Thesis title: Linking recent and older IEA studies on mathematics and science

University of completion: University of Gothenburg, Sweden

Author: Erika Majoros

Email address: erika.majoros@gu.se

Postal address: Department of Education and Special Education, Box 300, 405 30 Gothenburg, Sweden

Summary of the thesis *Linking recent and older IEA studies on mathematics and science*

Introduction

In this thesis, recent and older international assessments on mathematics and science are explored with the aim to link these surveys and scale the system-level educational outcomes onto a common metric. The main reason for linking the assessments is to provide researchers with comparable data of grade-eight mathematics and science achievement and motivation scales over a long time period.

The scales achieved in this thesis combined with powerful analytical approaches such as country-level longitudinal modeling techniques and advanced econometric methods allow for investigating changes in educational systems. For instance, educational reforms that take effect in the long term can be evaluated on the national level. In the comparative context, longitudinal studies are useful to explore global phenomena, such as trends toward a "global curriculum" (Johansson & Strietholt, 2019; Rutkowski & Rutkowski, 2009) or changes in the "socioeconomic achievement gap" (Broer et al., 2019; Chmielewski, 2019).

Two types of educational outcomes are the focus of this thesis. Firstly, cognitive outcomes, i.e., student achievement in mathematics and science in grade eight. Secondly, affective outcomes, i.e., how motivated students are for learning mathematics. These outcomes have been measured through international large-scale assessments (ILSAs) administered by the International Association for the Evaluation of Education Achievement (IEA).

The IEA has been maintaining trend scales of mathematics and science achievement since 1995. Before that, the IEA conducted four ILSAs in these subjects, but these early assessments have not been officially linked to the Trends in International Mathematics and Science Study (TIMSS) scales. In this thesis, the ILSAs administered before 1995 are referred to as the first-phase studies, while those after as the second-phase assessments (Gustafsson, 2008).

The decision not to link the studies from the two phases was motivated by the changes that have been made to the instruments, populations, and procedures between the early survey administrations (Martin & Kelly, 1996). Technological and methodological challenges at that time might also have constrained the feasibility of linking. During the decades since the first assessment, technical decisions have been made concerning e.g., sampling of items and test-takers or item wording. These decisions pose challenges to comparability and consequently to linking the assessments. However, recent technical and methodological advancements allow for tackling such challenges.

Previous research has shown that it is possible to link cognitive outcomes of the early IEA ILSAs to the recent assessments, with various linking approaches. One approach has been to link scales from surveys that include common items taking advantage of item response theory (IRT) modeling. Afrassa (2005) and Strietholt and Rosén (2016) linked cognitive outcomes of mathematics and reading achievement with this approach. However, the linking study on mathematics (Afrassa, 2005) remained limited in terms of evaluating the comparability with the TIMSS reporting scale and the scope of educational systems included in the linking.

Another linking approach has been applied to scores from different regional, national, or international assessments over a long period, which relies on classical test theory (CTT). In this approach, not all surveys have overlapping items, therefore, the linking is performed under stronger assumptions related to ability distributions (see e.g., Chmielewski, 2019; Hanushek & Woessmann, 2012).

The trend measurement of affective outcomes began with the 2011 administration of TIMSS. Certain context questionnaire scales that included common items across TIMSS 2011, TIMSS 2015, and TIMSS 2019 were linked to common metrics (Martin et al., 2012, 2016; Yin & Fishbein, 2020). To the best of my knowledge, there is no previous research on extending these longitudinal affective scales.

It can be concluded that with the recent methodological advancements and the increasing role that ILSAs play in educational systems, it is worth exploring the possibilities that lay in legacy data. The contribution of this thesis is twofold. First, the linking techniques may be applied to other large-scale assessments, in which changes have occurred between administrations. Second, the achieved scales are of potential use for future longitudinal studies.

This thesis is 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?

Method

The empirical work in this thesis is based on data from first-phase studies in mathematics and science that are listed in Table 1, from the populations representing 13-year-olds (FIMS and SIMS) and 14-year-olds (FISS and SISS). The data of the first-phase studies were processed differently compared to data from studies conducted later. This thesis took advantage of the work that has been done in the project titled Center for Comparative Analysis of Educational Achievement (COMPEAT). This project improved the conditions for secondary analysis by making the data and documentation from the early studies available online in updated formats[1]. Data and documentation for the TIMSS administrations were downloaded from the IEA Data Repository[2]. The data of eighth-grade students were used from all TIMSS cycles between 1995 and 2015.

