The current issue and full text archive of this journal is available at
www.emeraldinsight.com/0968-4883.htm

Understanding validity issues in international large scale assessments

Meiko Lin
Interdisciplinary Studies, Teachers College, Columbia University, New York, New York, USA

Erin Bumgarner
Tufts Interdisciplinary Evaluation Research, Tufts University, Medford, Massachusetts, USA, and

Madhabi Chatterji
Organization and Leadership, Teachers College, Columbia University, New York, New York, USA

Abstract

Purpose – This policy brief, the third in the AERI-NEPC eBrief series "Understanding validity issues around the world", discusses validity issues surrounding International Large Scale Assessment (ILSA) programs. ILSA programs, such as the well-known Programme of International Student Assessment (PISA) and the Trends in International Mathematics and Science Study (TIMSS), are rapidly expanding around the world today. In this eBrief, the authors examine what "validity" means when applied to published results and reports of programs like the PISA.

Design/methodology/approach – This policy brief is based on a synthesis of conference proceedings and review of selected pieces of extant literature. It begins by summarizing perspectives of an invited expert panel on the topic. To that synthesis, the authors add their own analysis of key issues. They conclude by offering recommendations for test developers and test users.

Findings – ILSA programs and tests, while offering valuable information, should be read and used cautiously and in context. All parties need to be on the same page to maximize valid use of ILSA results, to obtain the greatest educational and social benefits, and to minimize negative consequences. The authors propose several recommendations for test makers and ILSA program leaders, and ILSA users. To ILSA leaders and researchers: provide more cautionary information about how to correctly interpret the ILSA results, particularly country rankings, given contextual differences among nations. Provide continuing psychometric or research resources so as to address or reduce various sources of error in reports. Encourage policy makers in different nations to share the responsibility for ensuring more contextualized (and valid) interpretations of ILSA reports and subsequent policy development. Raise awareness among policy makers to look beyond simple rankings and pay more attention to inter-country differences. For consumers of ILSA results and reports: read the fine print, not just the country rankings, to interpret ILSA results correctly in particular regions/nations. When looking to high-ranking countries as role models, be sure to consider the "whole picture". Use ILSA data as complements to other national- and state-level educational assessments to better gauge the status of the country's education system and subsequent policy directions.

Originality/value – By translating complex information on validity issues with all concerned ILSA stakeholders in mind, this policy brief will improve uses and applications of ILSA information in national and regional policy contexts.

Keywords Validity, TIMSS, International assessments, International large scale assessments, PISA, Comparative international assessments

Paper type Research paper

Quality Assurance in Education
Vol. 22 No. 1, 2014
pp. 31-41
Emerald Group Publishing Limited
0968-4883
DOI 10.1108/QAE-12-2013-0050

Introduction

International large scale assessment (ILSA) programs, such as the well-known Programme of International Student Assessment (PISA) and the Trends in International Mathematics and Science Study (TIMSS), are rapidly expanding around the world today[1]. Launched in 1997, PISA tests have been administered to students in Australia, Asia, North America and Europe since 2000. ILSA programs like the PISA and the TIMSS are now being adopted in many developing nations in Asia and Africa. For more information on PISA and TIMSS, see www.oecd.org/pisa/ and www.iea.nl/

This policy brief, the third in the AERI-NEPC eBrief series "Understanding validity issues around the world", discusses validity issues surrounding ILSA programs. Countries are often ranked based on ILSA results, engendering inter-country comparisons. What do these country ranks really mean and does it make sense to use ILSA ranks as indicators of the quality of a country's education system or, by extrapolation, its economic performance? How can ILSA results be used productively and meaningfully for reforming education systems in aspiring regions?

In this eBrief, we examine what "validity" means when applied to published results and reports of large-scale, international assessments like the PISA. Following a discussion of the panelists' views, we present our own perspectives on future directions and for promoting more valid use of the ILSA results.

Who and what this eBrief speaks to

The validity issues discussed here have application to a wide variety of audiences. This eBrief could be helpful to ILSA program researchers by answering the following types of questions:

- How should we approach the tasks of validation – or of gathering of validity evidence to support defensible interpretations and uses of ILSA reports – and of educating users about ILSA programs, so that country-level users of ILSA results can take appropriate actions towards building, improving or reforming their education systems?

