Overview of Research Topic

Archival data from international large-scale assessments (ILSA) have opened up unique opportunities for examining cross-national trends in equality and equity in educational outcomes. Quasi-longitudinal studies taking advantage of the rich accumulation of ILSA data dating back to the 1960s (Chmielewski, 2019; Majoros et al., 2021) have not only facilitated countries with weak national evaluation or monitoring systems to make comparisons over time, but also enabled comparative studies for all participating states (Salmela-Aro & Chmielewski, 2019; Strietholt & Rosén, 2016) to analyse the impact of specific policy decisions on educational outcomes such as academic de-tracking and expanding education assess. The dual benefits of providing equitable access to policy insight and methodological advancement made systematic studies using historic ILSA data particularly attractive for both policy-makers and educational researchers.

Comparing studies administered decades apart, however, is logistically challenging. Concerns about possible coding errors, for example, is impossible to verify (e.g., FIMS 1964 US sample, Table 1 in Majoros et al. (2021)) due to limitation on early bookkeeping. Secondly, since it is not possible to define a target population balancing both age and grade due to differences in school entry ages across countries, switching target populations from age-based sampling (First International Mathematics Study (FIMS) 1964) to grade-based sampling (Second International Mathematics Study (SIMS) 1980, and subsequent Trends in International Mathematics and Science Study (TIMSS) from 1995 onward) introduced another layer of complexity for comparing different test waves and between ILSAs (Programme for International Student Assessment (PISA) uses age-based sampling). Thirdly, it is questionable whether the measured constructs carried the same meaning over several decades. This problem was particularly acute in early ILSA attempts with some critiquing FIMS for not being a study of mathematics education, but a study of schools and schooling with mathematics serving as a proxy for achievement (Husén (1967), as cited in Majoros et al. (2021)). Lastly, the inherent and stable differences between nation states questioned whether comparisons involving multiple countries is to "compar[e] the incomparable" (Husén (1983), as cited by Kaiser (1999), p.3, then by Majoros et al. (2021), p. 74). All these concerns highlighted the consensus that meaningful comparisons require comparable data.

Key Concepts

SES Achievement Gaps

One fruitful line of enquiry using archival ILSA data linked students' socio-economic status (SES) with their achievement gaps. Using a single-country study design and parental education as the SES measure, Salmela-Aro and Chmielewski (2019) observed a gradual reduction in SES achievement gaps in Finland between 1950s and 1980s but a clear reversal afterwards (Fig 9.2, p. 160). Chmielewski (2019) extended the SES measures using three different variables (a) parental education, (b)

parental occupation, and (c) number of books in the household and reported comparable magnitudes of achievement gaps (50, 55 and 40 percent, respectively) across most participating countries. For the purposes of quantifying achievement gaps, existing literature widely adopted the percentile-based approach (Reardon's (2011) method, as cited in Salmela-Aro and Chmielewski (2019)) in which *calibrated* score difference between the 90th and the 10th percentiles of the student cohort were used as an operationalisation for achievement gaps.

Calibration Criteria

In their respective studies, Strietholt and Rosén (2016) and Majoros et al. (2021) both followed Kolen and Brennan (2004, 2014) criteria for evaluating the degree of similarity between tests: inferences, populations, constructs, and measurement characteristics/conditions—referring to low- vs high-stakes tests, age- vs grade-based sampling, terminology shifting, and identical vs complex matrix test designs, respectively.

Causes for Missing Data

When dealing with missing data, Majoros et al. (2021) and Strietholt and Rosén (2016) distinguished three types of missing mechanisms in their studies: not-administrated, omitted, and not-reached items. Not-administrated items were treated as true missing data in estimating item and student parameters while omitted items were treated as incorrect responses in order not to award students for skipping an item. Not-reached items involved more professional judgement with consensus leaning towards recoding them as missing.

Document-type Reading Tasks

Some cycles of reading ILSAs (e.g., RLS-1991/2001) contained unique tasks involving locating information from structured document such as non-continuous tables, chars, graphs, maps or bus timetables. These tasks were later demonstrated to have introduced additional sources of variances into the tests (Gustafsson and Rosén (2006), as cited in Strietholt and Rosén (2016)). Resultantly, all document-type reading tasks were excluded during calibration for the interest of preserving maximum comparability.

