Measuring the Power of Learning.™

Research Report
ETS RR–18-27

Bridging Validity and Evaluation to Match International Large-Scale Assessment Claims and Country Aims

María Elena Oliveri

David Rutkowski

Leslie Rutkowski

December 2018

ETS Research Report Series

Since its 1947 founding, ETS has conducted and disseminated scientific research to support its products and services, and to advance the measurement and education fields. In keeping with these goals, ETS is committed to making its research freely available to the professional community and to the general public. Published accounts of ETS research, including papers in the ETS Research Report series, undergo a formal peer-review process by ETS staff to ensure that they meet established scientific and professional standards. All such ETS-conducted peer reviews are in addition to any reviews that outside organizations may provide as part of their own publication processes. Peer review notwithstanding, the positions expressed in the ETS Research Report series and other published accounts of ETS research are those of the authors and not necessarily those of the Officers and Trustees of Educational Testing Service.

The Daniel Eignor Editorship is named in honor of Dr. Daniel R. Eignor, who from 2001 until 2011 served the Research and Development division as Editor for the ETS Research Report series. The Eignor Editorship has been created to recognize the pivotal leadership role that Dr. Eignor played in the research publication process at ETS.

ETS Research Report Series ISSN 2330-8516

RESEARCH REPORT

Bridging Validity and Evaluation to Match International Large-Scale Assessment Claims and Country Aims

María Elena Oliveri,[1] David Rutkowski,[2,3] & Leslie Rutkowski[2,3]

1 Educational Testing Service, Princeton, NJ
2 Indiana University, Bloomington, IN
3 Centre for Educational Measurement, University of Oslo, Oslo, Norway

Fifty years after the first international large-scale assessment (ILSA), participation in these studies continues to grow, with more than 50% of the world's countries participating. Concomitant with growth in ILSAs is an expansion in the diversity of participant countries with respect to languages, cultures, and educational perspectives and goals. As educational aims might differ for new participants and goals among historic participants can be expected to shift over time, it is useful to understand the degree to which countries' expectations of ILSAs—as a means for understanding their educational system—align with the explicitly and implicitly stated purposes of these studies. In this presentation, we shift the conversation away from countries reporting ILSA shock and dissatisfaction with participation to fostering a productive conversation about the value and utility of participation. We propose a framework that combines notions from meta-evaluation to systematically test the evaluation tools—ILSAs and validity theory (in relation to test use and alignment with stakeholder needs) to help countries understand why they participate in ILSAs and the potential value in taking part. We develop this conceptual framework with the aim that countries can (a) systematically consider their educational goals and the degree to which ILSA participation can reasonably help countries monitor progress toward them; (b) use an argument model to analyze claims by ILSA programs against the background of a country's specific context; and (c) more clearly understand intended and unintended consequences of ILSA participation. The framework offers a tool to systematically think through a complex web of implicit and explicit purposes, goals, and actors related to ILSAs and educational systems. To demonstrate our proposed framework, we review national education agendas in several countries with differing educational traditions (e.g., the United States, Mexico, and Norway) against published ILSA frameworks. Using our proposed method would offer a set of general guidelines that national funders can use to chart a path forward in terms of future ILSA participation. It can also equip participating countries with the knowledge to engage in reasoned conversations with testing organizations regarding unmet needs from testing programs.

Keywords ILSA; consequences; validity; evaluation; fairness

doi:10.1002/ets2.12214

Fifty years after the first international large-scale assessment (ILSA), participation in these studies continues to grow, with more than 50% of the world's countries taking part (Kamens & McNeely, 2010). Concomitant with this growth is an expansion in the diversity of participating countries with respect to languages, cultures, educational perspectives, and goals. In line with this growth, ILSAs have evolved over time not only to assess mathematics, language, or scientific knowledge but also to include a range of questionnaires administered to students, teachers, principals, and parents. These questionnaires often serve a supporting role in the assessment by optimizing achievement estimates (Mislevy, Johnson, & Muraki, 1992) and contextualizing achievement results. Today, questionnaires that measure bullying, classroom climate, and teaching styles are gaining importance as outcomes in their own right—beyond serving as a context for understanding educational achievement, they serve as evaluative tools. For example, recent Trends in International Mathematics and Science Study (TIMSS) and Program for International Student Assessment (PISA) results were used to identify an epidemic of misbehaving students in Australia (Matthews, 2017). Beyond school-based surveys, the issues discussed in this report also apply to out-of-school and adult surveys.