Additionally, this eBrief could help policy makers and public users of ILSA reports (e.g. the media, educators and other public stakeholders) answer the following types of questions:

- What are some valid and appropriate ways to interpret and use ILSA reports in national policy contexts? Which interpretations and uses should we avoid making?
- What do the ILSA results mean for students, teachers and schools in my region versus internationally? What are the limitations in the information?

Method

This policy brief is based on a synthesis of conference proceedings and selected pieces of extant literature. It begins by summarizing perspectives of an invited panel of measurement and policy experts on the topic. To that, we add our own analysis of key issues. We conclude by offering our own thoughts and recommendations.

The content of this eBrief is derived from the keynote presentation and chapter by Michael J. Feuer (Feuer, 2013), as well as reactions by an invited panel of discussants (Backhoff, 2013; Laurie, 2013; Plisko, 2013; Wagemaker, 2013) and audience

discussions that followed at a 2012 conference held at Teachers College, Columbia University titled: "Educational assessment, accountability and equity: conversations on validity around the world". Following a summary of the main ideas from the panelists, we provide our own thoughts on the validity, fairness and test-use issues in these contexts.

A summary of main themes

Michael Feuer's main ideas

Michael Feuer (2013) endorses the benefits of ILSA programs and offers new guidelines for improving ILSA validation efforts. In discussing his concerns related to how ILSA results and reports tend to get used internationally, Feuer (2012, 2013) alerts us to three issues:

(1) ILSA results have an unavoidable impact in shaping the educational policy discourse in different nations, regardless of whether there is empirical validity evidence to support the conclusions made from the reports.

(2) There is a worldwide tendency to falsely but directly link ILSA results to a country's economic outcomes.

(3) There is a need for a comprehensive validation framework for ILSA programs that takes into account the core values underpinning education development efforts and policy goals in different nations.

Impact of ILSA results on the educational policy discourse

Providing examples from industrialized nations like the US, Germany and Japan, Feuer (2013) provides a compelling account of the significant role that ILSA reports play in global education reform discussions and national policy-making. Results from ILSA reports frequently make newspaper headlines, he observes, and are used to shape the public's thinking about national educational reforms. This happens regardless of any considerations as to whether the particular interpretations of ILSA results are meaningful or valid.

In 2000, for example, Germans were devastated to learn from the Organization for Economic Cooperation and Development (OECD) that teenagers in their country scored significantly lower than all other participating countries on the PISA tests in all subjects. In response, Germany designed dramatic, large-scale educational reforms that were implemented at a rapid pace. Feuer explains how Germans interpreted ILSA results, ignoring other – and in some ways contradictory – evidence on student achievement that suggested the German education system was not in fact declining in performance.

And Germany is not an exception in its response to ILSA reports. ILSA results have had important consequences for national policy reforms in both developed and developing countries.

Linking ILSA results falsely to economic outcomes

Next, Feuer (2013) examines whether ILSA results affect a country's economic future. This is a popular perception embraced by politicians around the world. Using international data, Feuer shows that any claims about a relationship between educational outcomes, as measured by ILSA programs, and economic outcomes of countries are largely without sufficient empirical support. Historical evidence fails to

show adequate links of ILDA data with macroeconomic outcomes of counties. Regardless of the seeming average or underperformance on ILSA programs by US students compared to other developed nations, for example, the country's economic indicators have generally remained steady and strong.

Feuer further asserts that to answer the question about the connection between educational and economic indicators, we must first make sure that:

- the interpretations of available economic and educational performance data are reasonable and valid in the first place; and
- the theory that education has a causal relationship with the economy of a nation measured with macro-level indicators has some empirical support (or validity).

Data on economic and educational indicators often provide only simple summary statistics, such as mean performance of 15-year-olds, which makes interpretations of the "full picture" impossible. For example, mean performance does little to explain how certain sub-populations are performing (e.g. the disadvantaged students, second language learners, or immigrants). Feuer concludes that large-scale international assessments in education are not adequately valid data sources for forecasting a nation's overall economic well-being.

