students from one another.) Thus, the ISC team put a great deal of effort into developing less-difficult assessments for some countries. These assessments—the TIMSS Numeracy and the PIRLS Literacy—are still aligned with the TIMSS and the PIRLS frameworks and scores can be reported on the TIMSS and the PIRLS scales.

THREE PRESSING DESIGN ISSUES

Leslie Rutkowski of the University of Oslo, Norway, wrote one of the commissioned background papers, *A Look at the Most Pressing Design Issues in International Large-Scale Assessments*, which explores three dilemmas:

- Cross-cultural comparability of items, particularly of those on the background questionnaires;
- Data quality and measurement error; and
- Potential inclusion of a longitudinal component in ILSA design.

We discuss each of these in turn.

Cross-Cultural Comparability

The first issue concerns cross-cultural comparability of both the achievement tests and the background questionnaires. This concern has become more acute as the number of countries included in the administration of ILSAs has expanded over time—now standing between 30 and 80, depending on whether the assessment is TIMSS, PIRLS, or the Programme for International Student Assessment (PISA; see Table 4-1 for the list of countries participating in each). Among the participating countries are not only Organisation for Economic Co-operation and Development (OECD) countries, but also less-developed countries. This breadth of participating nations leads to cultural differences that are ever more wide ranging, and these differences must be attended to when designing, administering, and interpreting the data.

Those who work with ILSA data are interested in the degree to which individual items, as well as indicators of latent variables (e.g., attributes not directly observable, such as a teacher's beliefs about teaching or mathematical proficiency), can be validly compared across populations. For example, the latent variable known as "problem-solving" might mean something different in the United States than it does in Botswana. If it does, as is likely, we risk errors of inference in these situations.

TABLE 4-1 Countries Participating in the Most Recent ILSA Administrations

	TIMSS 2015	PIRLS 2016	PISA 2015
	Africa		
Algeria			✓
Botswana	✓	✓	
Egypt	✓	✓	
Morocco	✓	✓	
South Africa	✓	✓	
Tunisia			✓
	Asia Pacific/Middle East		
Australia	✓	✓	✓
Bahrain	✓	✓	
China, People's Republic			✓
Hong Kong SAR	✓	✓	✓
Indonesia	✓		✓
Iran, Islamic Rep. of	✓	✓	
Israel	✓	✓	✓
Japan	✓		✓
Jordan	✓	✓	✓
Kazakhstan	✓	✓	✓
Korea, Rep. of	✓		✓
Kuwait	✓	✓	
Lebanon	✓		✓
Macao SAR			✓
Malaysia	✓		✓
New Zealand	✓	✓	✓
Oman	✓	✓	
Qatar	✓	✓	✓
Saudi Arabia	✓	✓	
Singapore	✓	✓	✓
Taiwan	✓	✓	✓
Thailand	✓		✓
Turkey	✓		✓
United Arab Emirates	✓	✓	✓
Vietnam			✓

continued

TABLE 4-1 Continued

	TIMSS 2015	PIRLS 2016	PISA 2015
	Europe		
Albania			✓
Armenia	✓		
Austria		✓	✓
Azerbaijan		✓	
Belgium	✓	✓	✓
Bulgaria	✓	✓	✓
Croatia	✓		✓
Czech Republic	✓	✓	✓
Denmark	✓	✓	✓
Estonia			✓
Finland	✓	✓	✓
France	✓	✓	✓
Georgia	✓	✓	✓
Germany	✓	✓	✓
Greece			✓
Hungary	✓	✓	✓
Iceland			✓
Ireland	✓	✓	✓
Italy	✓	✓	✓
Kosovo			✓
Latvia			✓
Liechtenstein			✓
Lithuania	✓	✓	✓
Luxembourg			✓
Macedonia			✓
Malta	✓	✓	✓
Moldova			✓
Montenegro			✓
The Netherlands	✓	✓	✓
Northern Ireland	✓	✓	
Norway	✓	✓	✓
Poland	✓	✓	✓
Portugal	✓	✓	✓
Romania			✓
Russian Federation	✓	✓	✓

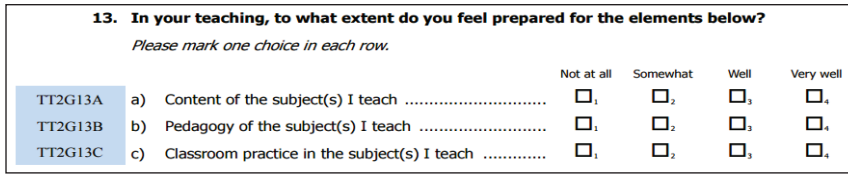
TABLE 4-1 Continued

	TIMSS 2015	PIRLS 2016	PISA 2015
Serbia	✓		
Slovak Republic	✓	✓	✓
Slovenia	✓	✓	✓
Spain	✓	✓	
Sweden	✓	✓	✓
Switzerland			✓
United Kingdom	✓	✓	✓
North America			
Canada	✓	✓	✓
Costa Rica			✓
Dominican Republic			✓
Mexico			✓
United States	✓	✓	✓
South America			
Argentina			✓
Brazil			✓
Chile		✓	✓
Colombia			✓
Cyprus	✓		
Peru			✓
Trinidad		✓	✓
Uruguay			✓

SOURCES: http://timss2015.org/timss-2015/about-timss-2015; https://nces.ed.gov/surveys/pirls/countries.asp; http://www.oecd.org/pisa/aboutpisa/pisa-2015-participants.htm.

To understand these issues, we need to introduce two further technical terms:

- **Differential item functioning** (DIF) refers to the phenomenon that a test item—even if its wording is as similar as it can be in different languages—may actually measure different things in different populations. DIF usually refers to differences across sub-populations within a single country; but in the ILSA context, it also refers to differences across populations in different countries.
- **Measurement invariance** is, in a way, the opposite of DIF: test designers want to know that a test as whole is measuring the same latent variable across different populations and sub-populations.



13. In your teaching, to what extent do you feel prepared for the elements below?

Please mark one choice in each row.

			Not at all	Somewhat	Well	Very well
TT2G13A	a)	Content of the subject(s) I teach	☐ 1	☐ 2	☐ 3	☐ 4
TT2G13B	b)	Pedagogy of the subject(s) I teach	☐ 1	☐ 2	☐ 3	☐ 4
TT2G13C	c)	Classroom practice in the subject(s) I teach	☐ 1	☐ 2	☐ 3	☐ 4

FIGURE 4-1 Sample question from teacher survey portion of TALIS.
SOURCE: Organisation for Economic Co-operation and Development: oecd.org/education/school/TALIS-2013-Teacher-questionnaire.pdf.

Evaluation of DIF and measurement invariance involves examining the behavior of an item, or set of items, across different groups identified on the basis of background characteristics (and in the case of ILSAs, different countries). Rutkowski illustrated these concepts using a set of three items that appears on the background questionnaire in the teacher survey that is part of the Teaching and Learning International Study (TALIS; see Figure 4-1).

Researchers have found that in some countries respondents do not use all of these categories and there is great variation across countries—not necessarily due to "true" differences in teachers' assessments of their level of preparation—that affect teachers' answers. She went on to note that "there are other types of cultural differences and response styles, such as acquiescent or extreme response styles that are culture-specific" that also can contribute to DIF and diminish measurement invariance.

Rutkowski noted that the organizations that field ILSAs are increasingly exploring the possibility of allowing more variation across countries at the expense of complete uniformity by allowing more "country-specific" background questions that are not necessarily comparable. Adjustments are even being made on the achievement tests themselves, with the goal of greater accuracy at the low end of the achievement continuum. In 2009, for example, PISA began adding blocks of easy items for lower-performing countries because researchers noted floor effects. Contrasting ceiling effects are being observed in some Asian countries but have yet to be addressed. Rutkowski supported the idea of introducing blocks of more difficult items for high-performing countries.

Measurement Error

Psychometricians use the term "measurement error" to refer to the difference between a variable's observed value and its "true value." ILSA administrators and researchers are correct to worry about the impact of measurement error, particularly in responses to background questions (much more attention is given to the psychometric properties of the

achievement tests). In PIRLS, for example, the background questionnaire asks students "About how many books are there in your home?" and the response choices include graphics of bookcases with approximate numbers of books. The parental background questionnaire uses the same stem, but its response choices are given as ranges of numbers, without the graphics. One might expect some differences between what students and their parents report (and neither may reflect the true value), but what is especially intriguing is the differences across countries in the magnitude of these discrepancies. Across all tested nations, the average correlation between student and parent reports is 0.66. But in Azerbaijan and Kuwait it is only 0.35, while in Portugal it is 0.76, and in Georgia it is 0.92. These discrepancies suggest some sort of measurement error is at play, but it is difficult to know the reason why the variation in correlations across countries is so great.

A similar discrepancy occurred in PISA in 2012 when 15-year-old participants and their parents were each asked whether the student had ever repeated a grade. In Hong Kong and Portugal, up to 20 percent of the paired responses differed: one respondent (either the student or his or her parent) answered "never," and the other failed to answer the item (which may also contain information, even if we do not understand what that information might be). Rutkowski argues that as with the question about the number of books in the home, the reasons for this discrepancy are also unknown. Without culture-specific insights into all of these countries, it is challenging to judge the reliability and the validity of responses. These two examples are likely just illustrations of the general problem, which happen to be observable because of the dual administration of the same question. It is not a far leap to argue that there is likely substantial measurement error in many other variables, even if we cannot quantify the magnitude of the problem.

In her commissioned paper, Rutkowski (2017) argues that future ILSA designs should identify a subset of the highest-priority reporting variables that are known to be highly susceptible to measurement error and cross-cultural comparability issues and prioritize developing better measures. She singles out socioeconomic and sociocultural status as two high-priority domains. In economically advanced countries that collect census-type data on educational systems (e.g., Norway and the United States), more reliable measures of school district supplemental educational services can be derived, such as using data from the U.S. Census's Small Area Income and Poverty Estimates program. Although this still leaves a gap between what we know about specific students and the schools they attend, these sorts of census measures would be better and finer-grained than anything used to date in international assessments. But Rutkowski agreed with Kaplan that many policy-relevant measures, such as socioeconomic or

BOX 4-2
Can We Measure Socioeconomic Status Comparably Across Cultures?

On the subject of cross-cultural comparisons and ILSA background questionnaires, David Kaplan of the University of Wisconsin–Madison highlighted difficulties in measuring two interrelated concepts: socioeconomic status and household affinity toward academic pursuits (see also van de Vijver, Chasiotis, & Breugelmans, 2011).

To quantify these concepts, PISA has developed a construct referred to as economic, social, and cultural status (ESCS), which is an index based on several background questions. The definition is as follows:

> The Programme for International Student Assessment (PISA) index of economic, social, and cultural status was created on the basis of the following variables: the International Socio-Economic Index of Occupational Status (ISEI); the highest level of education of the student's parents, converted into years of schooling; the PISA index of family wealth; the PISA index of home educational resources; and the PISA index of possessions related to "classical" culture in the family home. (OECD, 2003)

Family wealth is based on students' responses to questions on whether they had the following in their home:

- Room of their own;
- Link to the Internet;
- Dishwasher (only in some countries);
- DVD player;
- Number of cellular phones;
- Number of televisions;
- Number of computers;
- Number of cars; and
- Rooms with a bath or shower.

The second element in the construction of this variable is the number of "cultural possessions" in the home, including

sociocultural status, are conceptualized and operationalized differently in different countries. Although a universal measure is a laudable goal, individual countries can still develop and include locally relevant measures to maximize the utility of the assessments for both reporting and use in within-country statistical analyses (see Chapter 5). Workshop participants agreed that augmenting the questionnaire data with data from other sources, such as U.S. Census data or administrative data collected by school districts, was a good idea.

