Improving International Assessment Through Evaluation

David Rutkowski, University of Oslo

The goal of this special issue was to focus on empirical, conceptual, and theoretical perspectives on reading assessment in international studies. In that regard, the editors where able to bring together top scholars in the field to share a range of viewpoints and findings. Rather than comment on each individual paper, which are of high quality, I would like to expand the discussion of the special issue and focus on what I see as an emerging issue for international large scale assessments (ILSAs) more broadly: the need for a focused discussion of the merit and worth of ILSAs in the 21st century. To do this, I will draw on some of the findings from the papers in this special issue as examples of the ways the field is moving and how I believe the authors were limited, not by their expertise, but by the nature of the data. To that end, a close reading of the papers reveals an emergent and cohesive theme that could be described as unique methods applied in an effort to expand the usefulness of ILSA data. As a result, this special issue highlights the need for a larger discussion, one that calls for a reexamination of ILSAs and their use. In what follows I will briefly outline examples of how I believe these papers speak to the limits of ILSA data. Subsequently, I outline how national participation in a meta-evaluation of ILSAs can assist study administrators and ILSA users in creating data that fits the needs of a complex and diverse set of participants.

Limitations

Although each author in this special issue directly addresses the limitations of their own work, I would like to address three themes that arose as I read the papers. The first has to do with the quality of background questionnaires; the second, which is also related to the background questionnaires, has to do with measuring trend over time; and the third has to do with the thorny issue of making causal claims with cross sectional ILSA data. I discuss each of these separately. However, all of these issues are deeply connected. And my purpose is not to reiterate the limitations of the studies but rather to suggest that these limitations speak to a growing need by the ILSA community to understand what changes should be made to better meet the needs of both participating governments and the research community.

Background Questionnaire.

In general, the background questionnaires have two primary uses within ILSAs: (1) to help contextualize the assessed educational system; and (2) to optimize population and sub-population achievement estimation. The benefits of using background data to help estimate achievement are well documented (Mislevy, Beaton, Kaplan, & Sheehan, 1992). However, the former purpose is of specific concern for many of the authors in this special issue. For example, Shepherd (special issue) was faced with less than optimal information (low response rates and missing data on select variables). And as correctly noted by Caro, Kyriakides, and Televantou (special issue), the cross cultural comparability of the constructs, such as home background variables, is not well understood. In fact, Walzebug, Kasper, and Wendt (special issue) directly address concerns about the Progress in International Reading Literacy Study (PIRLS) scale for home resources for learning (HRL) when they state "the prediction effect of the HRL is not only sensitive with respect to its content validity but also with respect to methodological assumptions that are made by scaling of this index" (p. 5). The difficulty of creating a universal background questionnaire that includes the most important variables for all participating countries should not be understated. Further, making significant changes to any aspect of the background questionnaire brings with it a set of challenges, with one of the most important being the loss of trend between cycles on variables that are modified. This leads me to the next limitation, that of estimating trend.

Trend.

Two of the three largest and longest running international assessments, PIRLS and the Trends in International Mathematics and Science Study (TIMSS), incorporate important design elements that ensure stable and reliable achievement trend over each cycle of the study. On the other hand, until improvements were implemented in the 2015 cycle, the Programme for International Student Assessment's (PISA's) design is less suited for stable trend estimates over adjacent cycles. Instead, the PISA design from 2012 and earlier allows for stable and reliable achievement trend estimates on a given subject every 9 years (see Rutkowski & Rutkowski, 2016). In regards to trends on non-achievement indicators and scales, such as family background, a number of important questions arise. For example, are the variables that were important indicators of wealth in 1995 the same in 2016? Do the students have the knowledge necessary to answer the questions (e.g. do they know their parent's income)? Can you legally or ethically ask particular questions in all participating systems (e.g. religion or race)? Of course, when study