Table 1
IEA ILSAs on mathematics and science administered in the first phase

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

In the IEA ILSAs, the achievement tests have maintained a set of common items between consecutive administrations. The sets of common items serve as anchor tests between assessments. These items are referred to as *bridge items* in this thesis. Studies I (Majoros et al., 2021) and II (Majoros, n.d.) were concerned with linking mathematics and science cognitive test items of the assessments. Study III (Majoros et al., 2022) explored linking affective, i.e., intrinsic- and extrinsic motivational items.

[1] https://www.gu.se/en/center-for-comparative-analysis-of-educational-achievement-compeat

[2] https://www.iea.nl/data-tools/repository/timss

Evaluating the comparability of the outcomes

This section is a brief overview of the methods addressing longitudinal and cross-sectional comparability that were applied in Studies I-III. The degree of similarity across assessments to be linked determines the "utility and reasonableness" (Kolen & Brennan, 2014, p. 498) of linking. Kolen and Brennan (2014) proposed four criteria for evaluating similarity: inferences, populations, constructs, and measurement characteristics. Thus first, the goals need to be evaluated concerning the types of inferences drawn from the tests to be linked. Second, the alignment between the target populations of the assessments to be linked needs to be scrutinized. Third, the similarity of the measured constructs is to be evaluated. Finally, the measurement conditions, such as test length, test format, and administration need to be scrutinized.

After the evaluation of the similarity of the surveys, the comparability of bridge items over time was investigated. More specifically, bridge item behavior was evaluated across assessments with the delta plot method (Angoff & Ford, 1973). The delta plot is a method to identify differential item functioning (DIF) among dichotomously scored items (Magis & Facon, 2014). The delta plot method for detecting DIF works under the CTT framework. This choice of method in this thesis has been made for several reasons. Firstly, the delta plot is a not computationally intensive method. Secondly, this is a relative DIF method, i.e., bridge items were evaluated concerning all items comprising the bridge. Finally, some issues with the traditional DIF analysis methods persist, which have been discussed extensively in the literature (see e.g., Bechger & Maris, 2015; Cuellar, 2022; Cuellar et al., 2021; Doebler, 2019; Yuan et al., 2021).

Finally, the cross-cultural measurement invariance of the constructs was considered. The cognitive constructs were assumed to be invariant based on the numerous quality assurance processes applied in the assessments. The cross-cultural comparability of affective constructs was evaluated by applying multiple-group confirmatory factor analysis (MGCFA) for each time point. The confirmatory factor analysis (CFA) 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; Wu & Estabrook, 2016).

Model fit was evaluated by absolute and relative fit indices. $\chi 2$, the root mean square error of approximation (RMSEA), and the standardized root mean square residual (SRMR) values served as absolute model fit indices. The relative fit was indicated by the comparative fit index (CFI) and the Tucker-Lewis index (TLI). When evaluating these results, some caution needed to be taken. First, the $\chi 2$ values are sample size sensitive (Brown, 2015) Second, the presence of negatively worded items potentially causes one-dimensional CFA models to show a poor fit (see e.g., Marsh, 1996; Steinmann et al., 2021; Woods, 2006; Zhang et al., 2016). Finally, model fit values are influenced by many factors, such as estimation method or categorical/continuous specification, and Shi and Maydeu-Olivares (2020) suggested using only the SRMR because it is more consistent across these factors.

Linking approaches

Several approaches for linking were explored in the constituent empirical studies. Study I may be viewed as a feasibility study for Study II. Study II placed the mathematics and science

assessments on the TIMSS reporting scale. Study III is an exploratory investigation of linking affective scales, therefore, several different approaches were compared.

In 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. First, different IRT models were. The best fitting model combined 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.

Second, the item parameters were estimated via concurrent calibration. Thus, the item parameters were on the same IRT scale using data from all studies and four countries. Senate weights that sum to 500 for each country's student data were applied (stratum weights in SIMS were rescaled to sum to 500), thus, each country contributed equally to the item calibration. There were no weight variables in the FIMS 1964 datasets; therefore, individuals within a country were weighted equally, to sum up to 500.