The current state of ILSAs is the result of negotiations between educational jurisdictions (ministries of education) and testing organizations over the past 50 years. The original purpose of modern-day ILSAs was mainly research and was spearheaded by the early founders of the International Association for the Evaluation of Educational Achievement (IEA, 2016). However, some scholars contend that current ILSAs are also governing tools for global institutions, such as

Corresponding author: M. E. Oliveri, E-mail: moliveri@ets.org

the Organisation for Economic Co-operation and Development (OECD), where ILSAs serve as a vehicle through which international organizations set a global educational agenda (Sellar & Lingard, 2013). In either case, the responsible organizations administer a collection of tests and questionnaires, the results of which are used to judge the merit and worth of educational systems around the world.

As ILSAs continue to grow in terms of content domains, platforms, and especially in the case of the OECD's PISA study, policy prescriptions, current and future participants have varied expectations around what a given ILSA can provide. Furthermore, reasons for participation are also highly varied across participants. For example, OECD countries are expected to participate in PISA, whereas non-OECD countries volunteer or are otherwise incentivized to participate. In light of these varied motivations and aims, misalignments might occur between the participating countries' educational goals and what a testing organization is able to provide, leaving gaps in participants' needs. Such misalignments and shifts, however, can result in unintended consequences, which are defined as uses of test scores that lead to unplanned impacts on one or more stakeholders (Chatterji, 2013).

Examples of unintended consequences may include the inappropriate use, misuse, overgeneralization, or lack of knowledge of a test's limitations that leads test users to make score-based inferences that are more ambitious than the testing program can withstand. For instance, unintended (unforeseeable) consequences might arise when there is a shift in the characteristics of the test takers or stakeholders, such as a test being developed for a homogeneous population that through time becomes more diversified. This shift is likely to occur, given that ILSA participation has greatly increased with respect to the number of educational jurisdictions, cultures, languages, and educational paradigms of the participating OECD and non-OECD participants, giving rise to greater levels of diversification of the assessed population. This diversification presents possible threats to validity and possible unintended effects that participants need to understand as part of accurate score interpretation and use.

For instance, one threat to validity may be introduced through overlooked cross-cultural differences among participants. Such differences need to be considered as part of score interpretation. To illustrate, Oliveri and von Davier (2017) pointed out that partner countries might differ from OECD countries in relation to exposure to the item types used in the test, proximity of their language to the language in which the test is developed, and exposure to the curricula or opportunity to learn across countries. Additional threats might arise when tests are administered across learner populations representing a wide performance range, particularly when the test was constructed for a higher achieving population. Such complications might arise if new participating educational systems have a much lower proficiency level than the targeted proficiency distribution of the test. If test scores are used, for example, to inform policy for those countries, it is important to consider if the lower proficiency systems receive the information they seek (Backhoff, 2013). In fact, when a system's proficiency is much lower than the average testing population, the validity of any resulting score for the lower performing system could be compromised. Part of the mitigation plan may be to consider all other data collected during an assessment cycle and to contextualize the results through national or international documents, such as the IEA-produced encyclopedia that describes in some detail the policy in an educational context for each participating country.

Beyond possible problems of interpreting threats to validity as true variations in score meaning, Backhoff (2013) highlighted that unintended consequences might emerge if developing countries seek to clone first-world countries' policies and programs to improve their own educational results. This practice could lead to unintended consequences because developing countries' aims could differ substantially from those of developed countries and of the OECD. Adoption of ill-fitting policies might lead to wasted instructional interventions and wasted resources, both of which are unintended, negative consequences of test use.

Chatterji (2013) pointed out that unintended consequences could also arise from public users' possible misinterpretation or misuse of ILSA reports. This threat might arise if ILSA score rankings are taken at face value by the public and the media, without carefully attending to the purposes of the assessment program, the technical limitations of the information provided by the test, or local differences in educational systems. For example, in cases for which a national government is the primary data user, possible misuses could arise when scores are not appropriately contextualized. Issues could also arise if scores are kept away from the public or if they are used outside of stated intentions and possibilities (e.g., assigning individual student or school scores, sanctioning or rewarding teachers; Backhoff, 2013). Such uses clearly deviate from the stated ILSA purposes.

In light of the growing diversity of participating countries, organizations that administer ILSAs also face a number of constraints. For example, planning an assessment that includes a large number of diverse educational systems with

different cultures, resources, curricula, purposes for education, and academic achievement is a difficult task. The technicalities of placing these participants on a common scale are challenging and often require the organizations to focus the assessment on a specific subject (e.g., math, science) and subdomains of select subjects (e.g., numeracy, algebra) that are common across the majority of participating countries. Furthermore, testing organizations often face constraints by participating countries that require limiting testing burden on schools and students but also wish to meet the ever increasing needs of participating countries. In other words, the testing organizations must provide countries with the optimal amount of information given fixed resources. Such practical constraints that participating countries often place on testing organizations clearly limit how the test scores can be interpreted.

