Studies in Educational Evaluation 74 (2022) 101174

Contents lists available at ScienceDirect

Studies in Educational Evaluation

journal homepage: www.elsevier.com/locate/stueduc

Motivation towards mathematics from 1980 to 2015: Exploring the feasibility of trend scaling

Erika Majoros [a,*,1], Andrés Christiansen [b], Edwin Cuellar [c]

[a] Department of Education and Special Education, University of Gothenburg, Sweden
[b] KU Leuven, Belgium
[c] Cito, The Netherlands

ARTICLE INFO

Keywords:
International large-scale assessments
TIMSS
SIMS
Measurement invariance
Scaling and linking methods

ABSTRACT

The Trends in International Mathematics and Science Study (TIMSS) has been assessing students' attitudes every fourth year since 1995. The trend scaling of these constructs started in 2011, fueling interest in exploring how different education systems perform regarding affective outcomes of education. This study explored the feasibility of establishing long-term motivational scales extended with the Second International Mathematics Study administered between 1976 and 1982. We investigated whether cross-cultural comparability holds and how different methodological approaches influence the long-term scaling of motivation towards mathematics. We used grade eight data from five educational systems that have participated in every time point up to 2015. We followed three alternatives: an item response theory-, a confirmatory factor analysis-, and a market-basket approach. Our results show that the three methods provide similar trends at the country level and high correlations at the student level. We discuss methodological implications in the context of international large-scale assessments.

The Trends in International Mathematics and Science Study (TIMSS) includes affective constructs, such as students' attitudes towards mathematics employing student background questionnaires. Including trend scaling for affective constructs has only been started recently in TIMSS 2011 (Martin et al., 2016). It is important to explore the possibilities of extending these trend scales because country-level longitudinal data facilitates powerful analytical approaches to address causal research questions.

In the present study, we focused on the feasibility of extending the TIMSS trend scales of students' motivation towards mathematics. Following the model proposed by Eccles and Wigfield (2002), we distinguished motivation by its source. When individuals engage in an activity for instrumental reasons, i.e., receiving a reward, they are extrinsically motivated. Nevertheless, when individuals engage because they enjoy the activity itself, they are intrinsically motivated. We investigated the trend component of these two scales (i.e., intrinsic- and extrinsic motivation) via confirmatory factor analysis (CFA) and item response theory (IRT) scaling methods, as well as applied a market-basket approach scaling, while we studied relevant characteristics related to measurement bias and longitudinal linking.

1. Measurement bias and equivalence

The research of student outcomes across countries needs to consider cultural differences and the possibility of measurement bias. This seems obvious regarding the cross-cultural measurement of affective constructs; nevertheless, researchers face numerous statistical challenges. We employed a methodological framework proposed by van de Vijver (2018) to describe the types of bias and equivalence in the context of cross-cultural assessments measuring affective constructs. He identified three types of bias based on their sources: construct-, method-, and item bias. The presence of construct bias indicates that the construct measured is not identical across cultures. Method bias refers to confounding factors that originate in the sampling, structural characteristics of the instrument, or administration. Finally, an item is biased when it has a different psychological meaning across cultures.

In this framework, van de Vijver (2018) defined measurement equivalence of scales by the level of comparability and three types can be distinguished: construct-, measurement unit-, and full score equivalence (van de Vijver & Leung, 1997; van de Vijver, 2015, 2018). Construct equivalence is fulfilled when the same theoretical construct is

* Correspondence to: Department of Education and Special Education, Box 300, 405 30 Gothenburg, Sweden.
E-mail address: erika.majoros@gu.se (E. Majoros).
[1] https://orcid.org/0000–0003-4417–5016

https://doi.org/10.1016/j.stueduc.2022.101174
Received 25 September 2020; Received in revised form 1 December 2021; Accepted 11 May 2022
Available online 21 May 2022

measured in each group, i.e., configural invariance holds. Measurement unit equivalence corresponds to metric invariance, i.e., the scales have the same measurement unit but different scale origins. Finally, full score equivalence means the same as scalar invariance, i.e., the scales have the same measurement unit and origin.

Measurement equivalence or invariance can be tested in the structural equation modeling (SEM) framework with a measurement model applying a CFA approach as the psychometric equivalence of a construct across groups or in the IRT framework as the lack of differential item functioning (DIF). Putnick and Bornstein (2016), in their extensive review highlighted that the SEM framework using CFA is more commonly used than IRT. Numerous researchers (e.g., D'Urso et al., 2020; Kim & Yoon, 2011; Meade & Lautenschlager, 2004) have compared the two approaches and provided recommendations for measurement invariance testing in different assessment contexts. However, most of these studies were based on simulated data.

A recent report of an OECD conference on the cross-cultural comparability of questionnaire measures in large-scale assessments (van de Vijver et al., 2018) provided a broad and up-to-date discussion regarding techniques to investigate measurement invariance. Concluding the conference, Avvisati et al. (2018) pointed out that several participants observed how the distinction between the CFA and the IRT worlds is largely artificial, and, despite the most rigorous application of preventive measures, the assumption of full comparability of measurement instruments in ILSAs cannot be upheld (see also Davidov et al., 2014). The report indicated a consensus among participants that any procedure to address the possible violation of full measurement invariance needs to consider the non-comparability of scales as a possibility. This possibility imposes great challenges and potential limitations on longitudinal linking.

