Discussion of

A Look at the Most Pressing Design Issues in International Large-Scale Assessments

by Leslie Rutkowski

David Kaplan

Department of Educational Psychology

THE UNIVERSITY of WISCONSIN MADISON

NAEd Workshop Series on Methods and Policy Uses of International Large-Scale Assessment
June 17th, 2016, Washington, DC

- I am very pleased to have the opportunity to discuss this is a very interesting and important paper that reminds us of several essential facts of ILSAs
  - That ILSAs are constantly evolving under the selective pressures of changes in the policy landscape and innovations in methodology.
  - That ILSAs, as such, are imperfect, but represent the present state-of-the-art in survey sampling, adaptation and translation, and psychometrics.
  - That constant improvements are necessary, desirable, and can extend the boundaries of what is possible with ILSA data.

- Leslie's paper outlines three areas of methodological concern
  1. Cross-cultural comparability in the achievement test and the context questionnaire.
  2. Measurement error in the context questionnaire and specifically in key reporting variables
  3. Causal inference within the ILSA design framework.
- My brief discussion will be to simply reinforce and add a small bit of gloss to Leslie's major points.

Cross-cultural comparability

- It is virtually axiomatic that the believability and acceptance of ILSAs by the policy makers and the public rest on being able to verify cross-cultural comparability.
- This problem dominates internal technical discussions among contractors and experts charged with fielding an ILSA.
- A considerable amount of work on this problem occurs prior to the launch of the main study during the *field trial* and translation/adaptation stage of the ILSA, but main study official reporting often provides information on cross-country comparability after the main study data are collected.

- The problem is that conventional approaches to assessing cross-cultural comparability of items on the test and questionnaire necessitate statistical restrictions that are rarely supported by the data – so called invariance tests with so-called scalar invariance require to report mean differences.
- New developments in statistical methodology allow different approaches to addressing this problem.
- The OECD and the IEA have launched serious investigations into this problem – particularly for the context questionnaire.
- **Shameless plug warning**: See volume (forthcoming) by Kuger, Kleime, Jude, & Kaplan which emphasizes the PISA field trial.

- Leslie's focuses attention on cross-cultural comparability in the test design and illuminates three studies that show that allowing some cross-cultural specificity in the test design can lead to better data-model consistency.

- PISA 2009 and 2012 allowed countries to choose an option of having somewhat easier items, thus providing a less frustrating experience for the students as well being able to more fully describe achievement spectrum across countries.

- However, Leslie reports results showing that there are ceiling effects insofar as some very high performing countries are hitting the ceiling on some items. She therefore suggests that much harder items should be added to the assessment, again to cover the spectrum of country-level achievement.

- As Leslie notes, this requires a new matrix sampling design. Such a design could be tried out in the FT, where one could experiment with different matrix sampling configurations. Implications for the estimation of plausible values would also have to be assessed.

Improving Measurement of Key Reporting Variables

- ILSA's and national assessments such as NAEP routinely report achievement scores across policy relevant categories such as race/ethnicity, gender, and SES.
- An important point made in Leslie's paper concerns problems in the measurement error in key reporting variables.
- Leslie provides one example concerning the measurement and meaning of "grade repetition"
- Another example I would like to provide that underscores Leslie's point is the PISA measure of education, cultural, and social status (ESCS)

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

1. **Family wealth**: Based on the students' responses on whether they had the following at home: a room of their own, a link to the Internet, a dishwasher (treated as a country-specific item), a DVD player, and three other country-specific items; and their responses on the number of cellular phones, televisions, computers, cars and the rooms with a bath or shower.

2. **Cultural possessions**: Based on the students' responses to whether they had the following at home: classic literature, books of poetry and works of art.

3. **Home educational resources**: Based on the items measuring the existence of educational resources at home including a desk and a quiet place to study, a computer that students can use for schoolwork, educational software, books to help with students' school work, technical reference books and a dictionary.