programs adjust or omit background questions that were once collected, the ability to validly measure change over time on those constructs is challenged. Further, even something as simple as changing the response style to a given question (from a three to four point scale) can lead to problems in estimating change. For example, in this special issue, Lenkeit, Schwippert, and Knigge (this special issue) dummy coded many of the categorical variables in their analysis (losing statistical power and concealing nonlinearity) because response options differed across cycles. Again, there are clear tradeoffs between keeping the same questions over time or making changes (see L. Rutkowski, 2016). The specifics around particular tradeoffs are not the focus of the argument I wish to make here. Rather, I argue that in dialogue with their perspective research community, participating systems need to consider all changes made by the testing organizations to background questionnaires and provide constant feedback to test designers.

Causal Claims.

I, and others, have discussed the validity of making causal claims with ILSA data (D. Rutkowski & Delandshere, 2016; Schneider, Carnoy, Kilpatrick, Schmidt, & Shavelson, 2007). There is no clear consensus among researchers on the validity of using ILSAs to make causal claims; however, one area where most scholars seem to agree is that new and innovative designs will either enable and or strengthen that ability to make such claims. To that end, the OECD has included a goal in its longer term strategy for PISA of "strengthening the ability to draw causal inferences from the data" (OECD, n.d.). Some design suggestions by scholars in the field of ILSA call for the purposeful inclusion of instrumental variables (Pokropek, 2016), including causal questions into the study design and data collection (Kaplan, 2016), and as demonstrated in this special issue, the inclusion of prior achievement information from tested students to correct for omitted achievement bias (Caro, Kyriakides, & Televantou, this issue). Improving both the accuracy and validity of causal claims will take a great deal of resources and thought by participating countries and testing organizations. With these costs in mind, Kaplan (2016) recommends that those who participate in ILSAs "must first decide if addressing the effects of specific causes is a policy priority, and then to focus on a small set of priority causal questions" (online). Consider an example where policy makers want to understand whether the way that reading literacy is operationalized causes ILSA results to differ (see Solhein and Lundetrae, this special issue). In this situation assessing the same sample of examinees across PIRLS, PISA, and PIAAC would enable researchers to make stronger claims. Participating systems should consider

a cost versus benefit analysis of implementing design-based solutions for the purpose of answering causal questions. Such an analysis could be included in an evaluation of the merit and worth of national particpation in ILSAs. In the following section I outline a framework that may help initiate such an evaluation.

National Meta-evaluation of ILSAs

In what follows, I lay out a framework for taking a national meta-evaluative approach that may assist the developers of ILSAs to better meet the needs of its participants. Formal meta-evaluation in its simplest form is when the evaluation (in our case an ILSA), or parts of the evaluation process, are systematically evaluated based on an accepted set of criteria. Today, meta-evaluation, which can be fruitfully used to judge good and bad evaluations (see Stake, 1967; Stufflebeam, 1968) is well established in the field and has been included as an evaluation category by the internationally accepted Joint Committee on Standards for Educational Evaluation (see standards E2 and E3 in: Yarbrough, Shulha, Hopson, & Caruthers, 2010). By arguing for a meta-evaluation of international assessments, my goal is twofold. First, I aim to help ILSA participants understand goals/roles for ILSAs in the local context. By explicitly engaging in an evaluation of ILSAs at the system level, it one means for enabling national governments to determine whether ILSAs are meeting system-level goals and whether individual systems are well served by continued participation. Take, for example, participation of low performing systems such as Algeria in PISA. Low performance on the assessment simply tells policy makers in Algeria that student's do not know what is on the PISA assessment, which may or may not be relevant for their specific economic goals. Additionally, given the vast cultural and economic differences between Algeria and most OECD countries (PISA is tailored towards OECD countries) it may be that the background questionnaires are generally not meeting the needs of local policy makers. Meta-evaluation could assist the policy community from participating countries in understanding how ILSAs are informing their educational systems and if their needs are being met.