New validity framework for ILSA programs

Feuer (2013) proposes a new ILSA validation framework using Messick's (1995) concept of "consequential validity," to address the intended and unintended consequences of using ILSA reports to guide practices or policies. Feuer's framework consists of six principles and a series of questions that can guide both test developers and policy makers towards considering the issues of consequential validity through discussions, as they go about ILSA program adoption:

(1) Articulate the intended rationales for using ILSA data to guide policy, and consider whether those rationales are consistent with larger educational goals. Within the context of given regions and nations, users should evaluate to what extent the use of the proposed assessment, e.g. PISA or TIMSS, agrees with the core values of education. Users should also ask if by participating in ILSA programs, the intent is to foster public deliberations about core values in education. Similarly, users should evaluate to what extent using comparative ILSA program reports fits with a nation's values.

(2) Be transparent regarding the embedded assumptions and logic of relationships of ILSA results to policy choices. Before ILSA participation, users should evaluate whether the logic and empirical evidence on the meaning and quality of information from ILSA programs is sensible for national or regional purposes.

(3) Estimate potential benefits and risks of using ILSA results to guide educational practice and policy. Users should ask whether international rankings provide an adequate basis for development of education reform policies. "To what extent are downside risks of making reform decisions based on comparative scores taken into account by policy makers and educators? Do flaws in the comparative score data and their interpretation distort policy judgments and, if

so, how are the effects of such errors distributed across economically and socially diverse schools and school systems?" (Feuer, 2013).

(4) Acknowledge the compelling nature of international comparative rhetoric, and assess benefits and risks of ILSA participation to education systems. "To what extent does reliance on ILSA contribute to erosion of morale about the quality and prospects of genuine school reform and improvement of teaching and learning for all students? Is there a macro-level downside risk associated with exaggerated claims of decline and stagnation in educational performance, especially as it is implicitly or explicitly linked to long run economic performance?" (Feuer, 2013).

(5) Assess the benefits and costs of participation in ILSA programs. "What criteria should guide decisions by policy makers to invest in continued improvement of assessment programs and continued participation?" (Feuer, 2013).

(6) Think of ILSA validation as "procedural rationality" in the larger context of schooling. "How can a comprehensive approach to validity of ILSA promote and facilitate continuous improvement in teaching, learning, and education policy?" (Feuer, 2013).

Reactions to Feuer: main ideas

In their respective discussions of Michael Feuer's perspective, Hans Wagemaker, Eduardo Backhoff, Valena White Plisko, and Robert Laurie address the utility ILSA programs and of Feuer's validation framework in mitigating the undesirable consequences of misusing ILSA results and reports.

Wagemaker (2013), like Feuer, urges us to consider the broader context and the differences in the detail among nations participating in ILSA programs before using international rankings from ILSA reports to inform educational reforms. The purposes of ILSA are to provide data to inform policy reforms and educational improvement in different countries, but interpretations should be contextualized.

As an example of a confounding variable that challenges levels of validity in interpreting the 2009 PISA results, he points to the unreasonable assumption that adolescents of a given age are all exposed to same material in different nations, and thus equally prepared to take the same test. The PISA tests are typically administered to 15-year-olds internationally. Grade levels attended by those 15-year-olds vary greatly within and across countries. Wagemaker shows that while the majority of Polish 15-year-olds were in Grade 9 in 2009, the majority of 15-year-olds in New Zealand were in Grade 11. As such, country rankings on ILSA reports can be deceptive. Wagemaker thus urges ILSA program providers to offer support to countries so that they can correctly and meaningfully interpret reports.

Backhoff (2013) agrees with Wagemaker that inter-country comparisons cannot be based on simple rankings. He calls our attention to many other validity threats in nations where there are cultural and linguistic variations. These factors can cause unexpected changes in countries' measured learning outcomes and ILSA rankings. A few of these and other sources of error are:

- cross-cultural test translation errors;
- cultural differences in survey response styles;
- cultural differences in reading the meaning of some questionnaire items;

- limitations in samples of students tested, such as particular regional biases which do not reflect a country's overall performance; and
- students' opportunity to learn what is tested, and the meaning of resulting group-level averages and comparative reports.

Backhoff (2013) cautions that the biases embedded in ILSA reports will likely exert a negative effect on education policies. He, too, calls for more attention on reporting and interpreting ILSA results in appropriate and valid ways.