The third step was to use the item parameter estimates for person scoring. Five plausible values (PVs) were drawn per student using the expected a-posteriori method. The estimated abilities were converted to scale scores; thus, each PV was transformed to a metric with a mean of 500 and a standard deviation of 100 points across time. The transformed scores were used to compute the mean mathematics achievement for the respective country per study following Rubin's (1987) rule of pooling.

Study II compared two linking approaches for the mathematics scale. Firstly, *the four-country-all-time* (points) approach took advantage of the item parameters estimated by the method in Study I. The procedure started with the separate ability estimation for FIMS and SIMS, fixing the item parameters to the estimated values via the method in Study I. Then the distribution of the five PVs estimated for FIMS and SIMS was matched with the distribution of the reported TIMSS 1995 PVs.

Secondly, the *first-second-time* approach involved the concurrent calibration of item parameters using the first and second ILSAs on mathematics with the bridge items' parameters fixed to the values reported for TIMSS 1995. These item parameters were reported after a rescaling procedure in the 1999 assessment cycle (Martin et al., 2000). Then the student abilities were estimated separately for FIMS and SIMS, drawing five PVs per test-taker. To locate the student ability estimates on the TIMSS reporting scale, the original transformation constants used for the reported TIMSS 1995 scaling needed to be applied. These constants were acquired through Gonzalez (2022).

For constructing the science achievement scale, the first-second-time approach was chosen for several reasons. First, the IRT models are the same as those used in the TIMSS procedures, i.e., the 2PL, three-parameter logistic model (3PL), and GPCM. Second, this procedure is more economic although we use data from more countries (but fewer time points). If we compare the amount of information, i.e., the number of item responses used for item calibration in the two approaches, we can note on the one hand that the four-country-all-time concurrent calibration involves 893 items (1964-2015), while the first-second-time approach uses the items between 1964 and 1995, i.e., 373 items. On the other hand, the weighted number of item responses used for the link between SIMS and TIMSS 1995 is close to threefold in the first-second-time approach than those in the four-country-all-time method due to the larger number of countries (42) involved in the procedure.

Three methods were explored for constructing longitudinal affective scales in Study III: an IRT, a CFA, and a market-basket approach. Firstly, in the IRT approach, the GPCM model was found

to best fit the data. Then the item parameter estimation was conducted by concurrent calibration of all items in all studies, thus the parameters for all tests were automatically put onto the same scale. The parameters of the anchor items were assumed identical in each sample. Third, person scores were estimated and transformed onto a scale with a mean of five and a standard deviation of one.

Secondly, 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 by 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.

Finally, a market-basket approach was applied. The market-basket approach assumes that the items included in the assessment or survey define the construct. In this case, the assumption is that all the items across the time points, related to intrinsic and extrinsic motivation for learning mathematics, define each construct and can be considered as a market basket of representative items. The missing responses occur as a consequence of changes in the questionnaires across cycles. A measurement model was employed per country to generate plausible responses that fill the missing responses following the procedure suggested by (Zwitser et al., 2017). The measurement model was fit for each country separately to account for potential differences among countries. 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.

Summary of the results and discussion

Comparability of the outcomes

This section addresses the first overarching research question of this thesis: To what extent are the student outcomes comparable across the first- and second-phase IEA assessments on mathematics and science? The evaluation of the comparability across administrations shows that there has been a high level of stability concerning the inferences and measured constructs among the assessments. 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, to assess the assumption about the behavior of common items, the delta plot method was applied to all seven bridges between adjacent time points. A total number of 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. Furthermore, twelve non-anchor items were excluded due to missing answers in all countries. Overall, 893 items were included in the concurrent calibration.

In Study II, the delta plot method was applied to the six sets of bridges. These bridges consist of common items in the mathematics assessments between 1964-1980 (bridge 1), 1964-1995 (bridge 2), and 1980-1995 (bridge 3), and among the science surveys between 1970-1984 (bridge 4), 1970-1995 (bridge 5), and 1984-1995 (bridge 6). Two items in bridge 1, one item in bridge 3, and two items in bridge 4 were flagged for DIF. In the final, first-second-time linking procedure, these items were treated as unique items. To assess the assumption about the performance on the

anchor test and the whole test, Pearson's correlations were calculated. These correlations were moderate or high.