- Classical literature;
- Works of poetry; and
- Works of art.

The third element is home educational resources, and these include

- Desks;
- Quiet places to study;
- Computers that students can use for schoolwork;
- Educational software;
- Books to help with students' schoolwork; and
- Technical reference books and a dictionary.

These three elements are combined with the family breadwinner's occupational status and years of schooling to form the ESCS variable.

Kaplan was critical of this effort. First, a presumption is made that a particular culture actually has a tradition of a "classical literature," and that cultures would equally value works of poetry or art as a marker of status. Home educational resources measure the existence of desks and a quiet place to study. Kaplan quipped, "Those of you who have teenagers know that they don't study at their desk. They often will hang from the ceiling if they can to study, but nowhere near a desk." The flaw here is that family wealth, a checking-off of material possessions, cultural possessions, and home educational resources, is scaled under an assumption that there is a single underlying latent variable, and that this latent variable is the same in all countries. After those estimates are provided, they are combined with other variables into a single ESCS measure. "It's hard to justify an underlying latent variable generating a tick-off list of possessions, yet that's exactly what they do, and it's part and parcel of the ultimate component that defines ESCS," Kaplan said. He urged efforts to re-evaluate how supplemental educational services and related constructs are measured in ILSAs. Rutkowski concurred and added that the variable "doesn't work really well anywhere." They have also found a ceiling effect for the question about smartphones; for example, in a country like Norway, where every student has a smartphone, the question yields no variation.

Could (Some) ILSAs Be Designed to Include a Longitudinal Component?

Almost all ILSA data-collection efforts have been cross-sectional. Within each country, according to an agreed on sampling scheme, a statistically representative sample of students is selected for data collection, which occurs during a fixed finite time window. Even if it might appear as if an ILSA provides a type of longitudinal data—as when TIMSS assesses fourth graders one year and eighth graders 4 years later—they are not

truly longitudinal because the tested students are rarely the same, nor are two sets of data collected several years apart nationally representative samples of the same birth cohort because of immigration, emigration, students repeating a grade, and students dropping out.

Truly longitudinal data refer to designs in which the same students are tracked over time. Participants in both workshops noted the interesting data that would be available if some ILSAs expanded to include a truly longitudinal component. This was especially true following the discussion of Elizabeth Washbrook's paper (Bradbury et al., 2015; see Chapter 3), for which she and her colleagues were able to compare student trajectories of growth over time (as opposed to the analyses of status at a single point in time typical of most ILSA analyses). By restricting their analyses to data collected in four English-speaking countries, Washbrook and colleagues minimized many of the complex cross-cultural methodological issues raised earlier in this chapter, allowing them to exploit a fairly unique and powerful analytical opportunity.

Rutkowski suggested that TIMSS might provide one opportunity to collect truly longitudinal data. She highlighted TIMSS because it already tests fourth and eighth graders, so new assessments would not need to be developed. She suggested that in 2019, for example, it would be possible to test the same sample (or perhaps just a sub-sample) of fourth graders already assessed in 2015 using a set of linked items that would allow measurement of progress over time. Realistically, however, especially given the varying rates of transience and migration within and across different countries, this would require significantly more resources than are currently available to ILSA administrators.

There was also considerable disagreement among workshop participants about whether the difficulty of tracking large numbers of students as they progress through school would be worth the methodological payoff. Ina Mullis noted that longitudinal data in TIMSS "is an idea that has come up over and over and over." Mullis continued, "Every once in a while there's a small country where maybe they have a better idea of where students are from day to day, and we talk to them about the possibilities of doing a truly longitudinal design." As discussed in Chapter 5, several countries have run their own longitudinal follow-ups on ILSA samples, including Australia, Canada, Denmark, and Germany.

Even if feasible, truly longitudinal data are not a panacea as they cannot unequivocally yield causal inferences. Longitudinal data allow researchers to estimate rates of growth and change over and above ILSAs typical data about status. But even with truly longitudinal data, researchers cannot be certain that other plausible, intervening causes of growth and change are responsible for observed patterns. We return to this theme in Chapter 5.

IMPLICATIONS OF DIGITALLY BASED ASSESSMENT

Henry Braun of Boston College discussed how digitally based assessment (DBA) has the potential to improve ILSA designs. The improvements would not come so much from the better interpretation of results, but rather in the areas of improved administration, measurement, and data processing.

Administration

One benefit of DBA is that the platforms developed can be adapted and used in other testing settings. The emergence of a new testing infrastructure would allow developing nations to gain expertise in assessment, not only on the measurement side but on the administrative side, leading to improvements in the overall quality of assessment globally. Braun expects that DBA might be less error prone and that quality control may improve because test administrators could monitor students as they take the test in real time.

The challenge will be building and troubleshooting the new infrastructure, as well as training test administrators, both within and across countries. The shift from "paper-and-pencil" to DBA means that for a short transitional period, dual testing systems will have to be maintained, which complicates not only test administration and analysis, but also adds cost. PISA and the Programme for the International Assessment of Adult Competencies (PIAAC) have completed the shift to DBA, and according to Mullis, eTIMSS and ePIRLS are in the works.

Measurement

DBA can improve construct representation because it enables the development of new types of assessment tasks that can provide deeper and broader evidence of a student's knowledge and skills, compared with a test based solely on multiple-choice questions. This will be particularly useful as ILSAs move into measuring higher-order skills such as scientific problem-solving. DBA can also yield evidence on the amount of time it took for students to solve a problem, and in some cases, the specific strategies students used to do so.

DBA can improve the accuracy of assessment through adaptive testing. In adaptive testing, the difficulty of questions, or set of questions, is tailored to the current (i.e., time-varying) estimate of students' proficiency level. Generally, this strategy improves precision of proficiency estimates for a fixed amount of testing time.

But even with DBA, several challenges remain with regard to measurement. First, there are still constraints on the total amount of time that

each student spends taking an ILSA, which limits the number of tested domains per student, even with the booklet design described earlier. Second, the greater accuracy of a computer-adaptive design may be offset by the introduction of construct-irrelevant variance, that is, the introduction of extraneous factors tied to computer administration that change the nature of what the test question is measuring, at least for some students. For example, student performance may be inappropriately affected if students are unfamiliar with the devices used for test administration or faced with novel test item formats. Mullis stated that in many countries, students do a great deal of reading and other work online, so a reading assessment should also assess students' online reading capabilities.

Data Processing

DBA will inevitably bring more sophisticated and accurate data-management systems. Expert systems (e.g., software that uses artificial intelligence) can be used to computer-score open-ended responses in both written and graphical formats. Braun also expects further developments in test score scaling and reporting.

For Braun, the bottom line on technological developments in ILSAs is that in the near term, DBA is not likely to have a major impact on how ILSAs are used in educational policy making or in research (e.g., secondary analysis). Most improvements will be behind the scenes. However, the widespread introduction of DBA does present an opportunity for some creative rethinking of ILSA designs over the long term.

5

Analysis

Researchers who analyze international large-scale assessment (ILSA) data use a diverse array of analytical approaches, each with methodological advantages and challenges. In summarizing the state of ILSA data analysis, Judith Torney-Purta of the University of Maryland argued that most studies—whether using data from a single country or multiple countries—either focus on seemingly unique features of countries that are consistently high performing or rapidly improving, or seek to identify statistically significant predictors of student achievement. No wonder we see media reports highlighting one predictor after another as "the reason" one country (or set of countries) outperforms others. In 2015, Finland's high rankings, coupled with the lowest mean number of hours spent on homework, led many to suggest that U.S. schools should decrease the amount of homework assigned. In 2016, Shanghai's high rankings, coupled with its mastery-focused textbooks, led Great Britain to translate those books from Chinese into English and introduce them in almost half of its primary schools, as if new textbooks alone could change student achievement.

Workshop participants agreed that the most important educational research questions require more nuanced analyses that at least attempt to account for—or better yet, effectively rule out—the wide array of community, school, classroom, and home characteristics that affect student achievement. But analyses like these are extremely difficult to conduct using cross-sectional ILSA data. It does not take a statistician to see that

when it comes to many ILSA analyses and interpretations, nuance is in very short supply.

CAUSAL INFERENCE: THE HOLY GRAIL

For nearly 100 years, statisticians and philosophers have written treatises about the conditions necessary for drawing causal inferences. Even the seemingly simple word "cause" has multiple meanings that we will not explore here, but we instead cite R. A. Fisher's 1935 urtext, *Design of Experiments*, in which he argued that randomization of units to treatments would lead to the most reasoned basis for inference. Fisher's perspective evolved during his years at Rothamsted Experimental Station, where the units (i.e., mostly plants and plots) were quite amenable to randomization and the assignment of treatments (i.e., mostly fertilizers) could be easily controlled by researchers. Although context mattered as well—rainfall, for example, is one obvious important factor outside researchers' control—randomization of units to treatments ensured that, on average, with large enough sample sizes, it is safe to assume equivalence between treatment and control groups with respect to all observable, and even unobservable, characteristics.

This ability to assume equivalence has made randomized controlled trials (RCTs) the "gold standard" for drawing causal inferences. But in complex field settings such as education, RCTs are not without critics. Obviously, students are not plants and educational "treatments" are not fertilizers. Education is inherently social, and many factors outside school matter enormously in determining educational outcomes; some would argue these external factors matter more than what goes on in school. Others worry that the quest to isolate specific "causes" of educational outcomes has had the negative effect of substantially narrowing the focus of the educational research enterprise. Still others argue that the obsession with formal causal inference has crowded out other methodological concerns, especially about generalizability across time and place (Ginsburg & Smith, 2016).

In technical terms, RCTs privilege internal validity, that is, the ability to draw causal inferences for the sample of units under study, and ILSAs privilege external validity, that is, the ability to generalize from a sample to a population. Workshop participants did not debate the unanswerable question: which is more important? But methodologically attuned readers may have already noted that we consistently deferred the topic of causal inference from Chapter 4 on design to this chapter on analysis. Deferral was a strategic decision given that workshop participants generally agreed that ILSAs are unlikely to incorporate RCTs any time soon, except, perhaps, in the context of specific measurement questions, for which fea-

tures like item wording can be easily randomized. Educational policies and practices are hardly randomly assigned to countries or units within countries, nor are they ever likely to be. Instead, they are the product of several historical and cultural factors that vary over time. Thus, in the context of ILSAs, questions of how to draw causal inferences devolve to questions of analytical strategy, not research design (as in the design of data collection, for which many researchers would correctly argue that their analytical strategies are a type of research design).

Whether any sophisticated statistical analysis can yield inferences closely approximating causal statements was a matter of debate during both workshops. Some participants expressed hope that statistical methods might be able to estimate something approximating causal effects. In a 2016 article, for example, Rutkowski argued:

> With regard to international assessments and surveys, natural experiments also occur. . . . Take two relatively similar countries (Norway and Sweden) and consider a situation where Sweden chooses to privatize [its] educational system while Norway chooses not to follow suit. Under certain assumptions, we can treat this situation as a natural experiment and estimate the difference in some outcome between the two countries. (Rutkowksi, 2016, p. 4)

A quasi-experimental approach statistically matches the treatment group to the control group (e.g., students in two different countries) on important measured covariates that are known (or believed) to affect "treatment assignment." If the groups can reasonably be regarded as statistically equivalent on relevant covariates, average differences on the outcomes could plausibly be attributed to treatment differences. But of course, one person's "reasonable equivalence" may be another's "not even close."