Second, if governments were to participate in a meta-evaluation that is concerned with their ILSA participation they would better understand the needs of the research community who often spend a great deal of time engaging with the data to inform both research and policy. More active participation from stakeholders (policy and research) in developing what should be collected and why may be key to improving some of the short-comings of ILSAs as

demonstrated by this special issue. To be sure, engagement in a meta-evaluation is not a silver bullet and cannot work in isolation; however, my aim is to put evaluative power into the hands of ILSA participants so that they can – at least to some degree – determine the merit and worth of these international evaluations in their local context and thus create clear suggestions of what is needed to improve ILSAs as they move forward.

Scriven's (2012) "higher level" Meta-evaluation Checklist (MEC) frames the remainder of this discussion. It is "higher level" in the sense that it does not attempt to provide a purely prescriptive checklist but rather makes use of five criteria[1] in order to explain how meta-evaluation can assist in the evaluation of ILSAs. The five criteria include: validity, credibility, clarity, propriety, and cost-utility. I contend that these five areas form a useful basis from which to begin a conversation with national governments on the value of commissioning an internal or external meta-evaluation. Although, deep coverage of Scriven's guidelines is beyond the scope of this discussion, I attempt to outline how these five criteria can assist countries in understanding the usefulness of ILSA participation and, perhaps more importantly, provide participating countries with the knowledge to have clear and frank conversations with testing organizations on what they specifically need from ILSAs so that they and their respective research communities have the data they need to best inform their systems. Interested readers would be well advised to engage fully with Scriven's MEC. For now, let us briefly take up each criterion in relation to ILSAs and national meta-evaluation.

Validity.

Although a commonly used term, validity is much discussed, often misunderstood, and variably defined in the research community. As in the social sciences, the field of evaluation validates knowledge claims as one criterion to judge quality. Schwandt, (2001) writes that validity is an "epistemic criterion: To say that the findings of social scientific investigations are (or must be) valid is to argue that the findings are in fact (or must be) true and certain" (p. 319). Scriven (2012) argues that validity is the *key* criterion for meta-evaluation, where validity represents "the matter of truth" (p.2). Here "true" implies that the findings of the evaluation under study are accurately represented. Scriven outlines a number of topics that should be

[1] Scriven also adds generalizability as possible sixth criteria but argues that generalizability is something on the periphery and not a requirement for judging merit or worth. As such, I do not explicitly include generalizability in the following discussion. But it should be recognized that ideas concerning generalizability are present in the five criteria.

addressed, separating validity within meta-evaluation into two key components: "rules of the game" and "probable truth of the conclusion." The first is focused on setting the purpose of the meta-evaluation. In other words, to arrive at valid findings for the meta-evaluation, the evaluator has to determine what kind of evaluation was originally required (e.g., why did the system participate). In our case, the meta-evaluator would need to work with national actors to understand reasons for ILSA participation and the particular aspects of ILSAs that should be (meta-)evaluated. For example, do national representatives prefer to focus on the ILSA's conclusions, process, impact, or all three? And should the meta-evaluation be summative or formative? During this initial step, the level of detail that is sufficient for the meta-evaluation findings should also be decided. This is especially important in the case of ILSAs. For example, do national governments simply want some idea of how ILSAs have influenced education or do they want irrefutable evidence? The latter is especially useful for decision making but requires much more effort and resources. Similarly, meta-evaluation can be used to judge the merit and worth of ILSAs conclusions and processes, but the detail that is required will determine necessary resources. Important here is that the nation who commissioned the ILSA makes the determination without guidance from the organizations administering ILSAs. In other words, nations need to independently decide "rules of the game" so that at the national level the merit and worth of ILSAs processes, conclusions, and/or impacts within a national context can be independently determined. To that end, advances in technology, models, and methods to deal with measurement heterogeneity have proven useful for improving local usefulness while still maintaining international comparability (see L. Rutkowski & Rutkowski, 2017). A critical component of validity, as noted by Scriven (2012), "is the matter of the probable truth of the conclusion(s)" (p. 2). This part of the meta-evaluation process sets to examine whether relevant standards are being met. Importantly, standards at the national level may differ from standards at the international level. For instance, although geometry might be regarded as a sub-domain of an international 8th grade math curriculum, this topic might not be covered in a particular country until grade 9. In this situation, conclusion validity is weakened with respect to interpretations of an overall mathematics score in countries where students haven't had an opportunity to learn the topic. This is but one example of how focusing on validity at the national level can differ from the international level. The critical point is that judging validity will differ between national and international meta-evaluations. In other words, it may be that the coverage and correctness at the