1.1. Scaling affective items in international large-scale assessments

The TIMSS context questionnaire scales for trend measurement were constructed with IRT scaling using the Rasch partial credit model (PCM; Martin et al., 2016; Masters, 1982; Yin & Fishbein, 2020). To evaluate the context questionnaire scales, the Cronbach's alpha coefficient measuring internal consistency was computed for each scale for every educational system, and a principal component analysis of the scale items was conducted. Measurement invariance across countries was not evaluated, however, the scaling was done with a single-group design.

Questionnaire data surveying latent constructs may also be scaled within the SEM framework, with a CFA measurement model. An example of an ILSA employing CFA for scaling affective items is the Teaching and Learning International Survey (TALIS) administered by the Organisation for Economic Co-operation and Development (OECD) since 2008. The comparability across participating educational systems in all three cycles of TALIS was evaluated by measurement invariance testing with the multiple-group confirmatory factor analysis (MGCFA; Organisation for Economic Co-operation & Development, 2019).

1.2. Longitudinal linking

The process of adjusting, via statistical methods, two tests with differences in content or difficulty is known as linking. There is extensive research on linking cognitive outcomes in ILSAs over time, with various linking approaches. Linking can be achieved using IRT linking methods (see e.g., Afrassa, 2005; Johansson & Strietholt, 2019; Majoros et al., 2021; Strietholt & Rosén, 2016), which require a set of common items across tests among other preconditions. There have been also several attempts to link test scores from different regional, national, or international assessments assuming similar target populations and representative samples over a long period. These linking studies rely on IRT within the assessments and classical test theory across them because of the limited amount of overlapping items (see e.g., Altinok et al., 2018; Chmielewski, 2019; Hanushek & Wößmann, 2012).

1.3. The current linking practice in TIMSS for affective scales

Certain context questionnaire scales – constructed with IRT scaling – that maintained many of the same items across TIMSS 2011, TIMSS 2015, and TIMSS 2019 (see Martin et al., 2016; Yin & Fishbein, 2020), were linked through a two-step transformation process by applying the mean/sigma method. The first transformation placed the TIMSS 2019 logit scale scores on the TIMSS 2015 logit metric by applying the procedure described by Marco (1977) and referred by Kolen and Brennan (2014) as the mean/sigma method to the two sets of common item parameters. These sets were estimated by the separate calibration of TIMSS 2019 data and the TIMSS 2015 data. The mean and standard deviation of the estimates of the threshold parameters (Masters, 1982), i.e., the difference between item location and item step parameters, were used for all common items and all categories for each calibration. The second step was to transform the TIMSS 2015 Rasch logit scores on the TIMSS scale reporting metric (mean:10, standard deviation: 2). To assess the accuracy of the linking, item parameter estimates for the common items were compared across the two cycles by examining the differences between the TIMSS 2019 item parameter estimates after being transformed to the TIMSS 2015 logit metric and the TIMSS 2015 item parameter estimates on the 2015 logit scale. This linking procedure assumed full measurement invariance across countries at each time point.

1.4. The present study

In terms of method bias, we built on previous research (Majoros et al., 2021; Majoros et al., 2020) evaluating the comparability across the respective assessments, i.e., SIMS and every cycle of TIMSS between 1995 and 2015. To determine their overall similarity, the inferences, populations, measurement characteristics (i.e., method bias), and constructs (i.e., construct bias) were explored based on the scheme proposed by Kolen and Brennan (2014). A sufficient degree of overall similarity was found, therefore, we assumed that method bias was not severely impacting the trend scales in the present study.

To evaluate construct bias across countries, we followed the guidelines proposed by Svetina et al. (2020) to test measurement invariance across countries at each time point. They focused on selected solutions by Wu and Estabrook (2016) in terms of model identification and invariance testing. Thus, after establishing configural invariance, threshold invariance was tested first, followed by invariance testing for factor loadings. This approach differs from current practices of conducting measurement invariance testing, where a baseline model is established first and increasing parameter restrictions are subsequently imposed.

Item bias over time was evaluated focusing on the anchor items between time points. We used Angoff's delta plot method (Angoff & Ford, 1973) for investigating item parameter drift between time points. The delta plot is a score-based method that compares the proportions of correct responses in the reference group and the focal group. Items are flagged as biased when they change relative to the set of all items in the test. Magis and Facon (2014) argued that the main benefit of using relative methods is that the identification of problematic items relies on the particular items themselves. Moreover, we have a small number of anchor items and our major interest is in their overall trend.

We then performed the linking and scaling of the data with three methods applying different sets of assumptions. Since we attempted to link non-identical sets of items measuring the same constructs over time, only subsets of items were bridging over the assessments. Therefore, we made use of latent variable modeling in the IRT and SEM frameworks. In addition, we proposed a third alternative based on the manifest probabilities and plausible scores, the market-basket approach.

First, the IRT linking was achieved by concurrent calibration (Wingersky & Lord, 1984) of all items in all studies, thus the parameters estimated for each test were automatically put on the same scale. We have chosen the concurrent procedure because this method provides

smaller standard errors and involves fewer assumptions than other IRT procedures, and good linking may be achieved with as few as five common items or less (Wingersky & Lord, 1984). Item parameters were estimated simultaneously while the parameters of the anchor items were assumed identical across all time points and educational systems. We compared the PCM model applied in TIMSS with the generalized partial credit model (GPCM; Muraki, 1992). Second, in the SEM approach, we fit a single-group CFA model for each motivation scale and the estimate factor score for scaling the data. We assumed strong invariance across countries and over time.

