UNIVERSITY OF GOTHENBURG

Linking recent and older IEA studies on mathematics and science

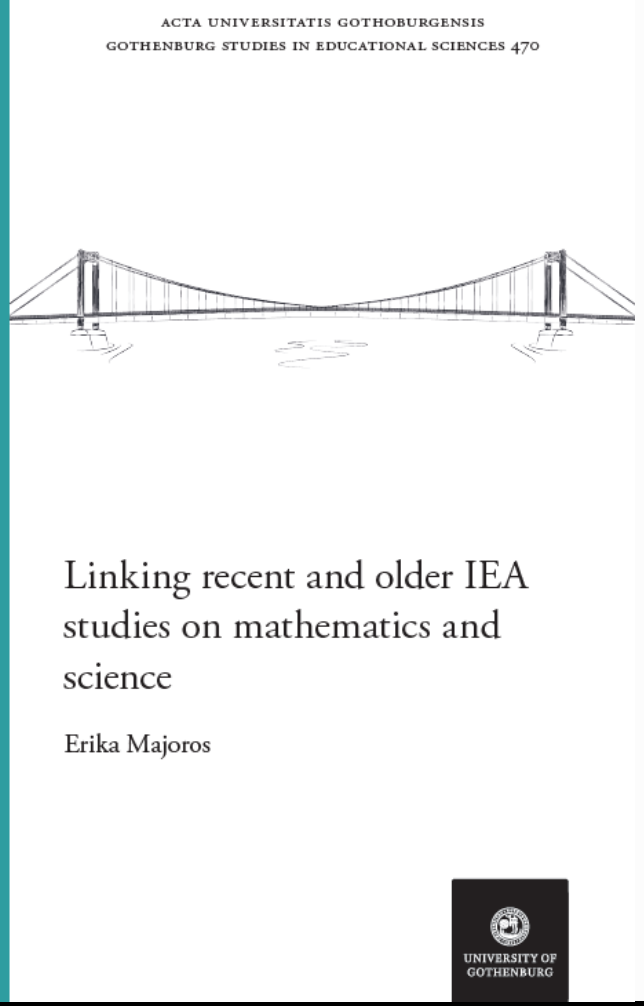
IEA General Assembly meeting
2023.09.27

Erika Majoros
erika.majoros@gu.se
Researcher/Lecturer
Department of Education and Special Education
University of Gothenburg, Sweden

The thesis project

www.etn-occam.eu
This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement no. 765400.

- The thesis is part of the OCCAM project
- The acronym OCCAM stands for “Outcomes and Causal Inference in International Comparative Assessments” in educational research
- OCCAM is a European Training Network (ETN) which is a sub call in the Marie Skłodowska-Curie Innovative Training Networks (MSCA ITN) of the European Commission’s Horizon 2020 framework
- OCCAM involved 15 PhD projects
- I have been working at
  - The Prerequisites of Educational Results (FUR) research environment at the University of Gothenburg (GU), Sweden
  - Educational Testing Service (ETS), Princeton, NJ, USA
  - International Association for the Evaluation of Educational Achievement (IEA) Hamburg, Germany



Linking recent and older IEA studies on mathematics and science

https://gupea.ub.gu.se/handle/2077/71965

- Supervisors: Monica Rosén and Jan-Eric Gustafsson
- The purpose of this thesis was to develop procedures that allow researchers to make reasonable comparisons of grade-eight mathematics and science achievement and motivation scales over a long time period, despite changes to the instruments, populations, and procedures between administrations
- The scales achieved combined with powerful analytical approaches such as country-level longitudinal modeling techniques and advanced econometric methods allow for investigating changes in educational systems
- The was guided by two overarching research questions
  1. To what extent are the student outcomes comparable across the first- and second-phase IEA assessments on mathematics and science?
  2. How do different linking approaches influence the descriptions of the system-level trends?