Others expressed doubt that the assessments themselves can support causal inferences. David Kaplan of the University of Wisconsin–Madison stated that the assessments would have to be completely redesigned with a small number of causal questions in mind. It is possible, Kaplan noted, that certain counterfactually constructed questions may be addressed by an existing ILSA, but it was doubtful that the necessary set of covariates would be available to support even quasi-experimental inference. Kaplan pointed out that ILSAs, in their current form, are conceptualized and designed as "monitoring indicator systems," providing a "bird's-eye view" of stability or change in the inputs, processes, and outcomes of education at the system level, and that the focus of attention should be on how to improve their utility for that purpose, both on the methodological side and on the reporting side. Drawing causal inferences with ILSAs in their present design structure, Kaplan felt, was misguided.

Similarly, longtime ILSA leader Ina Mullis of Boston College stated that ILSAs "are really not optimal for causal analyses." She expressed doubts that ILSA administrators would modify their designs to help move in this direction because of the costs and the difficulties faced by participating nations. "From a cost–benefit perspective, it's a huge burden for countries for what actually is perhaps a modest gain in the analysis and interpretation," she explained. While they have great potential for comparing populations of students, "they're ultimately observational." They tell us a lot about relative student achievement, but little about why students perform better in some nations than in others, and even less about what types of policies can be shown to work in one country and adopted in another.

Exploring different analytical perspectives on ILSA data is the central focus of this chapter. Our discussion is based largely on the commissioned paper by Anna Chmielewski and Elizabeth Dhuey (2017), both of the University of Toronto. Their review explores the claims researchers make, including some who argue that their analyses "come close" to supporting causal inferences that have the potential to provide useful policy guidance. We begin by outlining an array of analytical approaches and then we present critical evaluations of these strategies. In the interest of keeping the discussion accessible to a broad array of readers, we have prioritized accessibility over technical details; readers interested in more details would do well to delve into the Chmielewski and Dhuey (2017) background paper prepared for the workshop.

As will be clear by the end of this chapter, workshop participants disagreed on the fundamental question of just how close any of these analyses come to yielding credible causal inferences. But even when the evidence for a causal claim is lacking, most workshop participants agreed that such analyses can generate interesting hypotheses that might offer useful guidance for further research.

OVERVIEW OF ILSA ANALYTICAL METHODS

Chmielewski and Dhuey (2017) began by noting that ILSA data are "relatively underutilized by U.S. education policy researchers." Yet, in the United States and elsewhere, some researchers have used a variety of statistical methods to produce estimates that, they argue, describe the direction and the magnitude of causal relationships, as well as auxiliary results that provide some support for these causal claims.

Chmielewski and Dhuey organized ILSA analytical strategies into five broad categories based on the way in which researchers identify variation in policies or conditions, that is, the so-called "treatments" that students experience (see Table 5-1).

TABLE 5-1 Categories of Analytical Strategies Used by Researchers

Category	Strategy
1	Analyze policy variation across countries.
2	Analyze policy variation within countries.
3	Analyze repeated cross-sectional ILSA data to look at variation across birth cohorts or generations of students.
4	Analyze repeated cross-sectional ILSA data to look at variation across age within the same birth cohort within countries; known as "synthetic cohorts."
5	Analyze rare truly longitudinal ILSA data that follow the same students over time.

SOURCE: Chmielewski & Dhuey, 2017.

Categories 1 and 2 use cross-sectional data at a single point in time, while Categories 3, 4, and 5 examine policy variation over time. We explain and illustrate each of these five approaches below. We selected the corresponding examples because most participants agreed that they are creative attempts to approach making credible causal inferences using ILSA data. Some methods have been debated and critiqued; hence, their inclusion in this chapter does not represent the committee's endorsement. Rather, we included them because they collectively represent most of the interesting attempts to capitalize on the widespread availability of ILSA data.

Category 1: Analyze Policy Variation Across Countries

These kinds of studies—probably the most common analytical approach for ILSA data—use highly aggregated national test data to estimate correlations between policy characteristics (authentically described or measured at the country level) and student achievement. The correlations (or regression coefficients) can be uncontrolled or controlled but, regardless, they are still "just correlations." They cannot lead to causal inferences for many reasons. As noted above, policies are not randomly assigned to countries and are confounded with a wide variety of cultural and historical factors. As with all cross-sectional data, the causal arrow may go the other way; for example, the apparent "policy" in question could have been adopted as a response to patterns in student achievement, not vice versa. If analysts fail to document how long the specific policy being studied has been in place in each country analyzed—as most fail to do—their conclusions may simply be wrong.

Yet when certain policies that have been in place in numerous higher-

performing nations for a long period of time appear associated with higher achievement, researchers may have a slightly stronger argument that the policies in question have "led" to that higher achievement. Although most participants agreed that studies like these provide the weakest support for a causal link, one partial check would be to establish that the same policies are generally not in place in low-performing countries.

Some of these studies use a statistical technique known as instrumental variables analysis to take advantage of putative "natural experiments" with the goal of teasing out a clearer link. A valid instrumental variable is one that researchers can convincingly argue induces changes in a predictor but has no plausible independent effect on the outcome under study. Without delving into technical details, which are far too complex for this report, a successful instrumental variables analysis provides more compelling evidence that the predictor causes changes in the outcome. The interested reader would do well to consult Murnane and Willett (2010). We present an example of an instrumental variables analysis in Box 5-1.

Much of this research focuses on the relationship between economic factors measured at the country level—particularly levels of development, gross domestic product, or income inequality—and student achievement. For example, Chmielewski and Reardon (2016) have shown the strong association between achievement gaps and income inequality, as illustrated in Figure 5-1. In this figure, the size of each circle indicates the precision of each achievement gap estimate, and taking varying precision into account, we see that the higher the degree of income inequality in a nation, the larger the achievement gap.

Other studies in this tradition have focused on associations between student achievement and gender egalitarianism (Wiseman et al., 2009); curricular differentiation and tracking policies (Buchmann & Park, 2009; Chmielewski, 2014; Chmielewski, Dumont, & Trautwein, 2013; Chmielewski & Reardon, 2016; Marks, 2005; Pfeffer, 2008; Schmidt et al., 2015); and levels of socioeconomic segregation among schools (Willms, 2010). Occasionally, studies in this tradition go beyond achievement or economic matters. For example, Torney-Purta and colleagues (2008) used results from the International Association of the Evaluation of Educational Achievement (IEA) Civic Education Study to demonstrate an association between a country-level predictor (i.e., length of time the country had been a democracy) and students' knowledge and attitudes about human rights.

Category 2: Analyze Policy Variation Within Countries

Researchers can sometimes identify natural experiments when they can plausibly argue that the "treatments" within a country "approach"

BOX 5-1
Catholicism at the Turn of the Century and Present-Day Student Achievement

West and Wößmann (2010) used an instrumental variables analysis to study the relationship between public and private school competition (measured at the country level) and student achievement, exploring their hypothesis that school choice and competition improves educational outcomes. Their challenge was to identify a good instrument, a variable that would measure, at least in part, the level of school choice and competition in a nation at the time of data collection that is not plausibly also correlated with contemporaneous student achievement. In puzzling through options, the researchers decided on what many would believe to be an unusual choice (instrumental variables often seem unusual): the percentage of Catholics in each country in 1900.

Their rationale went as follows: at the end of the 1800s, local Catholic leaderships in both the United States and in Europe rebelled against the curricula in state-run schools. They formed their own parallel school systems. In nations where Catholicism was not a state-sponsored religion, private schools proliferated, so that countries with larger shares of Catholics in 1900 tended to have larger shares of privately operated schools today. The authors write:

> We [used] this historical pattern as a *natural experiment* (italics added) to estimate the causal effect of contemporary private competition on student achievement in cross-country student-level analyses. Our results show that larger shares of privately operated schools lead to better student achievement in mathematics, science, and reading, and to lower total education spending, even after controlling for current Catholic shares. (West & Wößmann, 2010, p.1)

Education costs were also lower in the historically Catholic nations. The authors assert that it was not Catholicism per se driving the results but the fact that the presence of Catholic schools opened the door to greater competition.

random assignment. Studies in this tradition identify variation within individual countries—but, once again, measured at a level of aggregation above that of the individual student—in terms of school effects, classroom size, and student-tracking policies, among others, and then examine variation in impacts across countries. Claims of causal linkages are bolstered if similar effects are found in multiple countries.

Category 3: Analyze Variation Across Birth Cohorts Within Countries

When a single country generates multiple observations over time, and when the results of parallel analyses are consistent, the argument that the link is causal is strengthened (at least somewhat, according to some workshop participants). Compared with Category 1, Category 3 "comparisons

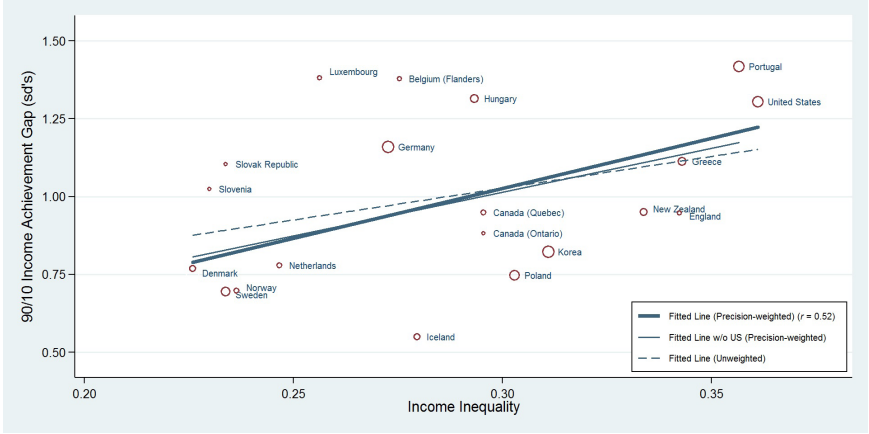


FIGURE 5-1 Association between income achievement gap and income inequality, wealthy Organisation for Economic Co-operation and Development countries, 2001 to 2012
NOTES: Pooled Progress in International Reading Literacy Study and Programme for International Student Assessment data. The size of each circle indicates the precision of the 90/10 gap estimate, with the larger circles indicating the most precisely estimated gaps.
SOURCE: Chmielewski & Reardon, 2016: http://journals.sagepub.com/doi/full/10.1177/2332858416649593.

are made within countries over time, rather than cross-sectionally at each time point for a changing set of countries." Of note, data like these are not truly longitudinal, which requires the same students to be tested at different time points.

Comparison of birth cohorts is most useful when some development or policy change occurs within a country that may affect student achievement. By examining different birth cohorts, researchers can identify whether these changes within a country may have affected achievement and see if this link holds in other countries as well. Box 5-3 presents one such example, but we caution that many workshop participants were not persuaded by the researchers' arguments that they have identified a causal link.

Category 4: Analyze Variation in Age Within the Same Birth Cohort Within Countries

Studies like these also use repeated cross-sectional data to examine change within countries over time, but the researchers are not interested

in change across different age cohorts, but rather in how variables evolve as students progress through the education system. Lacking student-level longitudinal data, studies in this tradition match different cross-sectional ILSAs by the birth year of the test population to create so-called "synthetic cohorts." Chmielewski and Dhuey (2017) argue that:

> since 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. 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. (p. 11)

Researchers then examine differences in trend lines between nations with and without a certain policy or attribute in place, using what is called a "difference in differences" approach.

BOX 5-2
When Should a Child Enter Kindergarten?

All nations (and in the United States, states and school districts) have a cutoff date for student birthdays that determines when a child will enter kindergarten; if a child's birthday is just after the cutoff date then he or she usually has to wait another full year to enter school. Some children will have their fifth birthday just before the cutoff date for kindergarten, while others will be almost 6 years old before they can enroll. This means that the age span in a given grade can be as large as nearly 1 year.