Laurie (2013) similarly finds Feuer's validity recommendations to be quite appropriate and applicable to the Canadian context. He cautions that care must be taken when using ILSA reports and results to guide educational policy at the regional or national levels. For example, since Finland has consistently performed well on the PISA, politicians, educators, administrators and journalists from around the world have rushed to Finland to see what they could learn from the Finnish education system. Too often, these "education tourists" fixate on singular aspects of Finnish education instead of the whole system. Laurie states that many successful components of the Finnish school system are interwoven with policies of the surrounding welfare state. Therefore, to simply transfer a singular aspect of the Finnish education model to a different educational system in a different country like Canada would likely not be successful. The many other contextual differences between the countries must be taken into account. Laurie's (2013) complete response includes an application of Feuer's six principles on ILSA validation to Canada's case based on the PISA results in 2000.

Plisko (2013) highlights the benefits of ILSA programs as providing standardized, benchmark assessment results that different nations could find useful. She explains the value of combining results from multiple ILSA programs to understand a country's overall performance, as each provides a different kind of information. For example, the TIMSS and PIRLS assess curriculum-based learning whereas PISA assesses applied knowledge and skills. In her view, each provides a distinct perspective on US student performance. Additionally, Plisko suggests that international assessment data can also be beneficially used in tandem with national testing programs such as the National Assessment of Educational Progress (NAEP) in the US. This would help evaluate progress towards desired educational outcomes in the long term against externally set benchmarks. The NAEP measures the performance of US students and also provides trends for individual states and demographic sub-populations.

Audience questions

Audience members acknowledged the benefits of ILSA programs, but expressed concerns nonetheless.

For example, a researcher from Iceland expressed concern about the use of ILSA data for analyzing or evaluating educational systems in a global context when not interpreted in a cultural context. Specifically, he raised concerns that Iceland is now standardizing the education system in order to focus more on improving its standing in ILSA program rankings, but not all students perform well on ILSA tests. Feuer (2013) responded by saying that a country's commitment to more inclusive education inevitably leads to some form of standardization of assessment metrics. On the other hand, the over-reliance on such metrics for policy actions can lead to negative consequences for the most disadvantaged sub-groups within the population, who tend

to not perform well. The way forward might be by inviting everyone to consider how standardized test programs can be developed and used to facilitate equal educational opportunity for all.

A second audience member, an education leader from Bangladesh, asked if his country should participate in ILSA programs. Feuer's primary response strongly endorsed Bangladesh's participation in ILSA because results from ILSA can offer valuable insights to a country's educational system. Feuer also acknowledged that research has been scarce concerning how ILSA results have been used or misused to date.

Conclusions

Our thoughts

The predicament around ILSA has less to do with the participation in ILSA but more to do with the inferences that can be legitimately and validly drawn from ILSA results. ILSA programs yield valuable information about a country's education system. Unfortunately, ILSA results are often interpreted solely in terms of inter-country rankings. As seen in the discussions of all the contributors here, focusing solely on country rankings can be subject to different degrees of invalidity when taken out of context. Misinterpretations could have negative consequences by spreading misinformation in larger national and societal contexts. When important contextual factors are ignored in generating the countries' average scores that are ranked, ILSA results will have little meaning or value.

Based on a content analysis of validity issues discussed in the full-length chapters based on the conference presentations, Chatterji (2013) identified some of the current challenges facing ILSA test makers/researchers and ILSA users at large, where the latter group includes educators policymakers, and the media. The challenges are summarized in Figure 1, with recommendations for ILSA program developers and users in Figures 2 and 3. We elaborate on a couple of these items below.

Stakeholders at different levels of national education systems may have multiple and different ILSA related information needs that remain unknown to ILSA program developers, who may end up designing the program components to serve a narrower set of purposes. In other instances, the diversity levels of ILSA test-takers shift drastically as the program expands to new regions, but too few or no changes occur in the test design, validation or reporting procedures. The new groups of test-takers perform differently than expected on the ILSA tests or particular test items, with the risk of less valid or biased results for these groups. All such oversights lead to collection of validity evidence that is too limited for the intended actions by regional decision-makers (see Figure 1).

Feuer's validation framework and Kane's (2013) argument based approach offer a viable and complementary solutions for addressing these issues systematically. In addition, to improve validity of interpretations of results from ILSA programs, Chatterji (2013) recommends that measurement and evaluation specialists use "systems-based logic models" to guide validation procedures and improve communications with stakeholders to forestall unintended consequences that could clash with well-intentioned goals of regional or national education systems (see Figure 3).