Aims of the thesis

- Student data were used from the Trends in International Mathematics and Science Study (TIMSS) and its four predecessors conducted before 1995, administered by the IEA
- The thesis aimed at
  1. Evaluating the degree of comparability of outcomes across these assessments
  2. Linking the cognitive test results onto the TIMSS reporting scale with the use of item response theory (IRT) modeling
  3. Exploring the feasibility of linking the motivational scales in these assessments with different approaches in the IRT and structural equation modeling (SEM) frameworks

Studies

Study I: 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

Study II: Majoros, E. (2023). Linking the first- and second-phase IEA studies on mathematics and science, *Large-scale Assessments in Education 11*(14). https://doi.org/10.1186/s40536-023-00162-y

Study III: Majoros, E., Christiansen, A., & Cuellar, E. (2022). Motivation towards mathematics from 1980 to 2015: Exploring the feasibility of trend scaling. *Studies in Educational Evaluation*, *74*, 101174. https://doi.org/10.1016/j.stueduc.2022.101174

Data

- The empirical work in this thesis was based on data of the populations representing 13-year-olds (FIMS and SIMS), 14-year-olds (FISS and SISS), and eighth-grade students (TIMSS cycles)

IEA ILSAs on mathematics and science administered before 1995

Assessment	Time of data collection	Number of participating educational systems
First International Mathematics Study (FIMS)	1964	12
First International Science Study (FISS)	1970-71	17
Second International Mathematics Study (SIMS)	1980-82	20
Second International Science Study (SISS)	1983-84	24

Evaluating the comparability of the outcomes

- The substantive basis of the three empirical studies lies in the evaluation of the extent of similarity across survey administrations
- The degree of similarity across assessments to be linked determines the “utility and reasonableness” (Kolen & Brennan, 2014, p. 498) of linking
- Four criteria for evaluating similarity (Kolen & Brennan, 2014)
  1. The goals need to be evaluated concerning the types of *inferences* drawn from the tests to be linked
  2. The alignment between the target *populations* of the assessments to be linked needs to be scrutinized
  3. The similarity of the measured *constructs* is to be evaluated
  4. The *measurement conditions*, such as test length, test format, and administration need to be scrutinized

Evaluating the comparability of the outcomes cont'd

- The achievement tests have maintained a set of common items (*bridge items*) between consecutive administrations, and these sets of common items serve as anchor tests between assessments
- Bridge item behavior was investigated across assessments with the delta plot method (Angoff & Ford, 1973; Magis & Facon, 2014)
  - To identify differential item functioning (DIF) among dichotomously scored items
  - The (transformed) proportion of correct answers (test items) or responses indicating positive endorsement (questionnaire items) is compared between the reference group and the focal group
  - It is a not computationally intensive method
  - It is a relative DIF method, i.e., bridge items were evaluated concerning all items comprising the bridge

Evaluating the comparability of the outcomes cont'd

- The cross-cultural comparability of affective constructs was evaluated by applying multiple-group confirmatory factor analysis (MGCFA) for each time point
  - This approach was chosen based on the suggestion by Meade and Lautenschlager (2004) that CFA is theoretically preferable over IRT methods when the number of items is small
  - The questionnaire items were treated as categorical variables and the students were grouped by country
  - The first step was to identify the baseline model and test for configural invariance among countries; after establishing configural invariance, threshold invariance was tested, followed by invariance testing for factor loadings (Svetina et al., 2020; H. Wu & Estabrook, 2016)

Linking approaches
Study I

- Using the pooled data of four countries that participated at each time point from FIMS to TIMSS 2015, the linking procedure involved three main steps

1. Different IRT models were tested to select the best fit, which were the two-parameter logistic model (2PL) for dichotomous items, i.e., multiple-choice items and constructed-response items for one score point, and the generalized partial credit model (GPCM) for polytomous items, i.e., constructed-response items for two or more score points
2. The item parameters were estimated via concurrent calibration, to which each country contributed equally by applying senate weights (stratum weights in SIMS were rescaled to sum to 500; and there were no weight variables in the FIMS 1964 datasets so individuals within a country were weighted equally)
3. Five plausible values (PVs) were drawn per student using the expected a-posteriori method; each PV was transformed to a metric with a mean of 500 and a standard deviation of 100 points across time

Linking approaches
Study II

1. In the four-country-all-time approach, previously (in Study I) estimated item parameters were used to fit a model on the pooled data of all countries
   - First, the test-takers' abilities were estimated separately for FIMS, SIMS, and TIMSS 1995, by fixed item parameters and drawing five plausible values
   - The distribution of the five PVs estimated for TIMSS 1995 was matched with the distribution of the reported TIMSS 1995 PVs by transformation constants, then the FIMS and SIMS scores were transformed
2. In the first-second-time approach
   - Item calibration by the concurrent calibration of FIMS and SIMS data of all countries combined with fixed item parameters of the bridge items to the values reported for TIMSS 1995 (3PL, 2PL, GPCM)
   - The student abilities were estimated separately for FIMS and SIMS, drawing five PVs per test-taker; and to locate the student ability estimates on the TIMSS reporting scale, the original transformation constants used for the reported TIMSS 1995 scaling were applied