What is the long-term impact of this age difference in kindergarten entry? The hypothesis is that, while there may be initial differences in student achievement due to differences in maturity, these should diminish as students progress through school. Bedard and Dhuey (2006) used the natural, within-country variation in age at entry into kindergarten to study the effects of school entry on student achievement. The researchers plausibly argued that a child's month of birth should be considered random, so we would expect no observable (or unobservable) differences among students born in different months. Using Trends in International Mathematics and Science Study (TIMSS) data from 19 countries, the researchers compared the test scores of older and younger students within a grade, both within and across countries. They found that age at entry predicts student achievement, and that this link occurs across a wide range of countries. Younger students—in both fourth and eighth grade—scored significantly lower, on average, on the TIMSS assessments than their older peers who missed the cutoff, showing that despite the hypothesis that differences would dissipate, they actually persisted into adolescence.

BOX 5-3
Computers in the Home: Good for Reading?

As one would expect, background questionnaires administered to students who take ILSAs indicate that increasing numbers of students have a computer at home, and the growth in households with computers is most pronounced in developed nations. It is assumed, and some studies have shown, that computer use at home can have a positive effect on student achievement, but overall the research literature is mixed on this matter.

Rosén and Gustafsson (2016) looked at performance on PIRLS from 1991 to 2001, and then from 2001 to 2006, in order to estimate the association between home computer use and reading achievement. Their hypothesis was that "increased computer use at home has a negative effect on reading achievement and that this can be explained by displacement theories . . . the reallocation of time in favor of computer activities [results] in less time being allocated to reading for enjoyment" (pp. 1–2).

The researchers used what they call a "longitudinal cross-cohort design." They write:

> When multiple observations of each unit are available, such as with a longitudinal design, this information can be used in such a way that units are made into their own controls. In that way, the effect of unit characteristics, which remain constant over time, are removed and information about these fixed characteristics can be omitted without causing any bias. (p. 5)

An argument like this is most compelling when the data are truly longitudinal—that is, the same students are followed over time—but these researchers argue that similar claims can be made when using aggregate data on different students (a claim not supported by all workshop participants).

Rosén and Gustafsson (2016) analyzed data aggregated to the country level for 19 developed nations. Their first year of data was 1991 when very few people had computers in their homes. They then looked at changes in responses to items

Category 5: Analyze Variation in Truly Longitudinal Data

Only one ILSA has attempted to follow the same students over time, creating a truly longitudinal data set: the Second International Mathematics Study (SIMS) of the early 1980s tested the same students twice, once when they were in seventh grade and again when they were in eighth. One noteworthy secondary analysis of these data was by Zimmer and Toma (2000), which found positive effects of having high-achieving peers in a classroom, especially for lower-ability students; this is the only ILSA longitudinal study of which we are aware that attempted to use truly

about computer use from 1991 to 2006. They found that, on average, countries where the percentage of students using a computer at home is higher have lower reading scores. However, the magnitude of the negative association diminished when the researchers controlled—once again, at the aggregate level—for responses to a background question that asked whether students borrowed and read books from the local library, which they argued served as a proxy for interest in reading. The negative association was most pronounced for the reading of literary texts (i.e., those which required sustained attention) as opposed to short informational texts, and the negative effect was larger for male students than female students.

Many researchers oppose using highly aggregated data like these to estimate what are often referred to as ecological correlations, or correlations among aggregate means that are intended to draw inferences about relationships among variables at the individual level. (Note that this issue is distinct from the issues raised earlier about estimating correlations between at least one variable appropriately measured only at the aggregate level, such as an educational policy decision.) Although the sociological literature identifying concerns about ecological correlations dates back more than 100 years (to Durkheim), it was Robinson's (1950) classic paper that highlighted these concerns by demonstrating why one can never infer the magnitude, let alone the sign, of a correlation estimated using individual-level data on the basis of one estimated using aggregate data. His dramatic example used 1930 U.S. Census data to show that at the individual level, the correlation between immigration status and illiteracy was small and positive (0.12), that is, immigrants were more likely to be illiterate. At the state level, not only was the correlation far stronger in absolute value, but it was negative (–0.53), that is, states with larger percentages of immigrants had higher literacy rates. This seeming discrepancy—which is not a discrepancy once one realizes that the two correlations describe different relationships—was because immigrants tended to settle in states where the native population was more literate. We return to these issues later in this chapter. Interested readers should consult Rosén and Gustafsson (2016) for their response to these concerns.

longitudinal data to make causal inferences. However, as Judith Singer noted, "two waves of data does not a longitudinal study make."

Over the years, some nations have taken it upon themselves to add a truly longitudinal component to an ILSA by following the same students who participated in the cross-national study into young adulthood or the workplace. Canada and Denmark retested PISA students, while Australia and Switzerland followed students and administered background questionnaires, but not achievement tests. These studies aimed to follow teens into young adulthood, looking at critical transitions from school to the

BOX 5-4
Does Tracking Work?

One well-known study of this type examined the effects of "ability tracking" in schools (Hanushek & Wößmann, 2006). In some nations, ability tracking begins as early as age 10, but usually before age 16 (e.g., France, Germany, Greece, and the Netherlands). Other nations have some tracking after age 16 or eschew it altogether (e.g., Hong Kong, Iceland, Norway, and the United States). Supporters of tracking argue that it is more efficient. By grouping students of similar ability in the same classroom, schools can have a more "focused curriculum and appropriately paced instruction that leads to maximum learning by all students.... The teacher does not have to worry about boring the fastest learners or losing the slowest learners" (pp. 1–2). Detractors of tracking argue that this practice perpetuates inequality because students in lower tracks are usually from lower socioeconomic backgrounds or otherwise disadvantaged populations.

The authors investigated the association between curricular tracking and country average performance and inequality. They matched assessments of fourth-grade students (i.e., TIMSS fourth grade and PIRLS) with those of lower secondary-school students (i.e., TIMSS eighth grade/age 15 in the Programme for International Student Assessment [PISA]). This difference-in-differences design allowed them to compare changes within country cohorts.

The authors measured inequality using several variables, such as the standard deviation of test scores and the differences between upper and lower percentiles. They found that in nations with tracking policies, inequality in scores increased in the upper grades; this was particularly true in the Czech Republic, Germany, and Greece. In most nations that do not track—including Canada, New Zealand, Turkey, and the United States—inequality in scores tended to be lower in the upper grades compared with that in the lower grades.

Results on achievement, in contrast, were mixed. Tracking was associated with lower reading and math performance (results were weaker for math in terms of statistical significance) but somewhat higher science performance. Outcomes tended to be worse for lower-performing students in the tracking countries than in non-tracking countries. Hanushek and Wößmann conclude that there is little to be gained in terms of efficiency from tracking policies, and the "results suggest that countries [that track] lose in terms of the distribution of outcomes, and possibly also in levels of outcomes, by pursuing such policies" (p. 14). But as with all ILSA analyses attempting causal inference, some critics—for reasons explained in this chapter's introduction—do not find their conclusions persuasive.

workforce (e.g., graduation, college attendance, first job, and satisfaction with various aspects of life).

Unfortunately, all of these studies have been plagued by attrition problems. Despite this issue, the countries are using these studies to inform policy-making within a country (i.e., not for cross-national research).

- **Canada's Youth in Transition Survey (YITS).** The survey population comprised 15-year-olds born in 1984. These students took the PISA in 2000, then YITS followed up with the same individuals every 2 years. In 2000, the original student sample size was 38,000; by 2010, it had shrunk to 11,011 (OECD, 2010).
- **Denmark's PISA Longitudinal.** Like Canada, Denmark extended the PISA 2000 assessment. The initial sample was about 4,000 young people born in 1984; by 2004, they were able to interview approximately 3,100 of them (Mejding & Roe, 2006).
- **Switzerland's Transitions from Education to Employment (TREE) survey** also followed the PISA 2000 sample. They started with approximately 6,000 15-year-olds who had taken the assessment and surveyed the same students every year until they were 23 years old, and again at ages 26 and 30. By 2014, they had managed to follow about 4,000 students from the original sample because their response rates were very good: 87 percent for the first survey, and 71 percent in 2014 when subjects were 30 years old (TREE, 2016).
- **Longitudinal Survey of Australian Youth (LSAY).** Australia is following cohorts that took PISA in 1995, 1998, 2003, 2006, 2009, and 2015. Participants are interviewed by phone each year until they are 25 years old. This is a huge undertaking with a sample size of about 10,000 for each cohort. By the end, the sample is about one-third that size. Several studies using these data are available (LSAY, n.d.).

Some researchers have tried to take these country-specific longitudinal datasets and combine them for cross-national comparisons. Of note, all of these studies are descriptive and do not attempt to make causal inferences. John Jerrim of University College, London, has led a number of these by taking national data from several non-ILSAs in four English-speaking countries and linking them with PISA. Studies based on these data have looked at the chances of low–supplemental educational services (SES) students entering college (Jerrim & Vignoles, 2015), entering selective universities (Jerrim, Chmielewski, & Parker, 2015), and the advantages of private secondary school attendance on educational and occupational attainment (Jerrim et al., 2016).

METHODOLOGICAL CHALLENGES

In Chapter 4 we raised some methodological concerns related to the design of ILSAs. Here, we focus on the statistical issues inherent in ILSA analyses that, of course, are intertwined with those design concerns. We

caution readers that many of these concerns are rooted in complex technical issues that we do not describe here. Interested readers are encouraged to consult Chmielewski and Dhuey (2017) for more detail.

- **Cross-cultural equivalence:** When researchers argue that their analysis may support causal claims, they must explicitly or implicitly assume that:
  - The research question and research design are equally valid in all countries;
  - All measures have equivalent meanings across all countries;
  - The fidelity of implementation is the same across countries; and
  - The treatment (i.e., the policy) itself must effectively be the same in all countries.

These are clearly strong assumptions that many critics argue can never be met, even if the researchers are well informed about the policy context in each country studied. In addition, as discussed in Chapter 4, care must be taken with background questionnaires and derived scales to ensure they are measuring the same items. Different countries have different meanings for seemingly objective terms such as "private school," and cultural differences have been noted in responses to Likert Scale questions as well as self-ratings of personal attributes. David Kaplan noted that, to be fair, all ILSAs take great pains to examine cross-cultural equivalence. Moreover, ILSA expert groups are heavily involved in basic research designed to improve methodologies for establishing cross-cultural equivalence.

- **Measurement error in estimating achievement:** As described in Chapter 4, ILSAs diminish the amount of time that individual students are tested by using booklets created through a matrix sampling design. This means that no student has scores on the entire set of items, but instead is assigned a set of plausible values (see Chapter 4). Although efficient for the assessments themselves, this methodology causes problems for analyses of individual student data (e.g., see Braun & von Davier, 2018, and Chmielewski & Dhuey, 2017).
- **Weighting:** ILSAs test samples of students, not the entire group of individuals in a target population. As is common practice in sampling, some groups of students are oversampled to ensure more accurate estimates at the group level. ILSA datasets include sample weights that "correct" for any oversampling; these weights add complexity and noise to ILSA analyses that researchers

need to account for when analyzing data both within and across countries.