Third, instead of reporting estimates on latent variable scales, we used a market-basket approach proposed by Zwitser et al. (2017). The main idea was that the constructs are defined as a large set of items, data are collected with subsets of items and reported in terms of summary statistics. To deal with incomplete data, we fit a measurement model (e.g., an IRT model) to generate plausible responses. When applying this approach, we assumed that the anchor items' parameter estimates are invariant over time within countries. To account for cross-cultural DIF, we fit a separate model per country. Another assumption was that, for each time point, the market basket of items in the survey represented the construct. Finally, we reported the results based on summary statistics, in this case, the expected sum scores over the completed set of responses, i.e., plausible scores.

2. Method

2.1. Data

The present study focused on grade eight (or equivalent) student questionnaire data in seven ILSAs on mathematics administered by the International Association for the Evaluation of Educational Achievement (IEA). Hence, we pooled the data of SIMS, administered in 1980, and all six cycles of TIMSS administered in every fourth year from 1995 to 2015. The data of SIMS were gathered from the Center for Comparative Analyses of Educational Achievement website (COMPEAT[2]). Data and documentation of the TIMSS studies were downloaded from the IEA Study Data Repository.[3]

We have selected the six educational systems that have participated in all time points: England, Hong Kong, Hungary, Israel, Japan, and the United States. The sample sizes are presented in Table 1. We can observe that in 1995, two adjacent grades were sampled in each country except for Israel. The sample size differences were taken into account with the use of senate weights. More details are provided in the analytical steps section.

2.1.1. Items

The items included in the present study correspond to intrinsic- and extrinsic motivation towards mathematics included in the students' questionnaire for each assessment (Appendices A and B). The overlapping items along with their variable names are presented for each scale in Tables 2 and 3. We can observe that there are identical and *similar* items across assessments. Nevertheless, item wordings have changed over time in some cases. In a few instances, it also meant shifting from positively worded statements to negatively worded items. These changes might influence comparability (for examining the effects of item wording changes see e.g., Dedrick et al., 2007; Schuman & Presser, 1996).

The number of overlapping items is summarized in Table 4. Overall, the pooled extrinsic motivation scale consisted of 15 items, while the intrinsic motivation scale comprised of an item pool of 19 questions.

The students had four response options to choose from in the case of all items in all TIMSS cycles: *strongly agree*, *agree*, *disagree*, and *strongly disagree* (the wording refers to 1995). However, in SIMS, they had a middle option: *undecided*. The proportion of *undecided* responses in the analyzed countries are shown in Tables 5 and 6. It is interesting to observe how these proportions vary between countries. In most of the cases, the Japanese students used the middle option considerably more frequently than students in other countries. Interestingly, the only exception was the item "I would like to work at a job that lets me use mathematics". To this item, only half of the Japanese students responded *undecided* compared to the other countries or the other items within the intrinsic motivation scale. This could be a preliminary indication of measurement non-invariance among the educational systems.

Due to the considerably large portion of middle responses, we have decided not to treat these responses as missing values. We can also observe that in most of the cases, the proportion of middle responses was the highest in Japan. We recoded these responses to random answers between the options agree and disagree. There were some cases when a student selected the middle option for all items. We excluded these cases, 0.95% of the sample for the extrinsic- and 0.71% of the intrinsic motivation scale.

2.1.2. Missing data

The proportion of missing responses ranged from 0.05% to 2.11% in the extrinsic- and 0.14–1.76% in the intrinsic motivation scales. One item presented only missing values for the Japanese sample in 1995.

2.1.3. Internal consistency of the scales

The internal consistency of the motivation scales varied across educational systems and over time. Appendix C shows the Cronbach's alpha coefficients that range between 0.47 and 0.94. The Japanese data from SIMS displayed unacceptable coefficients for both scales. Apart from these values, in most instances, the reliability was acceptable (>0.70; Cortina, 1993) and in all cases above.61.

2.2. Analytical steps

2.2.1. Comparability

Cross-cultural comparability. To test measurement invariance across countries, we performed an MGCFA for each time point, using Mplus 8. Students were grouped by country and the first step was to identify the baseline model and testing for configural invariance among countries. After establishing configural invariance, threshold invariance was tested, followed by invariance testing for factor loadings. The questionnaire items were treated as categorical variables and we followed the procedure outlined by Svetina et al. (2020). The WLSMV estimator was used to estimate factor models. This method produces a weighted least square parameter estimate by using a diagonal weight matrix, robust standard errors, and a mean- and variance-adjusted χ^2 test statistic (Brown, 2015).

Longitudinal comparability. We used Angoff's delta plot method (Angoff & Ford, 1973) for the detection of the item parameter drift between time points using the deltaPlotR package (Magis & Facon, 2014) for the statistics environment R (R Core Team). Under this method, the proportion of responses indicating positive endorsement are compared between the two groups. If there is no item parameter drift, these proportions should be located on a diagonal line. Items that are separated from that diagonal are flagged as biased items. For this step, we recoded the answers *strongly agree* to *agree* (1) and *strongly disagree* to *disagree* (0). Following the suggestion of Magis and Facon (2014), the threshold was derived by using a normality assumption on the delta points. Each item j has a pair of delta scores (Δ_{j0}, Δ_{j1}), referred to as the delta point. These delta points can be displayed in a scatter plot, called the diagonal plot, with the delta scores of the reference group on the X-axis and of the focal group on the Y-axis. The plot usually takes the form of an elliptical cloud of delta points. The items that substantially depart from the main axis of this ellipsoid can be flagged as DIF.

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

[3] https://www.iea.nl/data

Table 1
Sample Sizes.