Linking approaches
Study III

- Three methods were explored for constructing longitudinal affective scales

1. An IRT approach (GPCM)
   - Item parameter estimation and person scoring was conducted by concurrent calibration for each motivation scale on a pooled sample composed of data from all countries and cycles;
   - The person scores were transformed onto a scale with a mean of five and a standard deviation of one

2. A CFA approach
   - A CFA model was fit for each motivation scale on a pooled sample composed of data from all countries and cycles
   - Strong invariance of the anchor items across countries and over time was assumed
   - Factor scores were estimated applying maximum likelihood estimation with robust standard errors (MLR), while the items were treated as categorical variables
   - The factor scores then were transformed onto a scale with a mean of five and a standard deviation of one

Linking approaches
Study III cont'd

3. A market-basket approach
   - The market-basket approach assumes that all the items across the time points, related to intrinsic and extrinsic motivation towards mathematics, define each construct and can be considered as a market basket of representative items; and that the missing responses occur as a consequence of changes in the questionnaires across cycles
   - A measurement model per country was employed to generate plausible responses that fill the missing responses following the procedure suggested by Zwitser et al. (2017)
   - The measurement model was the GPCM model for consistency with the results from the IRT approach and the TIMSS procedure for linking contextual scales
   - Using the item parameters estimated by fitting the measurement models, missing responses were imputed five times per respondent
   - Then individual sum scores were calculated, thereby estimating five plausible scores per student; and the plausible scores were transformed onto a scale with a mean of five and a standard deviation of one

Summary of the results

- **Comparability of the outcomes of the assessments**
  - A high level of stability concerning the inferences and measured constructs among the assessments was found
  - The changes in the sampling and test conditions introduced challenges to the linking
  - Some of these challenges were handled to achieve a sufficient degree of similarity across the assessments
  - The rest remain as limitations of the scales
- **Bridge items**
  - In **Study I**, three items were flagged for DIF in the first two bridges, i.e., from FIMS to SIMS, and from SIMS to TIMSS 1995, respectively; no DIF items were detected in the rest of the bridges; the items showing DIF were excluded from the calibration