An additional complexity around goal alignment between ILSA results and participants' needs or wants is that educational aims are not static and change over time. Hence new or changing goals might no longer align with the stated aims of a given ILSA. These sorts of misalignments and unintended consequences present the need for an evaluation tool that countries can use to better understand the degree to which an educational jurisdiction's goals align with a test's particular capabilities. The tool would serve to elucidate (a) the expectations of educational jurisdictions for participating in ILSAs, (b) the degree to which jurisdictions' expectations of ILSAs — as a means for understanding their educational systems — align with the stated purposes of ILSAs, and (c) the importance of explicitly articulating (mis)alignments. These purposes seek to better guide jurisdictions in judiciously using ILSA data through their participation in the assessment. Unfortunately, few tools exist for these purposes. Thus we take up this gap in the literature and develop a framework that can be of use.

Specifically, we aim to shift the conversation away from countries reporting ILSA shock (Pons, 2012) and dissatisfaction with ILSAs. Instead, we seek to foster a productive conversation about the value and utility of participation for educational jurisdictions. More specifically, we address three objectives: (a) to guide educational jurisdictions to consider their national goals and the degree to which ILSA participation can reasonably help them monitor progress toward those goals, (b) to describe the use of a logic model to analyze claims by ILSA programs against the background of the educational system's specific context, and (c) to exemplify intended and unintended consequences of ILSA participation.

To address our objectives, we present a model that can be used by ILSA participants to systematically evaluate the extent to which their goals for participation in ILSAs can reasonably be addressed under the stated claims of ILSAs like TIMSS or PISA. The model can also be useful to researchers across a range of disciplines, such as policy formation and measurement that are increasingly interested in the use of national and international data from ILSAs. We demonstrate the proposed model and explain each of its steps by providing relevant examples. Our proposed approach offers a set of general guiding questions that educational system stakeholders can use to chart a path forward in terms of current and future ILSA participation. Additionally, we illustrate foundational elements needed to engage in reasoned conversations with testing organizations so that national governments can better explain what ILSAs can do to better meet their needs. We also provide examples of educational systems that have previously engaged in similar discussions and the types of actions they have implemented when addressing such questions. Examples like these might help generate future strategies for other educational systems to address possible areas of misalignment while learning from the approaches of other countries' educational systems.

International Large-Scale Assessments as an Evaluation System

We cast the conceptual model we present later in the report within an evaluation paradigm. Fitzpatrick, Sanders, and Worthen (2011) defined *evaluation* as "the identification, clarification, and application of defensible criteria to determine an evaluation object's value (worth or merit) in relation to those criteria" (p. 7). ILSAs — under this definition — are ideally situated to serve as evaluations: ILSAs can be used to help identify issues within, and provide clarity about, national educational systems in relation to evidentiary criteria to help determine their worth or value. The use of ILSAs as evaluations raises three questions that are fundamental to any evaluation: (a) What is the purpose of the evaluation? (b) How will the evaluation findings be used? and (c) How will the evaluation findings impact the various stakeholders?

Purposes of International Large-Scale Assessments

Fitzpatrick et al. (2011) noted that "determining and understanding the purpose of the evaluation is probably the most important job the evaluation sponsor or client will have in the course of an evaluation" (p. 260). This statement is particularly relevant in the ILSA context, because results are used for formative, summative, or dual purposes. For instance,

although Schleicher (2012) characterized PISA uses as formative, the focus on league tables and rankings also suggests a summative use (Takayama, 2008). This dual (summative and formative) role makes it even more complex for educational jurisdictions to understand the consequences of participation, increasing the need to ensure that the test can serve the participants' purposes within their own educational paradigms.

To elaborate, a summative evaluation determines the merit or worth of a program or system by informing a recommendation regarding retaining, altering, or eliminating a program. ILSAs inform summative aspects of educational systems because (whether intentionally or not) they provide the basis for making judgments about an educational system's overall worth or merit. For instance, in the case of the OECD and PISA, the evaluating organization makes summative recommendations on which aspects of the educational system should be retained, altered, or eliminated. The commissioned reports by Nusche, Earl, Maxwell, and Shewbridge (2011) are examples of the use of PISA results to inform holistic decisions about the disposition of the educational system and to advise, for example, Norway, on the ways in which its educational system could change. ILSAs might also serve a formative role by providing data to inform ongoing program improvements and to monitor change occurring as a result of educational interventions. Wandall (2013) exemplified that Nordic countries also use ILSA results as a data source to help inform pedagogy and didactics.

A clear purpose is also important in developing a contract between the evaluator (e.g., the OECD or IEA) and the assessment administrator (e.g., the participating educational system and, in the case of PISA, the OECD). We discuss this dual role subsequently. The goal is to articulate clearly, for both parties, what the evaluation is intended to achieve. Moreover, clarity is needed to safeguard countries from differing and complex purposes associated with participation. One source of complexity is that the evaluator might have its own purposes or needs for test development and use. For example, the OECD (i.e., the evaluator) uses PISA scores for various purposes, such as to inform (a) teacher-related policies (Organisation for Economic Co-operation and Development [OECD], 2005), (b) economic outlooks (OECD, 2010), and (c) the general well-being of each participating jurisdiction (OECD, 2015).