		1980	1995	1999	2003	2007	2011	2015
England	grade 7	–	1776	–	–	–	–	–
	grade 8	2583	1744	2833	2662	3938	3802	4718
Hong Kong	grade 7	–	3347	–	–	–	–	–
	grade 8	5362	3277	5144	4927	3431	3969	4111
Hungary	grade 7	–	2998	–	–	–	–	–
	grade 8	1753	2774	3168	3270	4076	5162	4869
Israel	grade 8	3343	1333	4060	4179	3133	4641	5416
Japan	grade 7	7697	5144	–	–	–	–	–
	grade 8	–	5108	4684	4831	3133	4355	4729
United States	grade 7	–	3772	–	–	–	–	–
	grade 8	6446	6944	8748	8777	7261	10,326	10,012

Table 2
Common Items of the Extrinsic Motivation Scales in SIMS and TIMSS.

SIMS 1980	TIMSS 1995	TIMSS 1999	TIMSS 2003	TIMSS 2007	TIMSS 2011	TIMSS 2015
It is important to know mathematics such as algebra or geometry in order to get a good job. (ymthjob)	I need to do well in mathematics to get desired job. (BSBMJOB)	I need to do well in mathematics to get the job I want. (bsbmjob)	I need to do well in math to get the job I want. (bsbmaget)	I need to do well in mathematics to get the job I want. (BS4MAGET)	How much do you agree that you need to do well in mathematics to get the job you want? (BSBM16M)	I need to do well in mathematics to get the job I want. (BSBM20D)
I can get along well in everyday life without using mathematics. (ygowo)	Do you think that mathematics is important to everyone's life? (BSBMLIFE)	Do you think that mathematics is important to everyone's life? (bsbmlife)	I think learning mathematics will help me in my daily life. (bsbmahdl)	I think learning mathematics will help me in my daily life. (BS4MAHDL)	How much do you agree that learning mathematics will help you in your daily life? (BSBM16J)	I think learning mathematics will help me in my daily life. (BSBM20A)
My parents want me to do very well in mathematics class. (ypwwell)	I need to do well in mathematics to please my parents. (BSBMPRNT)	I need to do well in mathematics to please my parents. (bsbmprnt)				My parents think that it is important that I do well in mathematics. (BSBM20H)
	I need to do well in mathematics to get into the school I prefer. (BSBMSCHL)	I need to do well in mathematics to get into the school I prefer. (bsbmschl)	I need to do well in math to get into the university of my choice. (bsbmauni)	I need to do well in mathematics to get into the university of my choice. (BS4MAUNI)	How much do you agree that you need to do well in mathematics to get into the university of your choice? (BSBM16L)	I need to do well in mathematics to get into the university of my choice. (BSBM20C)
	I think it is important to do well in mathematics at school. (BSBMSIP2)	I think it is important to do well in mathematics at school. (bsbmsip2)			How much do you agree that it is important to do well in mathematics? (BSBM14F)	It is important to do well in mathematics. (BSBM20I)
	My friends think it is important for me to do well in mathematics at school. BSBMFIP2)	My friends think it is important for me to do well in mathematics at school. (bsbmfip2)				
	My mother thinks it is important for me to do well in mathematics at school. (BSBMMIP2)	My mother thinks it is important for me to do well in mathematics at school. (bsbmmip2)				
			I need mathematics to learn other school subjects. (bsbmaoss)	I need mathematics to learn other school subjects. (BS4MAOSS)	How much do you agree that you need mathematics to learn other school subjects? (BSBM16K)	I need mathematics to learn other school subjects. (BSBM20B)

The major axis is computed with the following equation:

$$\Delta_{j1} = a + b\Delta_{j0}2, \quad (1)$$

in which

- *a is the intercept and*
- *b is the slope*

with

$$b = \frac{s_1^2 - s_0^2 + \sqrt{(s_1^2 - s_0^2)^2 + 4 s_{01}^2}}{2s_{01}} \quad and \quad a = \overline{x_1} - b\overline{x_0}, \quad (2)$$

in which

- $\overline{x_0}$ and $\overline{x_1}$ are the sample means of the delta scores,
- s_0^2 and s_1^2 are the sample variances, and
- s_{01} is the sample covariance of the delta scores.

The perpendicular distance D_j between the major axis given in equation (1) and the delta point (Δ_{j0}, Δ_{j1}), is computed as follows:

$$D_j = \frac{b\Delta_{j0} + a - \Delta_{j1}}{\sqrt{b^2 + 1}} \quad (3)$$

2.2.2. Scaling

CFA scaling. We fit a CFA model for each motivation scale on a

Table 3
Common Items of the Intrinsic Motivation Scales in SIMS and TIMSS.