- **Combining different assessments:** A number of researchers—for example, Hanushek and Wößmann (2006) as described in Box 5-4—have used various linking techniques to combine scores across ILSAs. The fact that TIMSS, PIRLS, and PISA have international mean scores of 500 and similar standard deviations does not mean that scores are equivalent. As outlined in Table 1-1, the tests are designed to assess knowledge and skills in different areas. For example, TIMSS is curriculum based, attempting to reflect what is actually taught in classrooms, while, according to Chmielewski and Dhuey (2017), "PISA assessments in these subjects depend less on formal knowledge of laws and formulas and more on application of competencies to real-world situations" (p. 20). When analyzing long-term trends in countries' performance, one must be cautious about comparing different ILSAs, or even earlier and later versions of the same ILSA. Braun and von Davier (2018) provide further discussion of this issue.
- **Who is actually assessed:** ILSA analyses may be complicated by differences across countries or across different ILSAs in the student populations from which samples are drawn. Across countries, the organizations that administer ILSAs have attempted to standardize policies as to who should be tested, but the reality is that complete standardization is unattainable. For example, participating countries are allowed to exclude students for many reasons, such as disability or lack of mastery of the tested language. Total exclusions may not exceed 5 percent of the student population. In studies combining different ILSAs, differences in the student populations tested are even more problematic. One important example is that PISA and the Programme for the International Assessment of Adult Competencies (PIAAC) use age-based samples, while TIMSS and PIRLS sample by grade. Some nations, like Japan, strictly enforce age cutoff dates and discourage grade retention, so there should be little difference between an age-based versus a grade-based sample. But most countries do not have a one-to-one correspondence between age and grade. Synthetic cohort analyses, of the type in Category 4 described above, require strict comparability of student populations in different ILSAs because repeated cross-sections must be drawn from the same population. However, students exiting school and immigrants must be accounted for in these types of studies. If immigrants are excluded, synthetic cohort data are more comparable, but on the other hand, the findings are less generalizable

to the whole population, as immigrants were educated within the country's educational system.

- **Changes in background questionnaire wording:** The wording of ILSA background questions have subtly changed over time, and sometimes the questions differ not only across countries but within countries. For example, the PISA 2003 background questionnaire asked how old the student was when he or she started his or her primary education (typically the first year after kindergarten) using the sentence How old were you when you entered elementary school? In 2009, with the obvious goal of improving clarity, the question was changed to How old were you when you entered first grade? In 2003, the mean response was 5.4 years old, but in 2009, it was 5.9; this is not surprising given that most U.S. elementary schools also include kindergarten classes.

ILSA questionnaires also include a number of adjustments to allow for differences among countries, such as in how grade levels are described. Post-secondary education takes a number of forms in different countries, and it is difficult to distinguish cross-nationally among vocational and technical certificates, community college, 2-year training programs, and associate's degrees, among others. "Careful researchers must be sure to compare questionnaires across years, surveys, and natural adaptations across countries to avoid making inappropriate comparisons" (Chmielewski & Dhuey, 2017, p. 24).

WHAT MIGHT BE POSSIBLE WITH LONGITUDINAL ANALYSES

Jan-Eric Gustafsson of the University of Gothenberg, Sweden, raised further questions about the possibility of causal inference, identifying this as the main limitation of ILSAs as currently designed and administered. There are "billions" of ways nations differ, and it is extremely difficult for researchers to account for all of them when trying to isolate factors affecting student achievement. Heterogeneity across countries makes it difficult to make causal inferences about the determinants of achievement from cross-sectional data.

With only cross-sectional data, we inevitably run into problems of biased estimates of causal effects. Gustafsson quipped, "Correlation is not causation, but it may be a hint, someone said, and that may be true, but it may also hint you in the wrong direction." Gustafsson highlighted three main problems:

- **Omitted variables:** All those factors that should be in researchers' models but are not; these are also referred to as "unobserved

heterogeneity," meaning that there are factors that we cannot or have not measured, or are not even aware of, that may differ systematically across countries and have an impact on the outcomes of interest (e.g., test performance).

- **Reverse causation:** Searching for causation can get things backward. For example, educational resource allocation is often compensatory. Lower-performing students may get more resources directed at them. Thus, if we find that greater resources of a particular type are correlated with lower performance, might we infer incorrectly that greater resources lead to lower achievement?
- **Errors of measurement:** Random error in observed variables used as predictors in regression is another major concern, typically causing estimates of relationships to be attenuated.

According to Gustafsson (but not all workshop participants agreed), the best available approach given the current constraints in the way ILSAs are designed and administered may be to use within-country change over time to identify potential causal relationships between putative determinants and educational outcomes (Category 3). What specific policies or attributes within a country worked to generate higher levels of student achievement? The single-country synthetic longitudinal approach (it is longitudinal in a sense that data trends are estimated over time, but it is synthetic because individual students are not tracked and tested through the grades) might address some of the shortcomings mentioned above. With synthetic longitudinal data, Gustafsson argues that countries can be their own controls, thereby removing estimation bias from omitted variables—at least those that are fixed characteristics of the countries for the time period of interest.

Gustafsson provided a simple example of this type of synthetic longitudinal approach. Educators and researchers know that there is a positive correlation between student age and achievement: older students are better readers. The PIRLS reading test is administered to students between the ages of 9.5 and 11 years, depending on the country. We would expect that countries that test at age 11 have higher scores. So, are scores higher for nations that test later? Gustafsson found that no, they are not. There were actually negative correlations for both 2001 and 2006; nations that tested older students did not have higher scores. Thus, when looking across nations, the data "[do] not really support the hypothesis of the positive correlation between age and reading achievement," which is counterintuitive.

However, a synthetic longitudinal within-country analysis yielded different results. The mean age of students tested differs somewhat at the country level at each administration. For example, among Russian

students tested in 2001, the mean age was approximately 10.25 years; among those tested in 2006, the mean age was approximately 10.75 years. As expected, the mean score for Russian students in reading went up between 2001 and 2006. Gustafsson estimated a correlation of 0.53 between age change and reading score change between the two testing occasions.

Why did the two approaches produce such different results? It is because the second method held constant all time-invariant omitted variables at the country level; countries were measured against themselves. The first approach did not adjust for the myriad variables that affect achievement across countries, hence the relationship between age and reading ability was obscured. Because this age and readability relationship was found not only in Russia but in a large number of other nations, the researchers argue that there is a causal relationship between age and reading ability, and one can make such an inference without having to track individual students. Gustafsson applied a similar method to looking at the relationship between the amount of time spent on homework and math achievement, and also found a positive correlation.

Despite the apparent promise of this approach, Judith Singer of Harvard University urged extreme caution when interpreting the results of synthetic longitudinal analyses conducted using aggregate data. As explained earlier in the section on ecological correlations, analyses conducted using individual-level data will not necessarily agree with analyses conducted using aggregate data. Singer went on to argue that ILSA data should be thought of as a complex, multilevel data structure, with countries at the highest level of aggregation, and the within-country multilevel data structures at lower levels. Proper analysis of multilevel data—whether cross-sectional or longitudinal—requires fitting statistical models that reflect this multilevel structure, with variables measured at the appropriate level of aggregation. Although discussion of these technical issues is beyond the scope of this report, the interested reader is referred to Singer (1998) and Singer and Willett (2003).

MORE WORK TO BE DONE ON METHODOLOGY

Are there settings in which any of the analyses described above offer credible evidence to support a causal claim? Certainly, some of the researchers whose work we cite do argue that their work "comes close" to supporting causal claims. But others argue that with very few exceptions, the threats to internal validity are just too large and nothing in the ILSA literature supports such strong conclusions. Judith Torney-Purta, in remarking on this body of research, stated that although no single approach is perfect, some rely on stronger assumptions than others. Researchers should carefully consider how methodological decisions

concerning such issues 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.

In the meantime, more methodological research is needed to clarify—and then judge—the relative advantages and disadvantages of the different analytical approaches. There are important roles to be played both by academic researchers and sponsoring organizations, such as the IEA, the Organisation for Economic Co-operation and Development (OECD), and their contractors, and by national governments. Importantly, IEA and OECD could enhance their roles in disseminating technical user guides, software programs, and macros that support methodological best practices for the increasingly complex analyses carried out by education policy researchers using ILSA data.

6

Summary and Key Messages

The National Academy of Education (NAEd) initiated this project to examine future directions for international large-scale assessments (ILSAs) from a variety of disciplinary perspectives. This report summarizes two related workshops (along with the commissioned work) organized to address this issue. The first workshop focused on methodological issues related to the design, analysis, and reporting of ILSAs, and the second workshop examined less-technical aspects of reporting, interpretation, and policy uses.[1]

Attendees at both sessions included specialists in educational policy, journalism, research design, and statistical analysis. Some participants had experience with ILSAs dating back to the 1970s, while others were relative newcomers to these debates. There were also individuals who fund studies, others who plan and conduct sampling and statistical analyses, and others who interpret ILSA results from the perspective of policy, economics, and sociology. Individuals with expertise in cross-national studies of health and aging, as well as in early childhood longitudinal studies, gave presentations that suggested new approaches and perspectives. Participants who addressed methodological topics were successful in making their presentations understandable by a general audience of educational researchers.

[1] Workshop materials, including agendas, videos, and commissioned papers, are available on the project website: https://naeducation.org/methods-and-policy-uses-of-international-large-scale-assessments.

Workshop participants agreed that ILSAs provide a great deal of useful information, but there was spirited debate and disagreement about what types of analyses are the most meaningful and what could be done to assure more sound interpretations. The goal of this project was not to reach consensus on these issues, but rather to highlight some of the strengths, limitations, and complexities of ILSAs, especially as a basis for informing future educational policy and practice. The workshops also considered the use of ILSA data for the improvement of policy and increased public awareness of the features of high-quality education.

A major theme that reappeared throughout the discussions was the need to be clear about the research questions we are trying to address with ILSA data. The research questions we want to answer and the research questions we can answer are often not the same. Some researchers have hope about what they would like to do with ILSAs in relation to their own research interests, as opposed to proposing the kinds of questions ILSAs, as currently designed, can realistically be expected to answer. Similarly, reporting on ILSAs also demonstrates a substantial gap between what the media and the public want—for example, headlines or basic information on student performance in relation to that in other countries—and what ILSA administrators and researchers want, which is a more nuanced view that focuses on patterns of student achievement and their correlates, with an eye toward informing policy discussions and decisions.

Several key messages emerged from this project and are highlighted below.

PURPOSES OF ILSAs

Different stakeholders—the media, the research community, policy makers, and ILSA administrators—have different goals and interests with regard to ILSAs. As outlined in Chapter 1, the six most important purposes of ILSAs are to:

1. **Describe and compare** student achievement and examine relevant contextual factors across nations (e.g., characteristics of countries, including their educational and social policies, and characteristics of students, including demographics and self-reported background data).
2. **Track changes over time** in student achievement, contextual factors, and their mutual relationships, within and across nations.
3. Disturb complacency about a nation's education system and to **spur educational reforms.**
4. **Create de facto international benchmarking** by identifying top-performing nations and jurisdictions, or those making unusu-

ally large gains, and suggesting ways to learn from this array of practices.
5. **Evaluate** the effectiveness of curricula, instructional strategies, and educational policies, while understanding that many of them are deeply contextualized.
6. **Explore causal relationships** between contextual factors (e.g., demographic, social, economic, and educational variables) and student achievement.

Workshop participants generally agreed that ILSAs can best serve the first three purposes: (1) describe and compare student achievement, (2) track changes over time, and (3) spur educational reforms. However, concerns and disagreements were widespread regarding the use of ILSAs for the last three purposes: (4) create benchmarking, (5) evaluate effectiveness, and (6) explore causal relationships.

The issue of whether causal inferences can be drawn from analyses of ILSA data was a recurring theme of this project. Purposes 4, 5, and 6 were controversial because they aim to use ILSA data to determine why some nations perform better than others; that is, which polices and practices lead to (i.e., cause) better educational outcomes. As discussed throughout this report, there are numerous methodological challenges that, for many scholars, stand in the way of making credible causal inferences. Most researchers believe causality cannot be firmly established without randomized controlled experiments or rigorous quasi-experiments, which are typically not realistic in educational settings at this scale. Experts also debated the extent to which it is possible to isolate specific factors or policies that contribute to improved student achievement in one nation that can be applied or adapted elsewhere. During the workshops, participants vigorously discussed the steps that would need to be taken to increase the likelihood of being able to make credible causal inferences. Given the large number of factors affecting student achievement and the dynamics among them, as well as the fact that nations are so different from one another with respect to history, culture, and politics, many participants argued that this is not, and possibly will never be, a realistic goal for ILSAs as presently designed. However, others made the case that it is possible in some instances to come close to identifying causal relationships using creative research designs, careful analyses, and plausible assumptions.