In Study III, the delta plot method was applied for each bridge between consecutive time points. The tests were conducted for each country separately as well as the pooled data and all these tests yielded no items flagged for DIF.

Cross-cultural invariance

In Study III, the measurement invariance was tested across countries at each time point. 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.

Trend descriptions by linking approach

This section presents results from the empirical studies addressing the second overarching research question of the thesis: How do different linking approaches influence the descriptions of the system-level trends? The results are discussed in terms of the measured constructs.

Mathematics achievement

In Study II, the first-second-time approach yielded consistently higher country means except for the low-performing countries in SIMS. There are three main differences in the linking approaches. First, 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. Second, 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. Finally, 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 between the two approaches.

Motivation for learning mathematics

The five countries included in Study III were treated as a single group both cross-sectionally and longitudinally in the CFA and IRT procedures. They were treated separately in the market-basket approach, but data were pooled into a single group model over time. 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 three methods yielded similar results at the individual- as well as country levels. The correlations between individual scores were high across methods for both motivation constructs, ranging between 0.96 and 1. The country-level trends show similar patterns across linking approaches.

Limitations

The final scales for the first-phase studies are publicly available on the COMPEAT website[3] along with the documentation of the scale linking. I would like to emphasize that 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 approach to account for these differences between time and countries is to treat age and grade level as plausible explanatory variables.

Many factors influence the quality of linking. These factors include the degree of similarity across assessments, the stability of the constructs in terms of content, meaning, and context, and the number and behavior of bridge items. This thesis addressed these influences or potential sources of bias from a substantive as well as a measurement point of view. However, there may be a degree of uncertainty in terms of evaluating these aspects because the linking involves legacy data.

A limitation of the longitudinal achievement scales lies in the within-country comparability because of the target populations of the assessments. The shift in the IEA sampling strategy was tackled with as good approximations of homogenous samples as possible. In further analyses using the new scale scores, age and grade level can be treated as control variables.

The relatively few common items in the bridges connecting to TIMSS 1995 are another concern. The ratio of bridge items in the affective scales is much less concerning than in the case of the cognitive scales. However, the concurrent calibration method provides the best approach to having only a few bridge items, as pointed out by Wingersky and Lord (1984).

Another limitation concerns the coding and treatment of different types of missing data in the achievement tests. In the first-phase studies, the not-reached type of missing responses was not distinguished. Therefore, those missing responses were treated as missing, unlike in the TIMSS scaling procedure. It would be possible to make this distinction and explore the influence on the results.

The affective scales analyzed in Study III until recently had not been designed for trend measurement. Modifications have occurred over time, e.g., the number of response options was changed in 1995. The way of handling the middle option in SIMS posed a limitation to the study. In addition, the linking methods in this study did not account for the differences in the motivation distributions over time since a single-group approach was applied.

Another limitation of the affective trend scales is that the standard errors of the means are underestimated. The reason for this is that because of the stratified multistage sampling design used in TIMSS, the simple random sampling assumed in the procedure for calculating standard errors of estimates did not apply (Rutkowski et al., 2010).

Finally, one of the most challenging remaining questions is whether changes in wording affect the internal relationships among motivational items (e.g., factor structure). Since non-identical sets of items were explored over time, the number of items varies at almost each time point, which makes the investigation of the effects of changes in item wording challenging.

Conclusions

With the newly established scales in this thesis, it is possible to examine long-term changes comparatively or within countries. System-level changes take time, therefore, evaluating school

[3]https://www.gu.se/en/center-for-comparative-analysis-of-educational-achievement-compeat/linking-projects/mathematics-and-science

reforms requires long-term data. Furthermore, as mentioned earlier, powerful statistical approaches to address causal research questions may be applied to system-level longitudinal data.

For instance, Strietholt et al. (2019) recently reviewed the international comparative literature on the impact of education policies on the socioeconomic achievement gap. The authors found that most of the existing research was descriptive, estimating simple correlations based on cross-sectional data. Further research into mapping indicators of socioeconomic background in the first-phase IEA surveys combined with the achievement scales could potentially contribute to this line of inquiry.