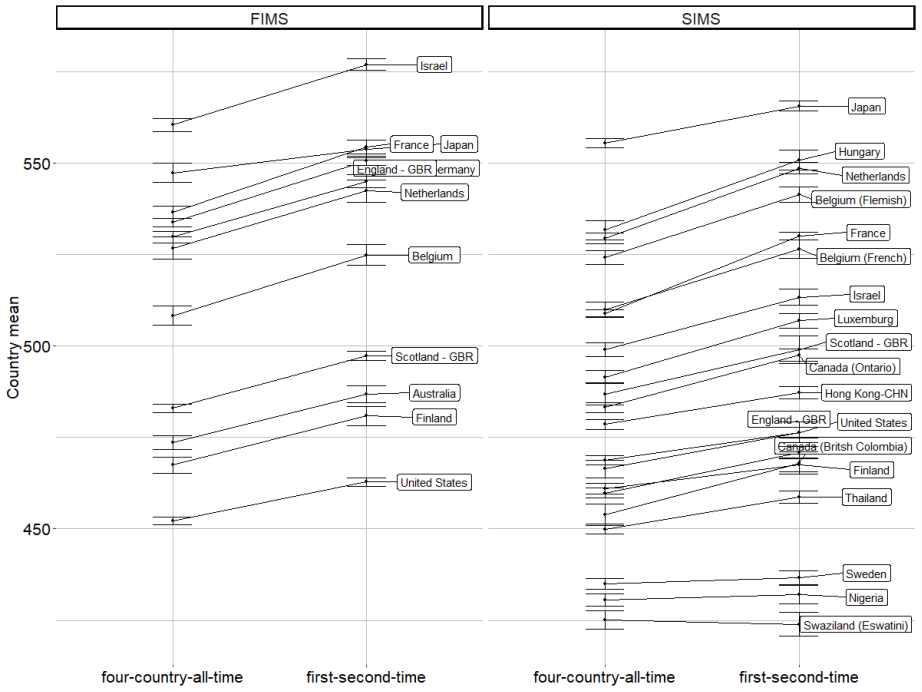
Summary of the results cont'd

- In **Study II**, the delta plot method was applied for the six sets of bridges: common items in the mathematics assessments between 1964-1980 (bridge 1), 1964-1995 (bridge 2), and 1980-1995 (bridge 3), and across the science surveys between 1970-1984 (bridge 4), 1970-1995 (bridge 5), and 1984-1995 (bridge 6); two items in the first, one item in the third, and two items in the fourth bridge were flagged for DIF; in the final, first-second-time linking procedure, these items were treated as unique items
  - In addition, Pearson's correlations were calculated the performance on the anchor test and the whole test; these correlations were moderate or high
- In **Study III**, the delta plot method was applied for each bridge between consecutive time points; for each country separately as well as the pooled data and all these tests yielded no items flagged for DIF
  - The MGCFA results for SIMS revealed that measurement invariance did not hold for Japan and all further analyses in this study were continued excluding data from this country.; the threshold and loadings equality constraints yielded an acceptable model fit at most time points for the five-country multiple-group model

Summary of the results cont'd

Trend descriptions by linking approach, mathematics achievement

- More item responses were used for the item calibration in the first-second-time approach than in the four-country-all-time approach; this implies more precision of the item parameters
- The item calibration is based on data from four educational systems in the four-country-all-time approach, while in the first-second-time approach, data from countries participating in FIMS, SIMS, and TIMSS 1995 were all used, a total of 50 countries; since in the IRT framework, item statistics are independent of the sample from which they were estimated, the differences in the samples should not influence differences in the scores
- In the first-second-time approach, a guessing parameter was included in the IRT model for multiple-choice items; the systematic difference seems to indicate that the IRT modeling mattered in the score estimation differences
- The rank order of the countries shows no difference in the two approaches



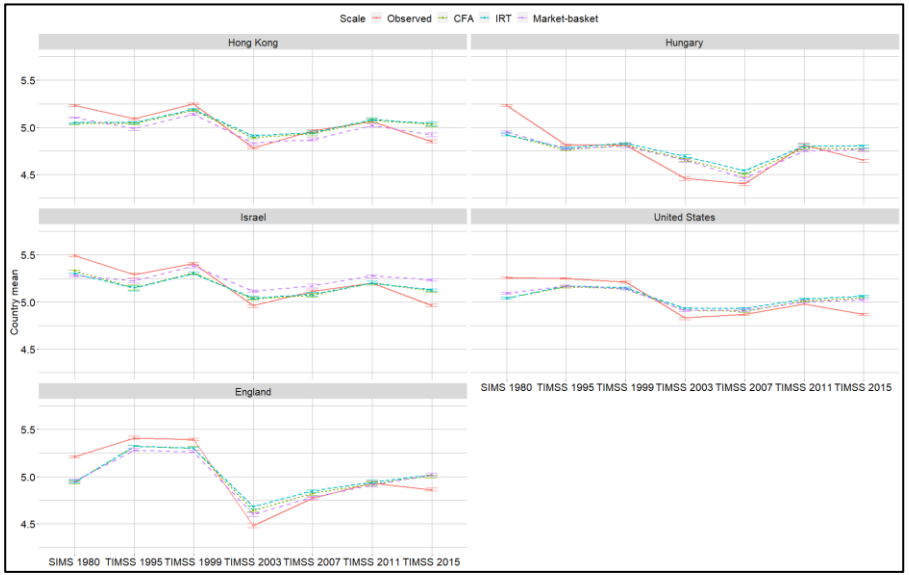
Summary of the results cont'd

Trend descriptions by linking approach, motivation for learning mathematics

- The observed scales were constructed by computing the sum of the scores per person at each time point divided by the number of answered items; then the standardized scores considering a mean of five and a standard deviation of one were calculated
- The country-level trends show similar patterns across linking approaches

Extrinsic motivation

Intrinsic motivation



Reporting the scales

- The first-second-time scales for the first-phase studies are publicly available at the Center for Comparative Analysis of Educational Achievement (COMPEAT) repository along with the documentation of the scale linking
- The sampling differences need to be considered when using the scales
- For instance, Strietholt et al. (2013) developed a correction model to improve comparability across countries and IEA studies on reading in terms of age and schooling; another suggestion is to account for these differences between time and countries is to treat age and grade level as plausible explanatory variables