Examples of ILSA Validity Challenges

Users' unreasonable assumptions of links between educational system outcomes based on students' ILSA scores and economic productivity of nations.

- Firm user beliefs that ILSA reports are good enough for local decisions and can be used *without* careful, context-based appraisals.
- Units of analysis and data aggregation levels that are unclear at different levels. E.g., Is it the nation, schools, teachers, or students that are being evaluated? To what degree are interpretations at all levels valid?
- Overlooked assessment context and regional dynamics unique to individual nations or regions.
- Levels and types of ILSA data use that are not clear or adequately specified as a part of test design, analysis and reporting programs.
- Factors outside the tested domains that interfere with, or are necessary to understanding of, what ILSA scores and program results really mean. e.g., student demographics, instrument errors due to translation biases, test or item biases in sub-groups, test administration issues, opportunity to learn issues, and sampling limitations.
- Conflicting assessment purposes, values, and information needs of different assessment stakeholders of ILSA programs within countries or internationally.

(Chatterji, 2013b, pp. 273-307)

Figure 1.
Example of ILSA validity challenges

Recommendations

As evident, involvement of all stakeholders at different levels of education systems is necessary to enhance validity in ILSA contexts. Our specific recommendations for test makers and ILSA program leaders are the following:

- First, provide more cautionary information about how to correctly interpret the ILSA results, particularly country rankings, given contextual factors in different nations. Remove the language of "comparative international assessments" in publications and research papers that are made public which inadvertently encourage comparative interpretations.
- Second, provide ongoing psychometric and research resources so as to continually address or mitigate various sources of cultural, linguistic or other biases in regional and national ILSA reports as the programs expand to different countries (this recommendation should be read in tandem with next recommendation).

How ILSA Program Developers Can Improve Validity

- Align goals and assessment purposes at all levels of ILSA, guided by appropriate validation frameworks (Feuer, 2013; Kane, 2013).
- Develop systems-based logic models to contextualize and guide ILSA program development, validation and use of reports/results.
- Collect empirical validity evidence to determine score interpretations and uses that are supportable.
- Educate users about ILSA programs, delineating limitations of reports.
- To evaluate intended and unintended consequences of ILSA use, map out potential feedback loops tied to test scores as a part of validation.
- Make validity evidence on ILSA reports public and understandable.

(Chatterji, 2013, pp. 273-307)

Figure 2.
How ILSA program developers can improve validity

How Can ILSA Users-at-Large Help Improve Validity?

- Seek out appropriate validity information on ILSA tests, reports and testing programs before taking actions.
- Make decisions consistent with the ILSA program's stated purposes, populations, tested domains, and validity evidence available to support decisions.
- Curb over-interpretation
- Attend to a test's or ILSA assessment report's limitations.

(Chatterji, 2013b, pp. 273-307)

Figure 3.
How can ILSA users-at-large help improve validity?

- Third, given that large amounts of research resources have already been invested in visible ILSA programs towards studying cultural and contextual differences, as documented by TIMSS and PISA, encourage policy makers in different nations to share the responsibility for ensuring more contextualized interpretations of ILSA reports and subsequent policy development.
- Third, raise awareness among policy makers to look beyond simple ranking and pay more attention to inter-country differences as reported by OECD and IEA.

For consumers of ILSA results and reports (e.g. policymakers and educators), we recommend the following:

- First, read the fine print, not just the country rankings. To what extent does the test match with what students were taught in the country or region? Who took the test? How old were students? What kinds of cultural differences should we pay attention to in understanding the ILSA reports better?

- Second, when looking to high-ranking countries as role models, be sure to consider the "whole picture" and not just singular programs or policies (i.e. what works in Finland may not necessarily work in US).
- Third, use ILSA data as complements to other national- and state-level educational assessments to better gauge the status of the country's education system.

In summary, ILSA programs and tests, while offering valuable information, should be read and used cautiously and in context. All parties need to be on the same page to maximize valid use of ILSA results, to obtain the greatest educational and social benefits, and to minimize negative consequences. The value of ILSA reports and results depends on the ability of test makers and test users to understand the principal purposes of adopted ILSA programs in particular nations.