A second source of complexity is that the OECD is acting as both evaluator and client, leading the organization to have both an agenda for ILSA development and a stated use for the evaluation. As such, it is ever more important for the participating educational jurisdiction to clearly understand the extent to which participation will provide information to help address its needs and which aspects of the evaluation and the evaluator's purpose may be at odds or in competition with national purposes (Backhoff, 2013). We illustrate a model to aid in understanding these possible tensions subsequently.

Using Evaluation Findings

A clearly stated evaluation purpose is important to help prepare clients or evaluation users to receive, interpret, and use the results of the evaluations. For example, stakeholders, such as ministries of education, national leaders, think tanks, and researchers, who will use findings from evaluations to inform relevant score-based decisions might be highly interested in communicating to all stakeholders that the evaluation results will be made public. Backhoff's (2013) portrayal of Mexico's participation in TIMSS is a case in point. Backhoff described that the nondisclosure of ILSA results by the Minister of Education for over a decade postparticipation might have led to negative consequences as potential users were unable to use the ILSA results to inform educational reforms. This issue sheds light on the importance of not only clearly stipulating the reasons for participation but also devising a communication plan for the ILSA results with relevant stakeholders. Additionally, it might help set limits on the types of uses (or misuses) of the ILSA results, such as using the data to impose sanctions or penalties for educators or laying off teachers, which are outside of the test's appropriate uses.

In cases when stakeholders want to use ILSA results to inform more than one purpose, a clear understanding of the original or intended purpose(s) for participation in the assessment can help determine the types of questions the assessment can inform and which questions would require other data sources or approaches to supplement ILSA scores. Otherwise, lack of clarity in how evaluation or, in our case, assessment findings will be used may lead to a number of unmet expectations regarding the extent to which ILSA results can address all educational systems' goals, leading to both misuse and underutilization of ILSA results (Breakspear, 2012; Grek, 2009; Hoplins, Pennock, Ritzen, Ahtaridou, & Zimmer, 2008).

Impact on Stakeholders

Generally, the primary stakeholders and funders of ILSAs are national governments and nonnational jurisdictions (e.g., Scotland, Shanghai, and Palestine). An exception is that over the years, the World Bank has funded ILSA participation

for a number of jurisdictions, acting as the client and the stakeholder. Because the primary stakeholder works with the testing organization to develop a clear purpose for the evaluation, which should include the relevant evaluation questions, identifying all critical stakeholders early on in the negotiation process is important. The next step after identifying the primary stakeholders is to clearly identify stakeholders' needs vis-à-vis ILSA purposes. As mentioned, the extent to which the ILSA goals align with stakeholders' national goals should be considered. In cases of misalignment, stakeholders should think through an action plan to supplement ILSA scores in other ways, such as the use of additional measures. For instance, participants may seek to augment ILSA data with local (national) assessments, such as is done in the United States, where the National Assessment of Educational Progress can be used to supplement US international assessment results to focus on particular aspects of the national educational system, which can also help inform resource allocation and policy development (White Plisko, 2013).

In the model description that follows, we assume that national and subnational jurisdiction governments are the primary clients for ILSAs. We reiterate that successful participation should involve clarity in the purpose of the evaluation and also in the participants' goals, despite the expectation of there being substantial variability across participants' goals, with more than 60 participants in each of the three largest ILSAs (Progress in International Reading Literacy Study, PISA, and TIMSS). A compromise between the various stakeholders and the testing organizations involved is needed, given the scale and scope of international assessments. A broad approach is needed to serve participants reasonably well.

The Evaluation Model

The evaluation model is designed to help national and jurisdiction-level policy makers (and other stakeholders) better understand the evaluative process of ILSAs and identify and elucidate areas of alignment and misalignment between an educational system's needs from ILSAs and the actual purposes and possibilities a given ILSA can offer. In instances when the assessment administrator (e.g., OECD) and participants' (educational systems') claims match, the model provides a logic argument that can be used by participants to help further clarify the extent to which this match will lead to the type of outcomes the participant (educational system) seeks to evaluate.

The model also aims to help participants identify inconsistencies between national needs and what the ILSA assesses and reports. It is important to identify such inconsistencies to help test users devise plans to supplement ILSA data with other data sources to help mitigate possible unintended consequences of test use, such as unforeseen consequences related to using poorly framed analytical frameworks to analyze ILSA data (Ercikan, 2009), or unforeseeable consequences related to not acknowledging rapid changes in the makeup of the assessed population (Oliveri & von Davier, 2016). Identifying the claims that match and those that do not can be a useful practice to help countries better understand, articulate, and prepare for both the intended and unintended consequences of ILSA participation. The proposed model is also designed to help test users identify possible issues and challenges that might arise with respect to determining the extent to which their aims match ILSA claims. The importance of this model can be highlighted through the presentation of instances in which such misalignments might have occurred and perhaps better outcomes could have been obtained through clearer and more purposeful preemptive matching of aims to claims.