SIMS 1980	TIMSS 1995	TIMSS 1999	TIMSS 2003	TIMSS 2007	TIMSS 2011	TIMSS 2015
I think mathematics is fun. (yfun)	Do you think that you enjoy learning mathematics? (BSBMENJY)	Do you think that you enjoy learning mathematics? (bsbmenjy)	I enjoy learning math. (bsbmtenj)	I enjoy learning mathematics. (BS4MAENJ)	How much do you agree that you enjoy learning mathematics? (BSBM14A)	I enjoy learning mathematics. (BSBM17A)
I would like to work at a job that lets me use mathematics. (yjobuse)	Do you think that you would like a job that involved using mathematics? (BSBMWORK)	Do you think that you would like a job that involved using mathematics? (bsbmwork)	I would like a job that involved using math. (bsbmajob)		How much do you agree that you would like a job that involves using mathematics? (BSBM16N)	I would like a job that involves using mathematics. (BSBM20E)
If I had my choice, I would not learn any more mathematics. (ynomore)			I would like to take more mathematics in school. (bsbmtmor)	I would like to do more mathematics in school. (BS4MAMOR)	How much do you agree that you wish you did not have to study mathematics? (BSBM14B)	I wish I did not have to study mathematics. (BSBM17B)
	Do you think that mathematics is boring? (BSBMBORE)	Do you think that mathematics is boring? (bsbmbore)		Mathematics is boring. (BS4MABOR)	How much do you agree that mathematics is boring? (BSBM14C)	Mathematics is boring. (BSBM17C)
	How much do you like mathematics? (BSBMLIKE)	How much do you like mathematics? (bsbmlike)		I like mathematics. (BS4MALIK)	How much do you agree that you like mathematics? (BSBM14E)	I like mathematics. (BSBM17E)
	I need to do well in mathematics to please myself. (BSBMSELF)	I need to do well in mathematics to please myself. (bsbmself)				
					How much do you agree that you learn many interesting things in mathematics? (BSBM14D)	I learn many interesting things in mathematics. (BSBM17D)

Table 4
Number of Selected Items in the Affective Scales of the Respective Studies.

	Extrinsic Motivation		Intrinsic Motivation	
	Scale	Overlap with previous	Scale	Overlap with previous
SIMS 1980	8	–	11	–
TIMSS 1995	7	3	5	2
TIMSS 1999	7	7	5	5
TIMSS 2003	4	3	3	2
TIMSS 2007	4	4	4	2
TIMSS 2011	5	4	6	4
TIMSS 2015	7	5	10	6

Table 5
Percentage of Middle Responses in SIMS, Extrinsic Motivation Scale.

Item	England	Hong Kong	Hungary	Israel	Japan	United States
ypwwell	10.26	21.58	32.74	3.89	36.68	9.68
ynouse	16.92	26.67	38.28	22.70	35.51	20.77
ymthjob	8.28	23.05	21.79	12.89	37.31	14.12
yuseday	13.63	20.74	25.39	14.81	32.75	17.86
ynoneed	10.34	18.54	12.09	14.99	47.69	14.07
ypract	24.70	19.79	25.33	21.12	51.55	21.28
ynotnec	10.03	20.76	23.62	23.00	44.20	10.10
ygowo	12.35	38.06	26.18	26.80	44.84	15.96

pooled sample composed of data from all countries and cycles. We assumed strong invariance of the anchor items across countries and over time. We used the estimated factor scores applying maximum likelihood estimation with robust standard errors (MLR) and the full information maximum likelihood (FIML) estimation in Mplus as a means of handling the missing data, while the items were treated as categorical variables. For the responses missing by design, we applied the pattern function in Mplus. This does not work together with the WLSMV estimation, but for items with more than three response options, the superiority of WLSMV over maximum likelihood estimation is less clear (Beauducel & Herzberg, 2006). We then transformed the factor scores onto a scale with a

Table 6
Percentage of Middle Responses in SIMS, Intrinsic Motivation Scale.

Item	England	Hong Kong	Hungary	Israel	Japan	United States
yiwant	8.48	13.32	15.86	4.07	50.84	6.67
yjobuse	39.45	42.06	42.56	40.29	20.77	41.70
yflgood	8.98	19.36	7.42	10.83	43.28	14.06
yhelpo	29.46	30.31	31.15	23.12	40.55	28.86
ynomore	14.21	27.04	17.80	13.97	52.86	19.30
ychall	20.21	18.52	32.80	24.98	43.11	22.29
ynotime	27.72	29.58	33.20	27.22	53.41	28.45
yhappy	42.20	40.77	28.18	28.42	42.80	44.45
yscared	22.69	35.34	19.91	15.79	48.02	17.90
yfun	31.17	23.72	38.33	31.71	50.51	28.00
ycalm	27.60	30.04	47.40	21.39	48.02	27.18
yinmaze	22.42	26.86	33.71	24.02	58.83	23.97
ymormth	34.92	20.55	40.50	27.40	58.36	28.45

mean of 5 and a standard deviation of 1. According to the SEM framework, an item y is predicted from the latent factor η as it is shown in the following equation:

$$y = \tau_y + \Lambda_y \eta + \varepsilon, \quad (4)$$

in which

- τ denotes the vector of item intercepts,
- Λ is the vector of factor loadings, and
- ε is the vector of residuals.

To estimate factor models from ordinal items, the MLR estimation procedure for continuous latent constructs was used because it is robust to non-normality. Mplus uses the maximum of the posterior distribution of the factor, which is known as the maximum *a posteriori* method (Muthén & Muthén, 1998–2017). The factor score estimate η for individual i is based on a regression method with correlated factors, where the factor score is computed as follows:

$$\hat{\eta_i} = \mu_y + C(v_i - \tau_y - \Lambda_y \mu_y), \quad (5)$$

in which

- μ is the mean vector of y items,
- C is the factor score coefficient matrix, and
- v_i is the vector of observations.