Acknowledgements

The first series of eBriefs "Understanding Validity Issues Around the World" is produced via a partnership between the Assessment and Evaluation Research Initiative (AERI) at Teachers College, Columbia University and the National Education Policy Center (NEPC) at the University of Colorado, Boulder. The inaugural AERI conference that generated the first series of eBriefs was co-sponsored by Educational Testing Service, Teachers College, and the National Science Foundation. Websites: www.tc.edu/aeri and nepc.colorado.edu

Note

1. Because this attempt to distill a great deal of information will necessarily lose some nuance and detail, readers are encouraged to access the original articles listed in the References section. AERI-NEPC eBriefs, or electronic versions of the items in the series are also available at the AERI and NEPC websites.

References

Backhoff, E. (2013), "Validity issues in international large scale assessment (ILSA) programs: thoughts for developing countries", in Chatterji, M. (Ed.), *Validity and Test Use: An International Dialogue on Educational Assessment, Accountability and Equity*, Emerald Group Publishing, Bingley, pp. 233-249.

Chatterji, M. (2013), "Insights, emerging taxonomies, and theories of action toward improving validity", in Chatterji, M. (Ed.), *Validity and Test Use: An International Dialogue on Educational Assessment, Accountability and Equity*, Emerald Group Publishing, Bingley, pp. 273-307.

Feuer, M. (2012), "Validity issues in international large scale assessments: truth and consequences", Educational Assessment, Accountability and Equity: Conversations on Validity Around the World Conference, Teachers College, Columbia University, New York, NY.

Feuer, M. (2013), "Validity issues in international large-scale assessments: 'truth' and 'consequences'", in Chatterji, M. (Ed.), *Validity and Test Use: An International Dialogue on Educational Assessment, Accountability and Equity*, Emerald Group Publishing, Bingley, pp. 197-215.

Kane, M. (2013), "Validity and fairness in the testing of individuals", in Chatterji, M. (Ed.), *Validity and Test Use: An International Dialogue on Educational Assessment, Accountability and Equity*, Emerald Group Publishing, Bingley, pp. 17-53.

Laurie, R. (2013), "Applying Feuer's validation framework in a Canadian context: a look at international large scale assessment programs", in Chatterji, M. (Ed.), *Validity and Test Use: An International Dialogue on Educational Assessment, Accountability and Equity*, Emerald Group Publishing, Bingley, pp. 263-271.

Messick, S. (1995), "Validity of psychological assessment: validation of inferences from persons' responses and performances as scientific inquiry into score meaning", *American Psychologist*, Vol. 50, pp. 741-749.

Plisko, V. (2013), "Validity and international large scale assessment programs: a reaction to Feuer's 'truth' and 'consequence'", in Chatterji, M. (Ed.), *Validity and Test Use: An International Dialogue on Educational Assessment, Accountability and Equity*, Emerald Group Publishing, Bingley, pp. 251-261.

Wagemaker, H. (2013), "International large scale assessment (ILSA) programs and the challenges of consequential validity", in Chatterji, M. (Ed.), *Validity and Test Use: An International Dialogue on Educational Assessment, Accountability and Equity*, Emerald Group Publishing, Bingley, pp. 217-231.

About the authors

Meiko Lin is a Senior Research Assistant at AERI and is currently studying for her Doctorate of Education in Inter-Disciplinary Studies at Teachers College, Columbia University. She holds an MA in Research Methods from the University of California at Los Angeles. Her interests include the application of measurement and evaluation techniques in educational, psychological, and health-related fields. Meiko Lin is the corresponding author and can be contacted at: ml2734@columbia.edu

Erin Bumgarner is project director of the Massachusetts Healthy Families Evaluation Project-2 (MHFE-2) at Tufts University. Her research interests focus on how policy and programs can improve the long-term outcomes of low-income children, especially for those from linguistically and culturally diverse backgrounds.

Madhabi Chatterji is Associate Professor of Measurement, Evaluation, and Education and the founding Director of the Assessment and Evaluation Research Initiative (AERI) at Teachers College, Columbia University. Dr Chatterji's publications focus on the topics of instrument design, validation, and validity; evidence standards and the "evidence debate" in education and the health sciences; standards-based educational reforms; educational equity; and diagnostic classroom assessment.