Note that the model we present next assumes that testing organizations can and should be taken at their word, that is, that the claims that an organization like the IEA or OECD makes about what a particular ILSA can achieve are based on reasonable evidence and previous research. Consequently, our proposed model is not intended to be used to evaluate the veracity of the claims international testing organizations make but rather the extent to which such claims fit with a particular client's goals.

A representation of the model can be found in Figure 1 and could be used as follows. First, we assume that relevant stakeholders in some educational system, say, Country A, have clearly articulated reasons for participating in a given ILSA (e.g., to monitor overall achievement, to measure socioeconomic status-based achievement gaps over time). Clearly this is no trivial task, and the development of these goals is exogenous to the model; however, Feuer (2013) provided guiding questions to help with the process of articulating the desired claims.

Then, the first step in the model is a matching exercise whereby a participating country compares its reasons for participating against the known aims of a given ILSA. As an example, imagine that a primary goal of Country A's participation in TIMSS is to monitor math achievement at the lower end of secondary schooling. This aim is clearly consistent with TIMSS's stated purpose, which is the "assessment of mathematics and science at the fourth and eighth grades" (IEA, 2013, p. 3). This reason is thus marked as *consistent* with the aims of TIMSS, and the next step in the model can be pursued.

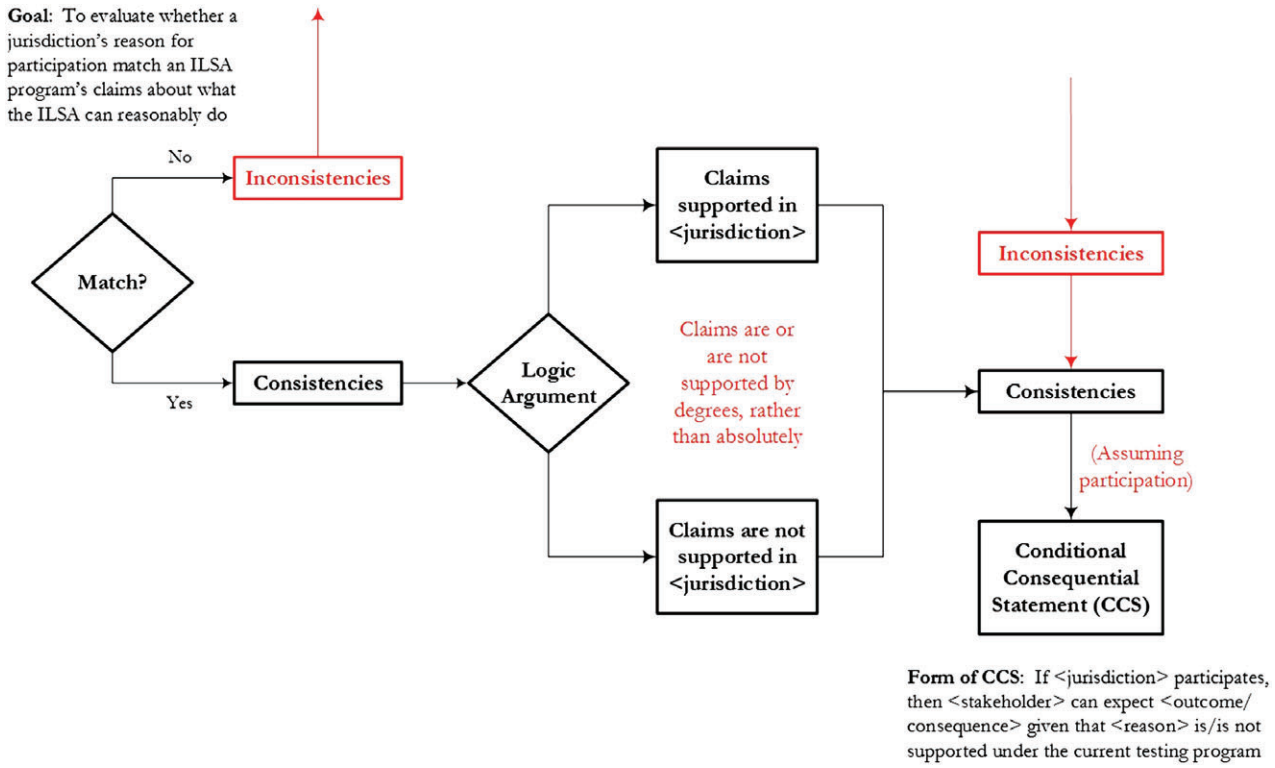

Figure 1 Model for valuing international large-scale assessment participation. Note that some text in this model is in red to call particular attention to those parts of the model that are either implicit (e.g., claims supported by degrees) or exogenous to the model (e.g., inconsistencies that leave the model, but return later for evaluation.