Concluding remarks

- The utility of linking the studies stems from the advanced econometric methods and country-level longitudinal modeling techniques that have already encouraged research involving ILSA outcomes that are on separate scales
- The main purpose was to facilitate future country-level longitudinal studies that include the first-phase IEA studies by the means of comparable measures of mathematics and science achievement in eighth grade
- Such studies might shed light on explanations for changes in the educational outcomes of participating countries
- However, drawing valid causal inferences from observational data is challenging (see e.g., Allardt, 1990; Rutkowski & Delandshere, 2016)
- Suggestions for advanced statistical methods for causal inferences based on ILSA data have been made by several researchers (see e.g., Gustafsson, 2008; Gustafsson & Nilsen, 2022; Robinson, 2013; Schlotter et al., 2014)

References

Allardt, E. (1990). Challenges for comparative social research. *Acta Sociologica*, *33*(3), 183–193. https://doi.org/10.1177/000169939003300302

Angoff, W., & Ford, S. (1973). Item-race interaction on a test of scholastic aptitude. *Journal of Educational Measurement*, *10*(2), 95–106.

Gustafsson, J.-E. (2008). Effects of international comparative studies on educational quality on the quality of educational research. *European Educational Research Journal*, *7*(1), 1–17. https://doi.org/10.2304/eerj.2008.7.1.1

Gustafsson, J.-E., & Nilsen, T. (2022). Methods of causal analysis with ILSA data. In T. Nilsen, A. Stancel-Piątak, & J.-E. Gustafsson (Eds.), *International handbook of comparative large-scale studies in education*. Springer International Publishing. https://doi.org/10.1007/978-3-030-38298-8_56-1

Kolen, M. J., & Brennan, R. L. (2014). *Test equating, scaling, and linking: Methods and practices* (3. ed.). Springer.

Magis, D., & Facon, B. (2014). DeltaPlotR: An R package for differential item functioning analysis with Angoff's delta plot. *Journal of Statistical Software*, *59*(Code Snippet 1). https://doi.org/10.18637/jss.v059.c01

Meade, A. W., & Lautenschlager, G. J. (2004). A comparison of item response theory and confirmatory factor analytic methodologies for establishing measurement equivalence/invariance. *Organizational Research Methods*, *7*(4), 361–388. https://doi.org/10.1177/1094428104268027

Robinson, J. P. (2013). Causal inference and comparative analysis with large-scale assessment data. In L. Rutkowski, M. von Davier, & D. Rutkowski (Eds.), *Handbook of international large-scale assessment* (pp. 521–545). CRC Press.

Rutkowski, D., & Delandshere, G. (2016). Causal inferences with large scale assessment data: Using a validity framework. *Large-Scale Assessments in Education*, *4*(1). https://doi.org/10.1186/s40536-016-0019-1

Schlotter, M., Schwerdt, G., & Woessmann, L. (2014). Econometric methods for causal evaluation of educational policies and practices: A non-technical guide. In R. Strietholt, W. Bos, J.-E. Gustafsson, & M. Rosén (Eds.), *Educational policy evaluation through international comparative assessments* (pp. 95–126). Waxmann.

Strietholt, R., Rosén, M., & Bos, W. (2013). A correction model for differences in the sample compositions: The degree of comparability as a function of age and schooling. *Large-Scale Assessments in Education*, *1*(1), 1. https://doi.org/10.1186/2196-0739-1-1

Svetina, D., Rutkowski, L., & Rutkowski, D. (2020). Multiple-group invariance with categorical outcomes using updated guidelines: An illustration using Mplus and the lavaan/semTools packages. *Structural Equation Modeling: A Multidisciplinary Journal*, *27*(1), 111–130. https://doi.org/10.1080/10705511.2019.1602776

Wu, H., & Estabrook, R. (2016). Identification of confirmatory factor analysis models of different levels of invariance for ordered categorical outcomes. *Psychometrika*, *81*(4), 1014–1045. https://doi.org/10.1007/s11336-016-9506-0

Zwitser, R. J., Glaser, S. S. F., & Maris, G. (2017). Monitoring countries in a changing world: A new look at DIF in international surveys. *Psychometrika*, *82*(1), 210–232. https://doi.org/10.1007/s11336-016-9543-8

Thank you for your attention!

https://www.gu.se/en/about/find-staff/erikamajoros
erika.majoros@gu.se