- Family wealth, cultural possessions, and home educational resources are scaled under an assumption that there is some single underlying latent variable.
- After estimates are provided, they are combined with the other variables into a PCA which then forms the single ESCS measure.
- So here also, we see a problem with measurement as discussed in Leslie's paper – namely that it is hard to justify a underlying latent variable generating a tick-off list of possessions. Yet, such a model is used and is part and parcel of the ultimate component that defines ESCS.
- Much discussion has taken place regarding better ways to get at SES, such as the 2012 expert panel report to NCES co-authored by our own Bob Hauser, Hank Levin, and Margaret Beale Spencer, among others. There is still a great deal of work and trial implementation required. Here again, the FT is the best place to try out different measurement methods.

- Finally, Leslie discusses the issue of causal inference with ILSA data.
- Her paper points out the very real challenges in drawing causal inferences with ILSA data and points (graciously) to my work, building on Rubin's ideas that causal inferential questions must be built into the design of the ILSA from the start.
- The causal field (what Leslie refers to as the context of the causal question), requires the selection of enough relevant covariates to help warrant the assumption strong ignorability.

- Mapping the causal field and selecting relevant covariates is a challenging problem.
- The difficult lies in the limited real estate and time available for administering the context questionnaire – precisely the place where the causal question(s) and relevant covariates are located.
- Two options are available to address this operational problem but both have been routinely rejected:
  1. Increase the amount of time (and therefore space) available to administer the context questionnaire.
  2. Within the current space/time constraints, matrix sample the context questionnaire items as is done with the achievement test.

- Another approach discussed by Leslie to improve causal inferences is to somehow add a longitudinal component to the overall framework.
- Regarding the longitudinal component, Leslie's suggestion regarding how TIMSS data could be used is quite intriguing.
- The idea is to create a synthetic longitudinal component via a cohort assumption.
- The idea is that for a given cycle of TIMSS, the eighth grade students are randomly equivalent to the fourth grade students, four years later.
- Random equivalence is an assumption, but it is not unreasonable, it is testable, and perhaps manageable through the careful use of matching algorithms.
- Nevertheless, given TIMSS cycles were not designed explicitly with a causal question in mind, the ability to try out Leslie's interesting suggestion is to try to manufacture a policy relevant causal question after the fact.

- So, is causal inference possible in the context of ILSAs?
- Others will disagree, but my view is that it is possible as long as it was part of the design of the ILSA from the beginning. In other words, ILSAs would have to be reconceptualized with the goal of causal inference in mind. And even then, it would be hard.
- Perhaps new ILSA programs could be developed around this area, but the current crop of ILSAs are not suited for this purpose.
- Rather, ILSAs in their current form were conceptualized and are designed as monitoring/indicator systems and our focus should be on how to improve their utility for this purpose - both on the methodological side and on the reporting side.
- Drawing causal inferences with ILSA data as currently designed is misleading at best.

- So to conclude, Leslie's paper is a reminder that although the methodology of ILSAs is sophisticated and rigorous, there are a large number of open questions informed by the demands of policy as well as new methodological/statistical research.
- Many of us who are heavily involved with ILSAs know full well how the bratwurst is made, so Leslie's paper is important insofar as itt reminds those outside the bratwurst factory that ILSAs are an evolving technology and that inferences and policies should recognize the strengths and current limitations of ILSAs.
- In the meantime Leslie's work pushes those of us on the inside to try out, and perhaps implement, new methodological ideas.

- And in a plea for more research support, international organizations such as the OECD and IEA, as well as national organizations such as NCES should expand their infrastructure to allow for long term methodological r & d. The IEA has this, in part through their Research and Analysis Division at the Data Processing and Research Center in Hamburg.
- The AERA Research Grants Program is another wonderful opportunity to engage in such research.
- Nevertheless, there should be more time and resources provided for methodological studies of the sort that Leslie discusses, so as to continually improve the validity and utility of ILSAs.

THANK YOU