international level differs for national level interpretation. As a consequence, the adequacy of inferences that are provided to support the international conclusions may not be supported or even needed at the national level. Finally, reliability needs to be assessed at the national level. Within ILSAs, for example, reliability of scales might meet some international standard; however, a measure can fall short in a particular country. A prime example is PISA's economic, social and cultural status (ESCS) scale. As noted in Rutkowski and Rutkowski (2017), reliability of this scale differs greatly among countries, calling into question the validity of this (and subsequently other) scales. Another example is the books in the home variable in PIRLS. Rutkowski and Rutkowski (2017) show that the agreement between what parents and children report on this variable vary by country. When parents and children disagree on the number of books in the home, without further in-depth research, it is impossible to know who is answering the questions accurately calling into question reliability and validity of this important indicator.

Credibility.

Scriven (2012) writes "the focus here is on matters of credibility *not* covered by directly checkable validity considerations. Obviously, the big issues here are *independence* and *relevant experience*" (p. 4). With respect to the former, primary issues regard independence of the evaluators (e.g. IEA or OECD) from all participants in decision making and whether they are able to independently include and exclude participants' needs and desires in the evaluation process based on what is best for the overall evaluation rather than what is best for one participant. As an example, the OECD makes a distinction between OECD member countries and non-member countries (termed *partner countries*). It is reasonable to ask then, whether the OECD has an equally independent relationship with OECD and partner countries. Further, do partner countries have an equal say in decisions around the PISA design, implementation, and reporting relative to member countries? That is, do all countries have equal weight in PISA? Where there are differences, it is important to query the degree to which this might have an impact on the relevance of the study for a given country.

Relevant experience is also important to notions of credibility. It may be the case, for example, that the OECD has a great deal of relevant experience in evaluating workforce knowledge but little experience in evaluating education for social justice. In a similar vein, the IEA may have the most relevant experience in evaluating curriculum although they demonstrate little relevant experience in evaluating financial literacy. In either case, how we judge credibility

of the evaluators in the meta-evaluation is closely connected to the meta-evaluation's main focus determined during the first part of the validity conversation. Important for our proposal of employing meta-evaluations through a national lens is that national actors are empowered to ultimately understand credibility based on their needs and reasons for commissioning and joining the given ILSA.

Clarity.

Clarity is a combination of comprehensibility and concision. Scriven (2012) argues that an evaluation should work toward producing the most concise findings that also ensure readability and understandability by the intended audience. For national meta-evaluations, ideas around clarity may differ depending on the context of the country. For example, one country may have the resources and ability to analyze and interpret the often complex results from ILSAs, while others may not. Further, some countries may be facing a situation where generalized results and ILSA achievement rankings may be used for political purposes or to blame specific schools for poor results (something most ILSAs are not specifically designed to do). My point here is simple. Any nationally focused meta-evaluation of ILSAs should consider clarity of a given ILSA's process, conclusions, and/or impact in the context of the national system and its own capacity to properly interpret the data and provided results. As such, a meta-evaluation of ILSAs at the national level should explicitly focus on the clarity of the initial findings and evaluation reports from the assessing organizations.

Propriety.