INTERPRETATION AND REPORTING

- **Media reporting of ILSA results tends to be superficial and, in many cases, misleading.**

ILSA results receive a good deal of media attention. In the United States, it is common to see alarming headlines about how the United States ranks poorly compared with other nations. But most ILSA researchers agree that press reports omit too much critical background information and fail to explain the results of more nuanced analyses. Consequently, a main concern discussed at the workshops was how to encourage and support the media to do a better job, not only of more thoroughly reporting results, but also presenting balanced interpretations of the findings.

Numerous differences in educational policies and practices usually go unmentioned when reporting ILSA results. Nations differ in terms of demographics, wealth, culture, beliefs about the value of education, the status of the teaching profession, and many other relevant factors. Knowing something about these contextual factors in all nations being compared should influence both score interpretation and policy implications.

The situation is exacerbated by declining revenue streams that make it difficult for news outlets to devote time and space to education stories. As a result, the public is presented with incomplete information about educational systems in other countries, which leads, as one participant noted, to a lot of "generalizing, hand waving, and anecdotal information" masquerading as hard evidence.

- **The organizations that administer ILSAs and release results would be wise to devote greater resources to preparing reporters and providing more guidance on what can and cannot be inferred from results.**

To improve media reporting, as highlighted above, one point of agreement was that it is impossible to expect strong reporting from the fast-paced cable news media and its short story cycles. Education researchers and ILSA administrators must work harder at presenting and interpreting results for the press and the public at large. Additionally, reporters can be provided with deeper and more advanced training on social sciences methodologies and background information on how results can be interpreted by a group such as the Education Writers Association.

- **An impartial, national board could be created and charged with providing ongoing guidance on ILSA design, analysis, reporting, and interpretation.**

Such a board could work with the Institute of Education Sciences and its National Center for Education Statistics. The Board of International Comparative Studies in Education (BICSE), which existed at the National Research Council when some of the early ILSAs were being planned, may

provide a model. Norman Bradburn of The University of Chicago, who chaired that committee, noted its value to the field. Jack Jennings, retired president and CEO of the Center on Education Policy, similarly proposed that there be a national report released each year by an independent and respected research group that summarizes what can be concluded from ILSAs about American education. It could explain the differences among the various ILSAs and interpret, with appropriate caveats, some of the findings.

- **Reporting results primarily at the country level obscures variation within countries that are often overlooked. States and other smaller jurisdictions often have different social and economic characteristics and varying educational policies.**

Most ILSA reports present only country-level statistics, but most countries are divided into smaller subunits, whether those subunits are locally responsible for education or not. Many participants argued against viewing nations as monolithic entities and were in favor of breaking down the data into smaller units of analysis. The United States is the extreme case with 50 states and 14,000 school districts. Where possible, scholars should explore within-country studies of characteristics of national and sub-national units of educational systems, rather than emphasizing high-level, cross-national comparisons.

Some workshop participants argued that much more could be learned from high-performing sub-jurisdictions (e.g., states and school districts) within one's own country rather than other countries that, as a whole, perform exceptionally well. Explanatory studies based on state differences or district differences would be more likely to identify policy ideas that can be better understood in the local context and be more palatable to citizens and politicians.

POLICY USES AND LIMITATIONS

- **ILSAs can provide useful information to make policy inferences and spur educational reforms only if carefully conducted analyses support these conclusions.**

Some workshop participants took the view that ILSAs—even if they cannot be used for causal inference—can have a positive impact on educational policy. Although there was disagreement about the overall utility of ILSAs for policy making, participants argued that the results can, and have been, used to spur educational reform. They can also be used to avoid complacency by highlighting differences between a focal coun-

try—be it Germany, Sweden, the United States, or any country—and other comparable nations. For example, the head of one prominent U.S. educational reform organization stated that advocacy organizations often use ILSA results to pressure policy makers. This happens primarily at the federal level but could also be effective at the state level (i.e., if sampling plans were adjusted and analyses could be conducted on data disaggregated into smaller subunits). In addition, secondary analyses of these rich datasets, often published in refereed journals, are an untapped reservoir of potentially relevant findings for policy and practice. Caution should be exercised, however, to ensure that results are not misinterpreted and do not lead to poor policies or a misallocation of resources.

- **Longitudinal studies employed in other fields serve as promising examples for making policy relevant inferences.**

Policy relevant, cross-national research in fields other than education can serve as examples of longitudinal studies that have impacted policy, particularly in the areas of health and human development. It was possible to draw policy relevant inferences because the researchers used strong quasi-experimental designs or truly longitudinal study designs that tracked individuals over time, yielding credible evidence of causal linkages. One question discussed at the workshop was whether such evidence suggests that it might be possible for some ILSAs to be redesigned with a longitudinal component.

DESIGN AND ANALYSIS

- **The issue of whether causal inferences can be drawn from analyses of ILSA data was a recurring and controversial theme of this workshop series and remains an important area for continued research and development.**

Education researchers disagree on whether causal inferences can be drawn from any designs other than randomized controlled trials (RCTs) and strong quasi-experimental designs. These differences of opinion are not confined to ILSAs, but they arose repeatedly at the workshops because there is such a strong temptation, due in part to political pressures, to draw causal inferences from ILSAs. At this time, perhaps the best approach for scholars using ILSAs to demonstrate causation is to present why they believe their particular analysis provides credible evidence that supports a causal claim, while simultaneously investigating all the reasons why critics would argue that it does not.

Although there was general agreement that truly longitudinal designs

would yield more useful data, there was considerable disagreement about the feasibility of tracking large numbers of students as they progress through school. There has been only one truly longitudinal ILSA effort conducted across a full array of countries. This was the Second International Mathematics Study (SIMS) of the early 1980s, which tested the same students in seventh and eighth grades. However, many countries do conduct truly longitudinal studies within their own jurisdictions. Some of these are based on original ILSA samples (Australia, Canada, Denmark, and Switzerland have longitudinally followed PISA participants), but more are homegrown longitudinal studies that were never intended to have any international component (e.g., the Early Childhood Longitudinal Study and the Education Longitudinal Study in the United States and the National Educational Panel Study in Germany). As explained in Chapters 4 and 5, attrition is a major concern for these efforts and, in some cases, the attrition is so severe as to raise a legitimate question as to whether the gain in credible information is worth the additional effort, time, and money. But with increased availability of online data collection, the possibility of truly longitudinal designs offers promise worth considering in the years ahead.

- **ILSA data would be more useful and accurate if questionnaire data (especially measures of socioeconomic status) are augmented with data from other sources, such as U.S. Census data or administrative data collected by school districts.**

Future ILSA designs should strive toward better measures of key background variables. For instance, in economically well-developed countries where census-type data are routinely collected, reliable measures related to school district supplemental educational services (SES) can be derived using sophisticated approaches such as the U.S. Census Small Area Income and Poverty Estimates program. These sorts of measures are more accurate and more fine-grained than anything currently available in international assessments.

- **Experts on computer-based assessment believe that it is inevitable that ILSAs will continue to move to a computer-based platform, which presents an opportunity for creative rethinking of ILSA designs over the long term.**

The benefits of digitally based assessments (DBAs) would come in the areas of improved measurement, ease of administration, and data processing. DBA can improve the accuracy of assessment through adaptive testing, whereby test questions are tailored to a student's level of profi-

ciency. In the near term, DBA is not likely to have a major impact on how ILSAs are used in educational policy making or in research; most of the improvements will be behind the scenes. However, the widespread introduction of DBA does present an opportunity for some creative rethinking of ILSA designs over the long term.

- **Emerging analytical approaches for the analysis of ILSA data that "come close" to supporting causal inferences may offer promising potential for providing useful policy guidance.**

RCTs are not currently feasible with ILSA designs, and the assessments largely do not track individual students over time. Nor do the assessments provide results for individual students at one point in time. This approach limits the ability to draw causal inferences from ILSA data. Some critics go so far as to assert that causal inference is simply not possible. Others argue that analyses can "come close" to supporting causal claims, despite dependence on cross-sectional observational data.

As outlined in Chapter 5, researchers have used five types of analytical approaches:

- Analyze policy variation across countries.
- Analyze policy variation within countries.
- Analyze repeated cross-sectional ILSA data to look at variation across birth cohorts or generations of students.
- Analyze repeated cross-sectional ILSA data to look at variation across age within the same birth cohort within countries; known as "synthetic cohorts."
- Analyze rare truly longitudinal ILSA data that follow the same students over time.

Scholars using these methods have looked at topics ranging from how being the youngest student in class impacts future academic outcomes, to the long-term impact of "tracking" students, to the relationship between computers at home and reading proficiency.

In addition to the aforementioned issues concerning causal inference, there are a number of other methodological issues that arise in discussions of research with ILSA data including reverse causation, omitted variables, weighting, cross-cultural equivalence, combining different assessments, quality of assessment items, measurement error, differences across countries in tested populations of students, and changes in the wording of background questions over time. By turning these dilemmas into research opportunities, these issues all constitute potential directions for future methodological work.

Ultimately, additional research is needed to weigh the relative advantages of the different methodological approaches in ILSA studies. To address these methodological issues, there are important roles to be played by academic researchers, as well as by sponsoring organizations such as the International Association for the Evaluation of Educational Achievement (IEA) and the Organisation of Economic Co-operation and Development (OECD) and their contractors, and by national governments and those conducting secondary analyses after the data have been released. IEA and OECD should enhance their roles in disseminating technical user guides, software programs, and macros that support methodological best practices for the increasingly complex analyses carried out by researchers using ILSA data.

THE FUTURE OF ILSAs: A FINAL NOTE

Despite the many issues raised during the workshops and described in this volume, ILSAs are here to stay. Indeed, not only are they here to stay, they are likely to become even more salient to educational policy discussions as the world becomes increasingly globalized. For this to be a good outcome, technical issues must be addressed and policy makers, the press, and the public must be more aware of the data's limitations. The technical and other issues discussed at the workshops and summarized in this report will become increasingly important and merit a substantial investment in further research—both theoretical and empirical. Thus, we hope that this report will serve as a springboard for greater attention from the broader research community and other stakeholders. We also hope that this volume will be useful to those who report on educational policy, whether on blogs or in the major media. The press and the public would benefit from a deeper understanding of how ILSAs work, what their shortcomings are, and how to interpret results. We hope this report is a step in that direction.

References

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.

Bradbury, B., Corak, M., Waldfogel, J., & Washbrook, E. (2015). *Too many children left behind: The U.S. achievement gap in comparative perspective.* New York: Russell Sage Foundation.

Braun, H., & von Davier, M. (2018). The use of test scores from large-scale assessment surveys: Psychometric and statistical considerations, *Large-Scale Assessments in Education*, *5*(1).

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.

Butler, P. (2016, September 20). No grammar schools, lots of play: The secrets of Europe's top education system. *The Guardian*. Retrieved from https://www.theguardian.com/education/2016/sep/20/grammar-schools-play-europe-top-education-system-finland-daycare.

Carey, K. (2014, June 28). America thinks we have the world's best colleges. We don't. *The New York Times*. Retrieved from https://www.nytimes.com/2014/06/29/upshot/americans-think-we-have-the-worlds-best-colleges-we-dont.html?_r=0.

Carnoy, M., & Rothstein, R. (2013). What do international tests really show about U.S. student performance? *Economic Policy Institute*. Retrieved from http://www.epi.org/publication/us-student-performance-testing.