IRT scaling. First, we compared two models using the R package mirt (Chalmers, 2012), employing an expectation-maximization algorithm to achieve marginal maximum likelihood estimates of the item parameters and person scores as outlined by Bock and Aitkin (1981). In the first model, item parameters were estimated using the PCM following the scaling procedure in TIMSS. The PCM gives the probability that a student with proficiency θ_s will have, for item i, a response x_{is} that is scored in the l^{th} of m_i ordered score categories as:

$$P_{is}\left(x_{is} = l|\theta_s, \quad b_i, \quad d_{i,l}, \ldots, d_{i,m_i-1}\right) = \frac{\exp[\sum_{y=0}^{1}(\theta_s - b_i + d_{i,v})]}{\sum_{g=0}^{m_i-1}\exp[\sum_{y=0}^{g}(\theta_s - b_i + d_{i,v})]}, \quad (6)$$

in which

- x_{is} is the response of student s to item i (0 or 1 if correct),
- θ_s is the ability of student s,
- b_i is the location/difficulty parameter of item i,

- m_i is the number of response categories for item i, and
- $d_{i,l}$ is the category l threshold parameter of item i.

In the second model, item parameters were estimated using the GPCM Muraki (1992). The fundamental equation of this model gives the probability that a student with proficiency θ_s will have, for item i, a response x_{is} that is scored in the l^{th} of m_i ordered score categories as:

$$P_{is}\left(x_{is} = l|\theta_s, \quad b_i, a_i, \quad d_{i,l}, \ldots, d_{i,m_i-1}\right) = \frac{\exp[\sum_{y=0}^{1} a_i(\theta_s - b_i + d_{i,v})]}{\sum_{g=0}^{m_i-1}\exp[\sum_{y=0}^{g} a_i(\theta_s - b_i + d_{i,v})]}, \quad (7)$$

in which a_i is the slope/discrimination parameter of item i.

The model comparison showed that the GPCM model fit the data better for both scales. The Akaike information criteria (AIC; Akaike, 1974) and the Bayesian information criteria (BIC; Schwarz, 1978) were calculated. Both information criteria indices indicate the better fit of the GPCM model that allows the items to vary in terms of discrimination in contrast with the PCM model employed in the TIMSS contextual trend scales.

The item parameter estimation was conducted by concurrent calibration of all items in all studies, thus the parameters for all tests are automatically put onto the same scale. Item parameters were estimated simultaneously while the parameters of the anchor items were assumed identical in each sample. Third, we estimated the person scores and transformed them onto a scale with a mean of 5 and a standard deviation of 1.

Market-basket approach. The market-basket approach assumes that the items included in the assessment or survey define the construct. In this case, the assumption was that all the items from 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. Here we did not have an incomplete design, but the missing responses occurred as a consequence of changes in the questionnaires across cycles. We followed the procedure described by Zwitser et al. (2017) using a measurement model per country as a tool to generate plausible responses and fill the missing responses related to items that were not included in each cycle. Using the item parameters estimated by fitting the measurement models, we imputed missing responses five times per respondent and calculated sum scores, thereby estimating five plausible scores.

It is worth pointing out three aspects of this method. Firstly, an IRT model is not required, and any kind of measurement model can be used to generate plausible responses. We have used a GPCM model for consistency with the TIMSS procedure. Secondly, this GPCM model was employed for each country separately, so differences among countries (i.e., DIF) did not influence the generation of plausible responses. Thirdly, the results and comparisons were based on a sum score over the market basket of representative items and not on estimated latent variables. In this way, the comparability across countries was not threatened by differences between them. We transformed the plausible scores onto a scale with a mean of 5 and a standard deviation of 1.

Observed Scores. We used the sum of the observed scores per person at each time point and divided them by the number of answered items. A higher score indicates a more positive attitude. We then standardized these scores considering a mean of 5 and a standard deviation of 1 for the whole sample. We used this scale for presenting the results of the scaling.

Weights. Senate weights that sum to 500 for each country's data were applied (stratum weights in SIMS were rescaled to senate weights) in the scaling procedures, thus, the sample size differences of each country were taken into account and each country contributed equally to the estimation of the scales. In TIMSS 1995, there were two grades sampled in each country except for Israel, thus, we rescaled senate weights so that each grade was weighted equally within a country.

3. Results

3.1. Cross-cultural comparability

We tested measurement equivalence across countries at each time point. The MGCFA invariance testing of SIMS revealed that four items out of eight in the extrinsic motivation scale had negative factor loadings in the case of Japan in the baseline model. The same pattern emerged from fitting the baseline model in the case of the intrinsic motivation scale, five items out of 11 showed negative factor loadings. We concluded that measurement invariance does not hold for Japan and continued all further analyses excluding this country.

The thresholds and loadings equality constraints yielded acceptable model fit at most time points for the five-country multiple-group model as presented in Appendix D, However, the root mean square error of approximation (RMSEA; Steiger & Lind, 1984, as referred in Brown, 2015) values are in many cases too high while the absolute and comparative fit indices are mostly acceptable (except for the sample size-sensitive χ^2 values).

A possible explanation for poor RMSEA values is that this absolute fit measure is considered as a parsimony correction index, and it yields in poor fit because we have relatively high numbers of freely estimated parameters in these models. In addition, the poor relative fit indices (CFI, TLI) in the early assessments could be attributed to the mixed-worded scales. The presence of negatively-worded items potentially causes one-dimensional CFA models to show a poor fit (e.g., Marsh, 1996; Steinmann et al., 2021; Woods, 2006; Zhang et al., 2016). Finally, as Shi and Maydeu-Olivares (2020) argue, model fit values are influenced by many factors, such as estimation method or categorical/continuous specification and they suggest using only the SRMR because it is more consistent across these factors. We concluded that the measurement of the constructs included in the present study was invariant across countries at each time point. However, this does not imply that there is full invariance across time points.

3.2. Longitudinal comparability

We evaluated the assumption of the invariance of the anchor items across time by employing the delta plot method for each bridge, i.e., 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 bias. For simplicity, the plots of the pooled data are shown in Appendices E and F.