Scriven (2012) defines propriety as "meaning ethicality, legality, and cultural/conventional appropriateness" (p. 5). For the most part, ILSAs follow stringent guidelines to adhere to ethical, legal, and cultural standards at the international level. Again, however, national contexts often differ and it is important that each nation fully consider this topic before, during, and after the ILSA is complete. Each stage brings a host of concerns. For example, an in-depth evaluation of the cultural appropriateness of ILSAs may differ greatly from nation to nation, depending on the context. Further, laws and legal structures are ever changing and what was legal in 1995 during the first release of TIMSS may differ from legal parameters today. Similarly, and relevant to many OCED countries, is how these assessments are culturally appropriate for the rapidly changing demographics faced by most countries.

Cost-utility.

Cost-utility includes being economical with the resources provided but also, and perhaps more importantly for the ILSA context, comprises a cost-benefit analysis, to include estimates for comparative cost-effectiveness. Estimating comparative cost-effectiveness may be made easier if meta-evaluations were completed for all major ILSAs and compared so that evaluators could see which is most cost-effective for national needs. That said, Scriven (2012) notes that the core idea of cost utility for a meta-evaluation is not to focus completely on cost-benefit analysis but to answer the following: "did the evaluation pay for itself…, or did it merely discharge an obligation…, or was it, de facto, an unnecessarily expensive gesture?" (p. 5). For example, it may be that participation in PISA is obligatory based on membership agreements to the OECD. In such a case this should be factored into any cost-utility analysis, especially when attempting to compare an ILSA to other studies. Under this criterion, a country may also consider if there are other more economical ways to produce the information they gain from participating in ILSAs. For example, does a given ILSA provide enough information to justify the frequency of administration? Or might it be more cost-effective to limit participation to every other cycle? Or is it advisable to urge the study architects to consider a longer lag between administrations? To that end, a national meta-evaluation might arrive at the conclusion that the knowledge gained from ILSA participation is worthwhile but that the cost of participating every three (or four) years outweighs the benefits gained by such frequent measurement. It is during this part of the meta-evaluation process that such questions can and should be analyzed.

Conclusion

Since 1995, which we could reasonably identify as the beginning of the modern era of international assessment, ILSAs have developed rapidly in terms of the number of assessment platforms, participating systems, measured domains, and assessed populations. In addition to monitoring the quality of education around the world, international assessments have also proven to be a useful resource for empirical researchers in education and beyond. As one example, this special issue exemplifies how researchers are exploring novel and innovative ways to better understand ILSA reading results. Indeed, the authors offer useful results with the potential to inform their respective fields. Nevertheless – as the authors are aware and unquestionably recognize, the degree to which ILSA data are suitable for definitively answering a given research question is necessarily limited. Changes to the surveys over time, the cross-sectional, observational nature of the data, and data quality are three such issues. Certainly these issues are

pervasive in much of the social sciences; however studies such as TIMSS, PISA, and PIRLS are somewhat unique in that they are recurring *evaluations* of the quality of education in dozens of participating systems. And the cyclical nature of these studies along with relatively short time span between studies (three to five years) necessitates that as these studies grow and change, tractable solutions must be found relatively quickly. As a result, there is little time to step back and take stock of the utility ILSAs for their intended stakeholders. But after more than 20 years of growth and development, such an evaluation is important.

It is reasonable to imagine findings from a meta-evaluation that point to a need on behalf of the research and policy community for prior achievement measures to strengthen knowledge claims (e.g. Caro, Kyriakides, and Televantou, special issue). Providing prior achievement is something that could be easily implemented in most participating systems and requires minimal work for either the testing organization or the national participant. Another finding from a meta-evaluation could be that ILSAs do not adequately contextualize achievement at the national level in certain systems. In such a situation, adding more questions about home resources, for example, could increase our understanding of how home reading resources are associated with achievement (Walzebug, Kasper, & Wendt, special issue). Although adding questions would seem to increase the response burden on examinees, well-developed design solutions can be brought to bear on this issue. PISA (2012), for example, employed a rotated background questionnaire, which allowed for a significant expansion of measured domains. Although this design was used in questionnaires for all countries, an alternative would be to use such an approach in a small subset of participating countries. A key example of the potential usefulness of an expanded questionnaire is Shepherd (special issue). In particular, South Africa and Botswana could have included a set of important language policy indicators to implement appropriate scales as national options. To capitalize on economies of scale, the two southern African countries could have collaborated to create a useful measure while sharing the cost of development. Similar arguments can be made for each limitation noted by any of the authors in this special issue.