Another potential area to take advantage of the long-term scales lies in issues related to the global educational reform movement (Fuller & Stevenson, 2019; Sahlberg, 2016). Such related phenomena involve privatization, free school choice, school competition, or teacher education.

An avenue for future research is to continue the exploration of long-term trends in the attitudes toward learning mathematics and science. There are considerably larger challenges with these outcomes than the achievement scales due to item-level changes. The market-basket approach employed in this thesis offers possibilities for linking with fewer assumptions of comparability than item-level linking.

Another interesting area is to explore the possibilities of linking the first- and second-phase IEA studies on science for the grade four population. Data are publicly available for the old studies. Furthermore, in the 1995 administration of TIMSS, the achievement tests of the younger and older populations were linked through anchor items (Martin & Kelly, 1996). It could be interesting to revisit this linkage for investigating developmental changes.

Furthermore, it may be useful to map other contextual indicators from the first-phase studies, which could facilitate more complex investigations in, for instance, gender differences. The relative proportion of females choosing a mathematical track in upper secondary and higher education or STEM-related professions is still unreasonably low and unrelated to mathematics achievement in many countries. However, current research on gender equality paradoxes based on ILSA data typically does not account for more complex group differences, e.g., based on socioeconomic background.

References

Afrassa, T. M. (2005). Monitoring mathematics achievement over time: A secondary analysis of FIMS, SIMS and TIMS: a Rasch analysis. In Alagumalai, Curtis, David D. & N. Hungi (Eds.), *Applied Rasch measurement: A book of exemplars* (pp. 61–77). Springer.

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

Bechger, T. M., & Maris, G. (2015). A statistical test for differential item pair functioning. *Psychometrika*, *80*(2), 317–340. https://doi.org/10.1007/s11336-014-9408-y

Broer, M., Bai, Y., & Fonseca, F. (2019). *Socioeconomic inequality and educational outcomes: Evidence from twenty years of TIMSS* (Vol. 5). Springer. https://doi.org/10.1007/978-3-030-11991-1

Brown, T. A. (2015). *Confirmatory factor analysis for applied research* (Second edition). The Guilford Press.

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

Cuellar, E. (2022). *Making sense of DIF in international large-scale assessments in education* [Doctoral dissertation]. University of Amsterdam.

Cuellar, E., Partchev, I., Zwitser, R., & Bechger, T. (2021). Making sense out of measurement non-invariance: How to explore differences among educational systems in international large-scale

assessments. *Educational Assessment, Evaluation and Accountability*, *33*(1), 9–25. https://doi.org/10.1007/s11092-021-09355-x

Doebler, A. (2019). Looking at DIF from a new perspective: A structure-based approach acknowledging inherent indefinability. *Applied Psychological Measurement*, *43*(4), 303–321. https://doi.org/10.1177/0146621618795727

Fuller, K., & Stevenson, H. (2019). Global education reform: Understanding the movement. *Educational Review*, *71*(1), 1–4. https://doi.org/10.1080/00131911.2019.1532718

Gonzalez, E. J. (2022, September 16). *Personal communication*.

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

Hanushek, E. A., & Woessmann, L. (2012). Do better schools lead to more growth? Cognitive skills, economic outcomes, and causation. *Journal of Economic Growth*, *17*(4), 267–321. https://doi.org/10.1007/s10887-012-9081-x

Johansson, S., & Strietholt, R. (2019). Globalised student achievement? A longitudinal and cross-country analysis of convergence in mathematics performance. *Comparative Education*, *55*(4), 536–556. https://doi.org/10.1080/03050068.2019.1657711

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

Majoros, E. (n.d.). *Linking the first- and second-phase IEA studies on mathematics and science* [Manuscript submitted for publication].

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

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

Marsh, H. W. (1996). Positive and negative global self-esteem: A substantively meaningful distinction or artifactors? *Journal of Personality and Social Psychology*, *70*(4), 810–819. https://doi.org/10.1037/0022-3514.70.4.810

Martin, M. O., Gregory, K. D., & Stemler, S. E. (Eds.). (2000). *TIMSS 1999 technical report*. TIMSS & PIRLS International Study Center, Boston College.