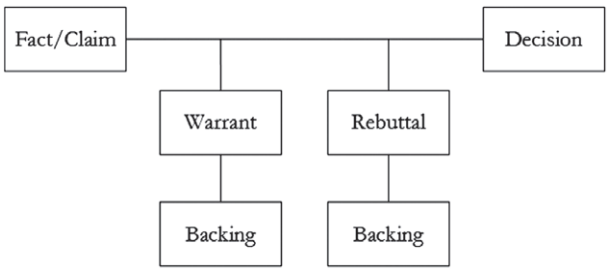

Figure 2 Logic argument from a model for valuing participation.

We return to this subsequently. In contrast, an aim that does not match with the mandate of the assessment is marked as inconsistent and, as pictured in Figure 1, exits the model until intended and unintended consequences are considered. An example of an inconsistent claim might be to measure mathematics achievement over time in a common cohort of students. As TIMSS is a cross-sectional study, measuring change over time is outside its scope.

The second step in the process is to submit the consistent reason(s) for participation in a given ILSA to a logic argument. Figure 2 illustrates a logic model that is a derivative of the Toulmin (2003) method of reasoning. The process involves a claim supported by a warrant and additional evidence, which is often provided through the backing statement. In contrast, rebuttals provide counterevidence against the claim. From this reasoning process, some conclusion is drawn.

We provide a worked example to illustrate the use of the logic model. To start, we submit the consistent purpose as a *fact* or *claim* about what an ILSA program says the test can reasonably accomplish: *TIMSS can be used to provide national-level achievement data at the lower end of secondary education in Country A*. This claim is supported by the following warrant: *TIMSS is designed to assess achievement in mathematics at the eighth grade internationally*. The backing provides concrete evidence: *TIMSS uses a curriculum-based, internationally agreed-upon mathematics framework to develop, administer, and*

report on mathematics achievement in a nationally representative sample of eighth-grade students. A plausible rebuttal to the original warrant and backing might be the following: *TIMSS does not fully represent the eighth-grade curriculum in Country A*. The backing for this rebuttal could be that *TIMSS dedicates 30% of the assessment to algebra concepts. Only 10% of the curriculum in Country A is spent on algebra, thus there is a certain degree of mismatch, in that the emphasis on algebra is greater on TIMSS than in Country A's curriculum*. Next, some conclusion must be drawn about the validity of the claim in relation to the participating country's goals: Country A either agrees or disagrees that mathematics achievement can be measured at the lower end of secondary education by TIMSS, using the degree of overlap in the country's curriculum and the TIMSS coverage as a basis for this decision.

The result of the logic argument leads to the next step in the model, which asks to what extent claims are or are not supported in a given educational jurisdiction. As we note in the model, this can be determined by degrees. On the basis of the preceding example, it is up to the national stakeholders to decide if the mismatch in the emphasis on algebra in the TIMSS framework and their national curriculum is acceptable. A reasonable conclusion could be as follows: *The IEA claims that TIMSS can monitor national mathematics achievement at the lower end of secondary education in Country A are reasonably supported; however, there is a degree of mismatch between our national curriculum and the TIMSS framework*.

The final step in the process is to use the logic argument conclusion and associated decision regarding whether claims are supported to develop an account of intended and unintended consequences that can arise from participation in TIMSS. Some reasonable expected consequences could be as follows: (a) The mismatch between the national curriculum and the TIMSS framework on algebraic emphasis will come at the cost of knowing less about other national curricular emphases, such as probability or geometry; (b) given that Country A is interested in introducing algebra earlier in the curriculum, understanding where possible gaps exist through participation in TIMSS can assist in this process. An unintended consequence could reasonably be stated as follows: *Teachers might shift teaching time to better reflect the TIMSS assessment design*. As a result, national educational priorities will not be given the intended treatment in the classroom. Each of these consequences leads to a final *conditional consequential statement* (CCS) of the form noted in Figure 2: *If Country A participates in TIMSS at the eighth grade, then national stakeholders will be given an opportunity to look more deeply into gaps in algebra education, given that TIMSS emphasizes algebra more heavily than the national curriculum*.

Each consequence should be stated as a CCS or placed in a list of related statements. In a similar fashion, *inconsistent* claims should be analyzed in terms of expected and unintended consequences. For example, an expected consequence of participating in TIMSS could be that nothing can be learned about change over time in a cohort of students, given the cross-sectional nature of TIMSS. The final evaluation can be summarized and used to determine the degree to which the ILSA meets national goals and whether the intended and unintended consequences meet national expectations, acceptability, and levels of tolerance.