Potential Research Questions

Chmielewski (2019) proposed the following research questions:

1. whether increasing SES achievement gaps are a global phenomenon,
2. whether some countries have avoided the trend, and
3. whether increasing SES achievement gaps can be explained by changing educational and social policies and conditions.

Salmela-Aro and Chmielewski (2019) explored the possible drivers behind the initial decline between 1950s and 1980s, following by subsequent increases in SES achievement gaps in Finland. Strietholt and Rosén (2016) and Majoros et al. (2021) both studied in-depth the calibration procedure linking old and recent ILSA data sets.

In addition to mathematics, reading and scientific literacy, interests in "21-Century skills" started to gain momentum in recent decades. Regular surveys of

these emerging skills and literacies such as PISA's financial literacy and global competency are also building up their rich libraries of data sources. It is natural to enquire whether the procedural insight gained from multi-decade of traditional literacies applies equally well to 21-Century skills. Potential research question may involve

1. whether SES achievement gaps also appeared in 21-Century skills such as financial literacy,

2. are there any country or country-group (emerging vs industrialised economies) differences in SES achievement gaps in new literacies, and

3. which macro- and microeconomic factors were strongly associated with such systematic variations in SES achievement gaps.

Methodological Approaches

IRT Approach to Calibration

Majoros et al. (2021) and Strietholt and Rosén (2016) applied similar item response theory (IRT) calibration procedures for linking different waves and types of ILSA data sets. Multiple-choice items and dichotomous constructed responses (i.e., 1 mark only with no partial credit given) were subject to 3- and 2-parameter logistic (3PL, 2PL respectively) IRT models to ascertain their difficulty levels. Polytomous items worth 2 or more marks underwent IRT modelling using (generalised) partial credit models (PCM or GPCM) for the same purpose. Standardised marks were then re-scaled to have means 500 and standard deviations 100. Table 3 in Strietholt and Rosén (2016) (p. 11) illustrated the effects of this calibration process by comparing the original and IRT scores.

Other Methodological Considerations

Missing data are the norm rather than the exception in empirical studies, particularly during syntheses of ILSA data sets. Existing publications applied multiple imputation (MI) by iterative chained equations and pulled the five plausible values together following Rubin (1987)'s rules. Recent advancement in both MI theories and software power opened up more options for missing data treatment such as Mplus's unrestricted variance-covariance model using Bayes estimators. The Bayesian procedure may also complement the bootstrap approach to standard error computation employed by existing literature. Mplus's recent upgrade (Version 8.5 and 8.6) combining with hardware infrastructure up to 64-core parallel processing reduced hierarchical growth curve model computation time from days to hours, the multilevel model used in Chmielewski (2019), greatly accelerating incremental model building.

References

Chmielewski, A. K. (2019). The global increase in the socioeconomic achievement gap, 1964 to 2015. *American Sociological Review*, *84*(3), 517–544. https://doi.org/10.1177/0003122419847165

Majoros, E., Rosén, M., Johansson, S., & Gustafsson, J.-E. (2021). Measures of long-term trends in mathematics: Linking large-scale assessments over 50 years. *Educational Assessment, Evaluation and Accountability*, *33*(1), 71–103. https://doi.org/10.1007/s11092-021-09353-z

Salmela-Aro, K., & Chmielewski, A. K. (2019). Socioeconomic inequality and student outcomes in Finnish schools. In L. Volante, S. V. Schnepf, J. Jerrim, & D. A. Klinger (Eds.), *Socioeconomic inequality and student outcomes: Cross-national trends, policies, and practices* (pp. 153–168). Springer. https://doi.org/10.1007/978-981-13-9863-6_9

Strietholt, R., & Rosén, M. (2016). Linking large-scale reading assessments: Measuring international trends over 40 years. *Measurement: Interdisciplinary Research and Perspectives*, *14*(1), 1–26. https://doi.org/10.1080/15366367.2015.1112711