3.3. Trend scaling

We treated the countries as a single group in the CFA and IRT scaling procedures and separate in the market-basket scaling. The three methods yielded similar results on the individual- as well as the country levels. The correlations between individual scores were high across methods for both motivation constructs, ranging between 0.96 and 1. Figs. 1 and 2 show the country-level means for extrinsic- and intrinsic motivation by scaling methods. It is striking in the country-level trends that both models (CFA and IRT) with assumptions for full cultural- and longitudinal- invariance resulted in very similar results to the observed scores. The market-basket approach was employed to account for differences across countries, but the trend results did not show large deviations.

To explore a recently proposed method that does not rely on latent variable modeling and measurement invariance across countries, we used the market-basket approach. The three methods produced similar results, which, in the case of the CFA and IRT framework, is not surprising. As we mentioned previously, the market-basket approach can be combined with any of these measurement models to get plausible responses. We showed that the correlations between individual scores were high across scaling approaches. Because of the stratified multistage sampling design used in TIMSS, the simple random sampling assumed in the procedure for calculating standard errors of estimates does not apply (Rutkowski et al., 2010). Therefore, a limitation of the trend scales is that we have underestimated the standard errors of the means.

4. Discussion and conclusions

We investigated the trend component of two affective scales (i.e., extrinsic- and intrinsic motivation) by employing three different scaling methods, while we explored relevant characteristics related to measurement bias and longitudinal linking. We applied two widely used latent variable modeling approaches on real data drawn from mathematics ILSAs spanning over 35 years. We tackled issues of cross-cultural measurement of affective constructs addressing the issues of method-, construct-, and item bias. We showed how the assumptions regarding measurement invariance affected the analytical process when we excluded Japan from the latent variable analyses. The popularity of the middle option in Japan in 1980 was consistently and considerably high, indicating a possible cultural difference compared to the other countries.

We performed the analyses for five countries that participated in the seven cycles of SIMS and TIMSS. It is worth mentioning that the analysis can be extended to the other cycles and participating countries. One additional aspect to consider in that extension is how each method can provide results when new data is included in the analysis. For instance, the market-basket approach can be applied to include more data in longitudinal scales without the need for recalibrating the original scales or employing equating methods.

We cannot ignore the fact that the affective scales analyzed here until recently had not been designed for trend measurement. Hence, the maxim introduced by Beaton and Zwick (1990): "When measuring change, do not change the measure" (p.10), did not entirely apply to the data in this exploratory analysis. The scales varied in length and the number of anchor items between them is smaller in the early years.

Fig. 1. Extrinsic Motivation Scale.

Fig. 2. Intrinsic Motivation Scale.

Furthermore, modifications occurred in the number of response options from 1995. The way of handling the middle option in SIMS posed a limitation on the study. Another limitation of the study from the aspect of cross-cultural measurement over time is the possibility of changes in the translation and cultural adaptation procedures. The psychometric validation of the attempted scaling could be extended by inspecting the instruments in their original language. Finally, one of the most challenging remaining questions is whether changes in wording affect the internal relationships among items (e.g., factor structure). Since we explored non-identical sets of items over time, the number of items at almost each time point varies, which makes the investigation of the effects of changes in item wording challenging.

We believe that despite these challenges, the *old* international mathematics studies – SIMS and even the First International Mathematics Study (FIMS) administered in 1964 – provide rich data for secondary analyses. It is important to evaluate the possibilities of linking these studies to the recent ones because the potential country-level longitudinal analyses that can stem from such trend scales might serve as powerful approaches to investigate causal research questions. Future research on taking a closer look at the changes over time could reveal mechanisms in the relationship between motivation and achievement across countries. For instance, 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. It is potentially interesting to explore the relationship between the (decreasing) trends of the gender gap in mathematics achievement (Mullis, Martin, & Loveless, 2016) and trends of gender differences in mathematics motivation.

Acknowledgments

In developing the ideas presented here, we have received great support from Monica Rosén and Jan-Eric Gustafsson from the University of Gothenburg, Sweden. We are also grateful for the helpful input of Eugenio J. Gonzalez from Educational Testing Service, Princeton, NJ, USA. 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.

Appendix A. Items in the extrinsic motivation scales

Year	Original ID	New ID	Question/Statement
1980	ypwwell	EM1980_01	My parents really want me to do well in mathematics.
1980	ynouse	EM1980_02	Most people do not use mathematics in their job.
1980	ymthjob	EM1980_03	It is important to know mathematics in order to get a good job.
1980	yuseday	EM1980_04	Mathematics is useful in solving everyday problems.
1980	ynoneed	EM1980_05	Mathematics is not needed in everyday living.
1980	ypract	EM1980_06	Most of mathematics has practical use on the job.
1980	ynotnec	EM1980_07	A knowledge of mathematics is not necessary in most occupations.
1980	ygowo	EM1980_08	I can get along well in everyday life without using mathematics.