Concluding Note

In closing, we suggest that the model proposed in this report could provide participants and potential users of ILSA data with a framework for understanding and comparing the various national and international assessments and that could help guide participation in and appropriate use of the data resulting from these surveys. The framework can provide an opportunity to participants to reformulate their goals to better align with ILSA purposes, negotiate changes to the test, or refrain from participating. Identifying possible threats to validity and unintended consequences of test use could also help with devising action plans to address them. In the ideal case, our model can help national or subnational governments clarify their aims, which can be made publicly available in transparent and understandable ways. Prior to participation, governments could share the results of this evaluation and announce which aspects of their educational systems they seek to evaluate and why. This approach could help prepare both primary and secondary users to receive and interpret ILSA information as relevant to their aims.

We acknowledge, however, that the proposed process is a significant undertaking and that there is a general lack of concrete examples to guide stakeholders. However, we contend that such an evaluation is valuable and worthwhile, particularly given the growing influence and importance of international assessments and a desire to effectively use limited resources. In instances where countries can clearly articulate their reasons for participation and testing organizations clearly and consistently communicate the goals of a given test, the proposed model should provide a fruitful base from which to evaluate alignment.

We also highlight the importance of clarity with respect to test purposes for the primary stakeholders. These issues include the need for ILSAs to clearly articulate their evaluative goals, for clients to understand the extent to which the ILSA goals align with their educational goals, and, in cases of misalignment, for clients to think through an action plan to supplement ILSA scores with other measures. All of these actions should help mitigate possible unintended consequences arising from possible misalignments that might not be made explicit or be clearly identified by participating countries and/or ILSA developers. We suggest that these issues are ever more important in light of the diversity of participating jurisdictions. For instance, Wandall (2013) described that participants with different educational profiles might differ in their reasons for participating in ILSAs. An intent to make accommodations for all participants may stretch an assessment's capacity beyond reasonable boundaries. Moreover, it is unlikely that there will be a perfect match between the educational goals of any one country and the assessment content or design, which highlights the need for a cost–benefit analysis to identify the elements that are aligned and those that are less well aligned with a country's goals prior to using test results to inform educational reforms.

Furthermore, assessments are typically developed for one primary purpose; the desire to include more than one primary use can lead to validity threats. Such threats may jeopardize data use and score interpretations (Feuer, Holland, Green, Bertenthal, & Hemphill, 1999). Indeed, "When the same test is used for multiple purposes, the validity of the inference that can be drawn from the results may be jeopardized" (Feuer et al., 1999, p. 77). Given testing time constraints and the varying test administration cycles of ILSAs, the extent to which ILSAs can provide results to inform ongoing formative change is questionable. A more reasonable expectation could be to utilize test scores and contextual questionnaires to enhance the interpretation of test scores and use the data in conjunction with other assessments (administered more frequently at more local levels) to inform formative decisions.

We suggest that future studies seek to develop general guiding questions to support educational system stakeholders in charting a path forward in terms of current and future ILSA participation. Stakeholders may wish to conduct such analyses using an organizational scheme to help structure claims and intended and unintended consequences for various test components, such as background questionnaires, cognitive instruments, and their alignment with national and international goals. Additionally, studies should seek to identify foundational elements needed to engage in reasoned conversations with testing organizations so that national governments can better explain what ILSAs can do to better meet their needs. Examples could include applying our evaluative tool to educational systems to identify the types of actions diverse nations tend to implement to appropriately make sense of ILSA data and reasonably inform educational change. We also note that although we describe a model for use by educational system–level stakeholders, researchers might also benefit from such an evaluation. Our proposed model could serve as a means for determining whether a particular (or any available) ILSA is well suited for answering the research question at hand.

References

Backhoff, E. (2013). Validity issues in ILSA programs: Thoughts for developing countries. In M. Chatterji (Ed.), *Validity and test use: An international dialogue on educational assessment, accountability and equity* (pp. 233–249). Bingley, England: Emerald Group.

Breakspear, S. (2012). *The policy impact of PISA: An exploration of the normative effects of international benchmarking in school system performance* (Education Working Paper No. 71). Paris, France: OECD. https://doi.org/10.1787/5k9fdfqffr28-en

Chatterji, M. (2013). Insights, emerging taxonomies, and theories of action towards improving validity. In M. Chatterji (Ed.), *Validity and test use: An international dialogue on educational assessment, accountability and equity* (pp. 273–308). Bingley, England: Emerald Group.

Ercikan, K. (2009). Limitations in sample to population generalizing. In K. Ercikan & M. W. Roth (Eds.), *Generalizing in educational research: Beyond qualitative and quantitative polarization* (pp. 211–235). New York, NY: Routledge.