Martin, M. O., & Kelly, D. L. (Eds.). (1996). *Third international mathematics and science study technical report: Design and development* (Vol. 1). TIMSS & PIRLS International Study Center, Boston College.

Martin, M. O., Mullis, I. V. S., Foy, P., Arora, & Alka. (2012). Creating and interpreting the TIMSS and PIRLS 2011 context questionnaire scales. In M. O. Martin & I. V. S. Mullis (Eds.), *Methods and procedures in TIMSS and PIRLS 2011*. TIMSS & PIRLS International Study Center, Boston College. https://timssandpirls.bc.edu/methods/t-context-q-scales.html

Martin, M. O., Mullis, I. V. S., Hooper, M., Yin, L., Foy, P., & Palazzo, L. (2016). Creating and interpreting the TIMSS 2015 context questionnaire scales. In M. O. Martin, I. V. S. Mullis, & M. Hooper (Eds.), *Methods and procedures in TIMSS 2015* (p. 15.1-15.312). TIMSS & PIRLS International Study Center, Boston College.

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

Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys*. John Wiley & Sons, Inc. https://doi.org/10.1002/9780470316696

Rutkowski, L., Gonzalez, E., Joncas, M., & von Davier, M. (2010). International large-scale assessment data: Issues in secondary analysis and reporting. *Educational Researcher*, *39*(2), 142–151.

Rutkowski, L., & Rutkowski, D. (2009). Trends in TIMSS responses over time: Evidence of global forces in education? *Educational Research and Evaluation*, *15*(2), 137–152. https://doi.org/10.1080/13803610902784352

Sahlberg, P. (2016). The global educational reform movement and its impact on schooling. In K. E. Mundy, A. Green, B. Lingard, & A. Verger (Eds.), *The handbook of global education policy* (pp. 128–144). John Wiley & Sons, Ltd.

Shi, D., & Maydeu-Olivares, A. (2020). The effect of estimation methods on SEM fit indices. *Educational and Psychological Measurement*, *80*(3), 421–445. https://doi.org/10.1177/0013164419885164

Steinmann, I., Sánchez, D., van Laar, S., & Braeken, J. (2021). The impact of inconsistent responders to mixed-worded scales on inferences in international large-scale assessments. *Assessment in Education: Principles, Policy & Practice*, 1–22. https://doi.org/10.1080/0969594X.2021.2005302

Strietholt, R., Gustafsson, J.-E., Hogrebe, N., Rolfe, V., Rosén, M., Steinmann, I., & Hansen, K. Y. (2019). The impact of education policies on socioeconomic inequality in student achievement: A review of comparative studies. In Volante & Melchior (Eds.), *Socioeconomic inequality and student outcomes. Education Policy & Social Inequality* (Vol. 4, pp. 17–38). Springer Singapore. https://doi.org/10.1007/978-981-13-9863-6_ 2

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

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

Wingersky, M. S., & Lord, F. M. (1984). An investigation of methods for reducing sampling error in certain IRT procedures. *Applied Psychological Measurement*, *8*(3), 347–364. https://doi.org/10.1177/014662168400800312

Woods, C. M. (2006). Careless responding to reverse-worded items: Implications for confirmatory factor analysis. *Journal of Psychopathology and Behavioral Assessment*, *28*(3), 186–191. https://doi.org/10.1007/s10862-005-9004-7

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

Yin, L., & Fishbein, B. (2020). Creating and interpreting the TIMSS 2019 context questionnaire scales. In M. O. Martin, M. von Davier, & I. V. S. Mullis (Eds.), *Methods and procedures* (p. 16.1-16.331). TIMSS & PIRLS International Study Center, Boston College.

Yuan, K.-H., Liu, H., & Han, Y. (2021). Differential item functioning analysis without a priori information on anchor items: QQ plots and graphical test. *Psychometrika*, *86*(2), 345–377. https://doi.org/10.1007/s11336-021-09746-5

Zhang, X., Noor, R., & Savalei, V. (2016). Examining the effect of reverse worded items on the factor structure of the need for cognition scale. *PloS One*, *11*(6), e0157795. https://doi.org/10.1371/journal.pone.0157795

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