To that end, my comment on this special issue was intended to illustrate meta-evaluation is one possible solution to understanding whether and how ILSAs are meeting the needs of policy makers and researchers. Although the stakeholders that are most relevant with respect to this special issue are empirical researchers and their readership, the approach I outline here is

suitable for a broad spectrum of ILSA developers and consumers. Take as one example the issue of causality in international assessment. This is an issue that can certainly be attended to via design solutions, including collecting prior achievement of participating students or through the thoughtful development and inclusion of questions geared toward answering targeted causal questions. Similarly, the utility and quality of background questionnaires can be enhanced through solutions like rotation that, when carefully implemented, make it possible to collect more information without overburdening participants. But well-designed changes to ILSA rely on understanding the ways in which these studies are or are not meeting stakeholders' needs. Finally, it's important to recognize that modifying these studies to better suit some needs can come with significant trade-offs, such as compromising trend estimates. To paraphrase an old adage, there is no free lunch; but a careful meta-evaluation offers the opportunity to know if the menu is satisfactory, who has allergies, and what people would like as the daily special.

References

Kaplan, D. (2016). Causal inference with large-scale assessments in education from a Bayesian perspective: A review and synthesis. *Large-Scale Assessments in Education*, *4*(1).

Mislevy, R. J., Beaton, A. E., Kaplan, B., & Sheehan, K. M. (1992). Estimating population characteristics from sparse matrix samples of item responses. *Journal of Educational Measurement*, *29*(2), 133–161.

OECD. (n.d.). Beyond PISA 2015: A longer-term stategy of PISA. OECD. Retrieved from http://www.oecd.org/pisa/pisaproducts/Longer-term-strategy-of-PISA.pdf

Pokropek, A. (2016). Introduction to instrumental variables and their application to large-scale assessment data. *Large-Scale Assessments in Education*, *4*(1), 4.

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

Rutkowski, L. (2016). A look at the most pressing design issues in international large-scale assessments. U.S. National Academy of Education. Retrieved from https://naeducation.org/wp-content/uploads/2016/12/Pressing-Methodological-Issues-in-International-Assessment-Rutkowski-2016_web-version.pdf

Rutkowski, L., & Rutkowski, D. (2016). A call for a more measured approach to reporting and interpreting PISA results. *Educational Researcher*, *45*(4), 252–257.

Rutkowski, L., & Rutkowski, D. (2017). Improving the comparability and local usefulness of international assessments: A look back and a way forward. *Scandinavian Journal of Educational Research*, *0*(0), 1–14.

Schneider, B., Carnoy, M., Kilpatrick, J., Schmidt, W. H., & Shavelson. (2007). *Estimating causal effects using experimental and observational designs*. Washington, DC: American Educational Research Association.

Schwandt, T. A. (2001). *The Sage dictionary of qualitative inquiry*. Thousand Oaks, CA: Sage.

Scriven, M. (2012). *Evaluating evaluations: A meta-evaluation checklist*. Retrieved from http://michaelscriven.info/images/EVALUATING_EVALUATIONS_8.1.11.pdf

Stake, R. (1967). The countenance of educational evaluation. *The Teachers College Record, 68*(7), 523–540.

Stufflebeam, D. L. (1968). Evaluation as enlightenment for decision-making. Retrieved from http://eric.ed.gov/?id=ED048333

Yarbrough, D. B., Shulha, L. M., Hopson, R. K., & Caruthers, F. A. (2010). *The Program Evaluation Standards: A Guide for Evaluators and Evaluation Users: A Guide for Evaluators and Evaluation Users*. SAGE Publications.