Feuer, M. J. (2013). Validity issues in international large-scale assessment programs: "Truth and consequences." In M. Chatterji (Ed.), *Validity and test use: An international dialogue on educational assessment, accountability and equity* (pp. 197–216). Bingley, England: Emerald Group.

Feuer, M. J., Holland, P. W., Green, B. F., Bertenthal, M. W., & Hemphill, F. C. (1999). *Uncommon measures*. Washington, DC: National Academy of Sciences.

Fitzpatrick, J. L., Sanders, J. R., & Worthen, B. R. (2011). *Program evaluation: Alternative approaches and practical guidelines*. Boston, MA: Pearson Education.

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

Hoplins, D., Pennock, D., Ritzen, J., Ahtaridou, E., & Zimmer, K. (2008). *External evaluation of the policy impact of PISA* (Report No. EDU/PISA/GB(2008)35/REV1). Paris, France: OECD.

International Association for the Evaluation of Educational Achievement. (2013). *TIMSS 2015 assessment frameworks*. Boston, MA: TIMSS/PIRLS International Study Center, Lynch School of Education, Boston College.

International Association for the Evaluation of Educational Achievement. (2016). *Brief history of the IEA*. Retrieved from http://www.iea.nl/brief-history-iea

Kamens, D. H., & McNeely, C. L. (2010). Globalization and international testing and national assessment. *Comparative Education Review, 54*, 5–27.

Matthews, A. (2017, March 16). Australian students behave badly, OECD report says. *ABC News*. Retrieved from http://www.abc.net.au/news/2017–03–16/australian-kids-behaving-badly-in-classrooms-says-oecd-report/8356506

Mislevy, R. J., Johnson, E. G., & Muraki, E. (1992). Scaling procedures in NAEP. *Journal of Educational and Behavioral Statistics, 17*, 131–154. https://doi.org/10.3102/10769986017002131

Nusche, D., Earl, L., Maxwell, W., & Shewbridge, C. (2011). *OECD reviews of evaluation and assessment in education: Norway*. Paris, France: OECD.

Oliveri, M. E., & von Davier, A. A. (2016). Psychometrics in support of a valid assessment of linguistic minorities: Implications for the test and sampling designs. *International Journal of Testing, 16*, 205–219. https://doi.org/10.1080/15305058.2015.1099534

Oliveri, M. E., & von Davier, M. (2017). Analyzing invariance of item parameters used to estimate trends in international large-scale assessments. In H. Jiao & R. W. Lissitz (Eds.), *Test fairness in the new generation of large-scale assessment* (pp. 121–146). Charlotte, NC: Information Age.

Organisation for Economic Co-operation and Development. (2005). *Teachers matter*. Retrieved from http://www.oecd-ilibrary.org/content/book/9789264018044-en

Organisation for Economic Co-operation and Development. (2010). *The high cost of low educational performance*. Retrieved from http://www.oecd-ilibrary.org/content/book/9789264077485-en

Organisation for Economic Co-operation and Development. (2015). *How's life? 2015*. Retrieved from http://www.oecd-ilibrary.org/content/book/how_life-2015-en

Pons, X. (2012). Going beyond the 'PISA shock' discourse: An analysis of the cognitive reception of PISA in six European countries, 2001–2008. *European Educational Research Journal*, *11*(2), 206–226.

Schleicher, A. (2012). *Use data to build better schools*. Retrieved from http://www.ted.com/talks/andreas_schleicher_use_data_to_build_better_schools

Sellar, S., & Lingard, B. (2013). The OECD and the expansion of PISA: New global modes of governance in education. *British Educational Research Journal*, *40*, 917–936. https://doi.org/10.1002/berj.3120

Takayama, K. (2008). The politics of international league tables: PISA in Japan's achievement crisis debate. *Comparative Education, 44*, 387–407.

Toulmin, S. E. (2003). *The uses of argument* (Rev. ed.). Cambridge, England: Cambridge University Press.

Wandall, J. (2013). Education, testing, and validity: A Nordic comparative perspective. In M. Chatterji (Ed.), *Validity and test use: An international dialogue on educational assessment, accountability and equity* (pp. 137–161). Bingley, England: Emerald Group.

White Plisko, V. (2013). Validity and international large scale assessment programs. In M. Chatterji (Ed.), *Validity and test use: An international dialogue on educational assessment, accountability and equity* (pp. 251–262). Bingley, England: Emerald Group.

Suggested citation:

Oliveri, M. E., Rutkowski, D., & Rutkowski, L. (2018). *Bridging validity and evaluation to match international large-scale assessment claims and country aims* (Research Report No. RR-18-27). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/ets2.12214

Action Editor: Rebecca Zwick

Reviewers: Irwin Kirsch