Chappell, B. (2013, December 3). U.S. students slide in global ranking on math, reading, science. *National Public Radio*. Retrieved from http://www.npr.org/sections/thetwo-way/2013/12/03/248329823/u-s-high-school-students-slide-in-math-reading-science.

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., & Dhuey, K. (2017). *The analysis of international large-scale assessments to address causal questions in education policy.* Paper commissioned by the National Academy of Education.

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.

Dillon, S. (2010, December 7). Top test scores from Shanghai stun educators. *The New York Times.* Retrieved from http://www.nytimes.com/2010/12/07/education/07education.html?mcubz=1.

Dobbins, M., & Martens, K. (2012). Towards an education approach à la Finlandaise? French education policy after PISA. *Journal of Education Policy*, *27*(1), 23–43. http://doi.org/10.1080/02680939.2011.622413.

Duncan, A. (2013, March 12). The threat of educational stagnation and complacency. U.S. Department of Education. Retrieved from http://www.ed.gov/news/speeches/threat-educational-stagnation-and-complacency.

Egelund, N. (2008). The value of international comparative studies of achievement: A Danish perspective. *Assessment in Education: Principles, Policy & Practice*, *15*(3), 245–251. http://doi.org/10.1080/09695940802417400.

Ertl, H. (2006). Educational standards and the changing discourse on education: The reception and consequences of the PISA study in Germany. *Oxford Review of Education*, *32*(5), 619–634. http://doi.org/10.1080/03054980600976320.

Finn, C. E., Jr. (2010, December 8). A Sputnik moment for U.S. education. *The Wall Street Journal*. Retrieved from http://online.wsj.com/article/SB10001424052748704156304576003871654183998.html.

Ginsburg, A., & Smith, M. S. (2016). *Do randomized controlled trials meet the "gold standard"? A study of the usefulness of RCTs in the What Works Clearinghouse*. Washington, DC: American Enterprise Institue.

Grek, S. (2009). Governing by numbers: The PISA "effect" in Europe. *Journal of Education Policy*, *24*(1), 23–37.

Hanushek, E. A, & Wößmann, L. (2006). *Does educational tracking effect performance and inequality? Differences-in-differences evidence across countries*. Working Paper 11124. Cambridge, MA: National Bureau of Economic Research. nber.org/papers/w1124.

Hanushek, E. A., & Wößmann, L. (2010). The economics of international differences in educational achievement. *Handbook of the Economics of Education, 3*, 89–200. The Netherlands: North-Holland.

Heim, J. (2016, December 8). Finland's schools were once the envy of the world. Now, they're slipping. *The Washington Post*. Retrieved from https://www.washingtonpost.com/local/education/finlands-schools-were-once-the-envy-of-the-world-now-theyre-slipping/2016/12/08/dcfd0f56-bd60-11e6-91ee-1adddfe36cbe_story.html?utm_term=.f8379258716e.

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.

Loveless, T. (2013, October 9). *PISA's China problem.* The Brookings Institution. Retrieved from https://www.brookings.edu/research/pisas-china-problem.

LSAY (Longitudinal Surveys of Australian Youth). (n.d.). *Briefing papers.* Retrieved from https://www.lsay.edu.au/publications/briefing-papers.

Marks, G. N. (2005). Cross-national differences and accounting for social class inequalities in education. *International Sociology, 20*(4), 483–505.

Mejding, J., & Roe, A., (Eds.). (2006). *Northern lights on PISA 2003. A reflection from Nordic countries.* Copenhagen: Nordic Council of Ministers.

Mullis, I. V. S., Martin, M. O., Ruddock, G. J., O'Sullivan, C. Y., & Preuschoff, C. (2009). *TIMSS 2011 assessment frameworks*. Boston, MA: TIMSS & PIRLS International Study Center, Lynch School of Education, Boston College. Retrieved from http://www.eric.ed.gov/ERICWebPortal/detail?accno=ED512411.

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

NASEM (National Academies of Sciences, Engineering, and Medicine). (2017). *Communicating science effectively: A research agenda.* Washington, DC: The National Academies Press. https://doi.org/10.17226/23674.

NCSL (National Conference of State Legislatures). (2016). *No time to lose: How to build a world-class education system state by state.* Washington, DC: NCSL. Retrieved from http://www.ncsl.org/documents/educ/EDU_International_final_v3.pdf.

OECD (Organisation for Economic Co-operation and Development). (2010). *Pathways to success: How knowledge and skills at age 15 can shape future lives in Canada.* Retrieved from http://www.oecd-ilibrary.org/education/pathways-to-success/introduction-the-case-for-linking-pisa-with-longitudinal-studies_9789264081925-2-en.

OECD. (2003, January 30). PISA Index of Economic, Social and Cultural Status (ESCS). *OECD.* Retrieved from https://stats.oecd.org/glossary/detail.asp?ID=5401.

Partanen, A. (2011, December 29). What Americans keep ignoring about Finland's school success. *The Atlantic*. Retrieved from http://www.theatlantic.com/national/archive/2011/12/what-americans-keep-ignoring-about-finlands-school-success/250564.

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

Reardon, S., Kalogrides, D., & Ho, A. (2016). *Linking U.S. school district test score distributions to a common scale, 2009–2013*. CEPA Working Paper No. 16-09. Stanford Center for Education Policy Analysis. Retrieved from https://cepa.stanford.edu/sites/default/files/wp16-09-v201604.pdf.

Rohwedder, S., & Willis, R. J. (2010). Mental retirement. *Journal of Economic Perspectives*, 24(1):119–138. Retrieved from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2958696.

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. Retrieved from https://largescaleassessmentsineducation.springeropen.com/articles/10.1186/s40536-016-0020-8.

Rosero-Bixby, L., & Dow, W. H. (2016). Exploring why Costa Rica outperforms the United States in life expectancy: A tale of two inequality gradients. *Proceedings of the National Academy of Sciences of the United States of America*, *113*(5):1130–1137. Retrieved from https://www.ncbi.nlm.nih.gov/pubmed/26729886.

Rutkowski, L. (2016) Introduction to special issue on quasi-causal methods. *Large-scale Assessment in Education, 4*(8), 1–6.

Rutkowski, L. (2017). *A look at the most pressing design issues in international large-scale assessments.* Paper commissioned by National Academy of Education.

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.

Singer, J. D. (1998). Using SAS PROC MIXED to fit multilevel models, hierarchical models, and individual growth models. *Journal of Educational and Behavioral Statistics, 24*(4), 322–354.

Singer, J. D., & Braun, H. I. (2018, April 6). Testing international education assessments: Rankings get headlines, but often mislead. *Science, 360*(6384). doi: 10.1126/science.aar4952.

Singer, J. D., & Willett, J. B. (2003). *Applied longitudinal data analysis: Methods for studying change and event occurrence.* New York: Oxford University Press.

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.

TREE (Transitions from Education to Employment). (n.d.). Retrieved from http://www.tree.unibe.ch/index_eng.html.

van de Vijver, F., Chasiotis, A., & Breugelmans, S. M. (2011). Fundamental questions in cross-cultural psychology. In S. M. Breugelmans, A. Chasiotis, & F. J. R. van de Vijver (Eds.), Fundamental questions in cross-cultural psychology (pp. 9–34). Cambridge, UK: Cambridge University Press.

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, UK: Emerald Group Publishing Limited.

Zimmer, R. W., & Toma, E. F. (2000). Peer effects in private and public schools across countries. *Journal of Policy Analysis and Management*, 75–92.

Appendix A

Workshop Agendas and Participants

WORKSHOP I:
DIRECTIONS FOR IMPROVING ILSA DESIGN AND ANALYSIS

National Academy of Sciences Building, Room 125,
2101 Constitution Avenue, NW, Washington, DC
Friday, June 17, 2016

AGENDA

8:15–8:45 a.m.	**Breakfast**
8:45–9:05 a.m.	**Welcome and Project Goals** *Michael J. Feuer*, National Academy of Education and The George Washington University *Judith Singer*, Harvard University *Peggy G. Carr*, National Center for Education Statistics, U.S. Department of Education

Panel 1: Design

9:05–9:10 a.m.	**Introduction by Panel Chair** *Marshall "Mike" Smith*, Carnegie Foundation for the Advancement of Teaching

9:10–9:40 a.m.	**Overview of ILSA Design Issues** *Leslie Rutkowski*, University of Oslo, Norway
9:40–10:00 a.m.	**Implications of Computer-Based Testing** *Henry Braun*, Boston College
10:00–10:30 a.m.	**Discussants** *Ina Mullis*, Boston College *David Kaplan*, University of Wisconsin–Madison
10:30–11:00 a.m.	**Audience Q&A** Led by *Mike Smith*

Panel 2: Voices from Other Fields

11:00–11:05 a.m.	**Introduction by Panel Chair** *Richard Durán,* University of California, Santa Barbara
11:05–11:35 a.m.	**Child Development** *Elizabeth Washbrook*, University of Bristol, United Kingdom
11:35 a.m.–12:05 p.m.	**Aging** *John Haaga*, National Institute on Aging, U.S. Department of Health and Human Services
12:05–12:30 p.m.	**Audience Q&A** Led by *Richard Durán*
12:30–1:15 p.m.	**Lunch**

Panel 3: Analysis

1:15–1:20 p.m.	**Introduction by Panel Chair** *Judith Torney-Purta*, University of Maryland
1:20–1:50 p.m.	**Overview of ILSA Analysis Issues** *Anna Katyn Chmielewski*, University of Toronto *Elizabeth Dhuey*, University of Toronto

1:50–2:10 p.m.	**Longitudinal Analysis and the Potential for Causal Interpretations** *Jan-Eric Gustafsson*, University of Gothenberg, Sweden
2:10–2:40 p.m.	**Discussants** *Daniel Koretz*, Harvard University *Eric Hanushek*, Hoover Institution, Stanford University
2:40–3:10 p.m.	**Audience Q&A** Led by *Judith Torney-Purta*
3:10–3:25 p.m.	**Wrap-Up** *Judith Singer*
3:25–4:00 p.m.	**Reception** (Great Hall)

PARTICIPANTS

Chris Averett, Westat
Norman Bradburn, NORC at The University of Chicago
Henry Braun, Boston College
Peggy Carr, National Center for Education Statistics
Madhabi Chatterji, Teachers College, Columbia University
Anna Katyn Chmielewski, University of Toronto
Naomi Chudowsky, National Academy of Education
Mary Coleman, National Center for Education Statistics
Elizabeth Dhuey, University of Toronto
Richard Durán, University of California, Santa Barbara
Laura Engel, The George Washington University
Ebru Erberber, American Institutes for Research
Michael Feuer, The George Washington University and the National Academy of Education
Joshua Glazer, The George Washington University
Jan-Eric Gustafsson, University of Gothenberg, Sweden
John Haaga, National Institute on Aging
Clarisse Haines, National Center for Education Statistics
Eric Hanushek, Hoover Institution, Stanford University
Robert Hauser, The National Academies of Sciences, Engineering, and Medicine
HyoJung Jang, The Pennsylvania State University
David Kaplan, University of Wisconsin–Madison

Dana Kelly, National Center for Education Statistics
Judith Koenig, The National Academies of Sciences, Engineering, and Medicine
Daniel Koretz, Harvard University
Jorge Ledesma, National Center for Education Statistics
Lydia Malley, National Center for Education Statistics
Maureen McLaughlin, U.S. Department of Education
David Miller, American Institutes for Research
Ina Mullis, Boston College
Ruth Neild, Institute of Education Sciences
Oren Pizmony-Levy, Teachers College, Columbia University
Stephen Provasnik, National Center for Education Statistics
Taslima Rahman, National Center for Education Statistics
Leslie Rutkowski, University of Oslo, Norway
Judith Singer, Harvard University
Marshall "Mike" Smith, Carnegie Foundation for the Advancement of Teaching
Bernhard Streitwieser, The George Washington University
Sheila Thompson, National Center for Education Statistics
Judith Torney-Purta, University of Maryland
Elizabeth Washbrook, University of Bristol, United Kingdom
Katrina Weil, U.S. Department of Education
Gregory White, National Academy of Education
James Williams, The George Washington University

WORKSHOP II: REPORTING, INTERPRETATION, AND POLICY USES

National Academy of Sciences Building, Lecture Room, 2101 Constitution Avenue, NW, Washington, DC
Friday, September 16, 2016

AGENDA

8:15–8:45 a.m. **Breakfast**

8:45–9:05 a.m. **Welcome and Project Goals**
Michael J. Feuer, National Academy of Education and The George Washington University
Peggy G. Carr, National Center for Education Statistics, U.S. Department of Education
Judith Singer, Harvard University

Panel 1: Media Perspectives

9:15–9:35 a.m.	*Nicholas Lemann*, Columbia University
9:35–9:55 a.m.	*Kevin Carey*, New America
9:55–10:15 a.m.	*Brad Wible*, *Science* magazine
10:15–10:45 a.m.	**Audience Discussion**
10:45–11:00 a.m.	**Break**

Voices from Other Fields

11:00–11:20 a.m.	*Ellen Nolte*, London School of Hygiene & Tropical Medicine (via Web)
11:20–11:35 a.m.	**Audience Discussion**

Panel 2: Policy Perspectives

11:35–11:55 a.m.	*Jack Jennings*, Center on Education Policy
11:55 a.m.–12:15 p.m.	*Michele McLaughlin*, Knowledge Alliance
12:15–1:00 p.m.	**Lunch**
1:00–1:20 p.m.	*Marc Tucker*, National Center on Education and the Economy
1:20–1:50 p.m.	**Audience Discussion**

Panel 3: Research Perspectives

1:50–2:10 p.m.	*Sean Reardon,* Stanford University
2:10–2:30 p.m.	*Norman Bradburn,* The University of Chicago
2:30–2:50 p.m.	**Break**
2:50–3:10 p.m.	*Henry Levin,* Columbia University

3:10–3:40 p.m. **Audience Discussion**

3:40–4:00 p.m. **Wrap-Up**
Judith Singer

4:00 p.m. **Adjourn**

PARTICIPANTS

Chris Averett, Westat
Norman Bradburn, NORC at The University of Chicago
Henry Braun, Boston College
William Bushaw, National Assessment Governing Board
Kevin Carey, New America
Peggy Carr, National Center for Education Statistics
Anna "Katyn" Chmielewski, University of Toronto
Naomi Chudowsky, National Academy of Education
Mary Coleman, National Center for Education Statistics
Dian Dong, National Academy of Education
Richard Durán, University of California, Santa Barbara
Ebru Erberber, American Institutes for Research
Michael Feuer, National Academy of Education and The George Washington University
Matthew Frizzell, Center on Education Policy
Robert Hauser, The National Academies of Sciences, Engineering, and Medicine
Jack Jennings, Center on Education Policy
David Kaplan, University of Wisconsin–Madison
Dana Kelly, National Center for Education Statistics
Nicholas Lemann, Columbia University
Henry Levin, Columbia University
Laura LoGerfo, National Assessment Governing Board
Lydia Malley, National Center for Education Statistics
Dan McGrath, National Center for Education Statistics
Michele McLaughlin, Knowledge Alliance
Melissa Menzer, National Endowment for the Arts
David Miller, American Institutes for Research
Ellen Nolte, London School of Hygiene & Tropical Medicine
Oren Pizmony-Levy Drezner, Teachers College, Columbia University
Stephen Provasnik, National Center for Education Statistics
Taslima Rahman, National Center for Education Statistics
Sean Reardon, Stanford University
Scott Sargrad, Center for American Progress

Judith Singer, Harvard University
Marshall "Mike" Smith, Carnegie Foundation for the Advancement of Teaching
Bernhard Streitwieser, The George Washington University
Sheila Thompson, National Center for Education Statistics
Judith Torney-Purta, University of Maryland
Marc Tucker, National Center on Education and the Economy
Elizabeth Washbrook, University of Bristol, United Kingdom
Gregory White, National Academy of Education
Brad Wible, *Science* magazine
Holly Xie, National Center for Education Statistics

Appendix B

Biographical Sketches of Steering Committee Members

Judith D. Singer, Ph.D. (*Chair*), is the senior vice provost for Faculty Development and Diversity and the James Bryant Conant Professor of Education at Harvard University. An internationally renowned statistician and social scientist, Dr. Singer's scholarship focuses on improving the quantitative methods used in social, educational, and behavioral research. Her publications include numerous papers and book chapters, as well as three co-authored books: *By Design: Planning Better Research in Higher Education*, *Who Will Teach: Policies that Matter*, and *Applied Longitudinal Data Analysis: Modeling Change and Event Occurrence.* Dr. Singer is a member of the National Academy of Education, a fellow of the American Statistical Association, and a fellow of the American Educational Research Association. She has also been honored with a fellowship at the Center for Advanced Study in the Behavioral Sciences. In 2012, her nomination by President Obama to serve as a member of the Board of Directors of the National Board of Education Sciences was confirmed by the U.S. Senate. She received her B.A. in mathematics from the State University of New York at Albany in 1976 and her Ph.D. in statistics from Harvard University in 1983.

Henry I. Braun, M.S., Ph.D., holds the Boisi Chair in Education and Public Policy in the Lynch School of Education at Boston College. He also serves as a distinguished presidential appointee (retired) at Educational Testing Service (ETS) in Princeton, New Jersey. Dr. Braun joined ETS in 1979 as a research scientist in the Division of Measurement, Statistics,

and Data Analysis Research. In 1990, he was named vice president of research management and was responsible for a staff of more than 200 and a budget of $25 million. In 1999, Dr. Braun stepped down from his role as an officer to become an ETS distinguished presidential appointee and became more involved in education policy issues. Among the more recent reports Dr. Braun authored or co-authored are *Reconsidering the Impact of High-Stakes Testing* (2003); *Using Student Progress to Evaluate Teachers: A Primer on Value-Added Models* (2005); *A Portrait of Advanced Placement Teachers' Practices* (2005); and *An Introduction to the Measurement of Change Problem* (2005). He was elected as a fellow of the American Statistical Association in 1991 and is a member of the National Academy of Education. He is also a co-recipient of the 1986 Palmer O. Johnson Award from the American Educational Research Association and a co-recipient of the National Council for Measurement in Education's 1999 Award for Outstanding Technical Contribution to the Field of Educational Measurement. Prior to his work at ETS, Dr. Braun was a faculty member in the Department of Statistics and the Office of Population Research at Princeton University. Dr. Braun earned his bachelor's degree in mathematics from McGill University. He received his master's degree and doctorate, both in mathematical statistics, from Stanford University.

Anna Katyn Chmielewski, Ph.D., is an assistant professor in the Department of Leadership, Higher, and Adult Education at the University of Toronto. Her research examines macro-level trends in educational inequality, both cross-nationally and over time; specifically, socio-economic disparities in academic achievement, school segregation, curricular streaming/tracking/ability grouping, and university access, as well as the consequences of childhood inequality for adult skills, educational attainment, and income. She uses a sociological lens and quantitative methods, including multi-level modeling and methods for measuring segregation and achievement gaps. Much of her research draws on data from international large-scale assessments (ILSAs), such as the Programme for International Student Assessment (PISA), the Trends in International Mathematics and Science Study (TIMSS), and the Programme for the International Assessment of Adult Competencies (PIAAC). She received her Ph.D. in education from Stanford University in 2012.

Richard Durán, Ph.D., is a professor in the Department of Education at the University of California, Santa Barbara. After obtaining his Ph.D. in psychology from the University of California, Berkeley, in 1977, he worked at Educational Testing Service (ETS) in Princeton, New Jersey, conducting investigations and publishing research findings on the validity of the SAT, the GRE, and the TOEFL tests. As a result, he developed

a strong interest in how more effective instruction could be designed to assist academic outcomes for culturally and linguistically diverse students who do not perform well on standardized tests and who come from low-income families. His recent research has investigated how classroom interaction leads to the construction of learning expertise, how teachers design and implement constructivist learning activities for students, and how students' self-awareness of their performance leads to new notions of assessment. He is also pursuing research on student learning in after-school computer club settings and is working with immigrant parents to help them acquire knowledge of how to use computers and how to work with their children on research and publication projects. He served on the Board on Testing and Assessment of the National Academy of Sciences from 1996 to 2000. From 2012 to 2014, he served on the English Language Learner Advisory Committee for the Smarter Balanced Assessment.

David Kaplan, Ph.D., is the Patricia Busk Professor of Quantitative Methods in the Department of Educational Psychology at the University of Wisconsin–Madison. Dr. Kaplan holds affiliate appointments in the University of Wisconsin's Department of Population Health Sciences and the Center for Demography and Ecology, and is also an Honorary Research Fellow in the Department of Education at the University of Oxford. Dr. Kaplan's program of research focuses on the development of Bayesian statistical methods for education research. He is actively involved in the Organisation for Economic Co-operation and Development's (OECD's) Programme for International Student Assessment (PISA). He served on its Technical Advisory Group from 2005 to 2009 and its Questionnaire Expert Group from 2004 to present; he is now chair of the Questionnaire Expert Group for PISA 2015. Dr. Kaplan also sits on the Design and Analysis Committee and the Questionnaire Standing Committee for the National Assessment of Educational Progress (NAEP). He is a member of the National Academy of Education, a recipient of the Humboldt Research Award, a fellow of the American Psychological Association, and was a Jeanne Griffith Fellow at the National Center for Education Statistics. Dr. Kaplan received his Ph.D. in education from the University of California, Los Angeles, in 1987.

Marshall "Mike" Smith, Ed.D., is a Senior Fellow at the Carnegie Foundation for the Advancement of Teaching. He recently retired from the federal government, having served for 17 months as a senior counselor to the U.S. Secretary of Education in the Obama administration. From 2001 to 2008, he was the program director for education at the William and Flora Hewlett Foundation in Menlo Park, California. He served as acting deputy secretary and undersecretary for 7 years in the U.S. Depart-

ment of Education during the Clinton administration. During the Carter administration, he was chief of staff to the U.S. Secretary of Education and assistant commissioner for policy studies in the U.S. Office of Education. During the Ford administration, he was the associate director at the National Institute of Education. When not in government, he was an associate professor at Harvard University, a professor at the University of Wisconsin–Madison, and a professor at Stanford University. He also served as dean of the School of Education at Stanford University. He is a member of the National Academy of Education.

Judith Torney-Purta, Ph.D., is a developmental and educational psychologist who is Professor Emerita of Human Development at the University of Maryland. She has conducted interdisciplinary research for nearly 50 years on young people's knowledge of democracy and on the social and political attitudes necessary to maintain it. The International Association for the Evaluation of Educational Achievement (IEA) appointed Dr. Torney-Purta as the international chair of the Steering Committee for its landmark IEA Civic Education Study. Over a 10-year period with colleagues from 30 countries, she led a rigorous study of how young people are prepared for their roles as citizens in democracies and societies aspiring to democracy. Recently, Dr. Torney-Purta has focused on enhancing the public's and policy makers' understanding of methods of assessment and findings from cross-national studies of U.S. students' achievement across subject areas. She has also led efforts to promote effective international collaboration among researchers in the social sciences and education as a member of the U.S. National Committee for Psychological Science at the National Academy of Sciences. She is a member of the National Academy of Education.