(*continued on next page*)

(*continued*)

Year	Original ID	New ID	Question/Statement
1995	BSBMFIP2	EM1995_01	My friends think it is important for me to do well in mathematics at school
1995	BSBMJOB	EM1980_03	I need to do well in mathematics to get desired job
1995	BSBMLIFE	EM1980_08	Do you think that mathematics is important to everyone's life?
1995	BSBMMIP2	EM1995_04	My mother thinks it is important for me to do well in mathematics at school
1995	BSBMPRNT	EM1980_01	I need to do well in mathematics to please my parents
1995	BSBMSCHL	EM1995_06	I need to do well in mathematics to get into the school I prefer.
1995	BSBMSIP2	EM1995_07	I think it is important to do well in mathematics at school
1999	bsbmfip2	EM1995_01	My friends think it is important for me to do well in mathematics at school.
1999	bsbmjob	EM1980_03	I need to do well in mathematics to get the job I want
1999	bsbmlife	EM1980_08	Do you think that mathematics is important to everyone's life?
1999	bsbmmip2	EM1995_04	My mother thinks it is important for me to do well in mathematics at school.
1999	bsbmprnt	EM1980_01	I need to do well in mathematics to please my parents
1999	bsbmschl	EM1995_06	I need to do well in mathematics to get into the school I prefer.
1999	bsbmsip2	EM1995_07	I think it is important to do well in mathematics at school.
2003	bsbmaget	EM1980_03	I need to do well in math to get the job I want
2003	bsbmahdl	EM1980_08	I think learning mathematics will help me in my daily life.
2003	bsbmaoss	EM2003_03	I need mathematics to learn other school subjects.
2003	bsbmauni	EM1995_06	I need to do well in math to get into the <university> of my choice.
2007	BS4MAGET	EM1980_03	I need to do well in mathematics to get the job I want
2007	BS4MAHDL	EM1980_08	I think learning mathematics will help me in my daily life.
2007	BS4MAOSS	EM2003_03	I need mathematics to learn other school subjects.
2007	BS4MAUNI	EM1995_06	I need to do well in mathematics to get into the <university> of my choice.
2011	BSBM14F	EM1995_07	How much do you agree that it is important to do well in mathematics?
2011	BSBM16J	EM1980_08	How much do you agree that learning mathematics will help you in your daily life?
2011	BSBM16K	EM2003_03	How much do you agree that you need mathematics to learn other school subjects?
2011	BSBM16L	EM1995_06	How much do you agree that you need to do well in mathematics to get into the <university> of your choice?
2011	BSBM16M	EM1980_03	How much do you agree that you need to do well in mathematics to get the job you want?
2015	BSBM20A	EM1980_08	How much do you agree with these statements about mathematics? I think learning mathematics will help me in my daily life
2015	BSBM20B	EM2003_03	How much do you agree with these statements about mathematics? I need mathematics to learn other school subjects
2015	BSBM20C	EM1995_06	How much do you agree with these statements about mathematics? I need to do well in mathematics to get into the <university> of my choice
2015	BSBM20D	EM1980_03	How much do you agree with these statements about mathematics? I need to do well in mathematics to get the job I want
2015	BSBM20F	EM2015_05	How much do you agree with these statements about mathematics? It is important to learn about mathematics to get ahead in the world
2015	BSBM20G	EM2015_06	How much do you agree with these statements about mathematics? Learning mathematics will give me more job opportunities when I am an adult
2015	BSBM20H	EM1980_01	How much do you agree with these statements about mathematics? My parents think that it is important that I do well in mathematics
2015	BSBM20I	EM1995_07	How much do you agree with these statements about mathematics? It is important to do well in mathematics

Appendix B. Items in the intrinsic motivation scales

Year	Original ID	New ID	Question/Statement
1980	yiwant	IM1980_01	I really want to do well in mathematics.
1980	yjobuse	IM1980_02	I would like to work at a job that lets me use mathematics.
1980	yflgood	IM1980_03	I feel good when I solve a mathematics problem by myself.
1980	yhelpo	IM1980_04	I like to help others with mathematics problems.
1980	ynomore	IM1980_05	If I had my choice I would not learn any more mathematics.
1980	ychall	IM1980_06	I feel challenged when I am given a difficult mathematics problem.
1980	ynotime	IM1980_07	I refuse to spend a lot of my own time doing mathematics.
1980	yhappy	IM1980_08	Working with numbers makes me happy.
1980	yscared	IM1980_09	It scares me to have to take mathematics.
1980	yfun	IM1980_10	I think mathematics is fun.
1980	ycalm	IM1980_11	I usually feel calm when doing mathematics problems.
1995	BSBMBORE	IM1995_01	Do you think that mathematics is boring?
1995	BSBMENJY	IM1980_10	Do you think that you enjoy learning mathematics?
1995	BSBMLIKE	IM1995_03	How much do you like mathematics?
1995	BSBMSELF	IM1995_04	I need to do well in mathematics to please myself.
1995	BSBMWORK	IM1980_02	Do you think that you would like a job that involved using mathematics?
1999	bsbmbore	IM1995_01	Do you think that mathematics is boring?
1999	bsbmenjy	IM1980_10	Do you think that you enjoy learning mathematics?
1999	bsbmlike	IM1995_03	How much do you like mathematics? *REVERSED*
1999	bsbmself	IM1995_04	I need to do well in mathematics to please myself.
1999	bsbmwork	IM1980_02	Do you think that you would like a job that involved using mathematics?
2003	bsbmajob	IM1980_02	I would like a job that involved using math.
2003	bsbmtenj	IM1980_10	I enjoy learning math.
2003	bsbmtmor	IM1980_05	I would like to take more mathematics in school.
2007	BS4MABOR	IM1995_01	Mathematics is boring.
2007	BS4MAENJ	IM1980_10	I enjoy learning mathematics.
2007	BS4MALIK	IM1995_03	I like mathematics.
2007	BS4MAMOR	IM1980_05	I would like to do more mathematics in school
2011	BSBM14A	IM1980_10	How much do you agree that you enjoy learning mathematics?
2011	BSBM14B	IM1980_05	How much do you agree that you wish you did not have to study mathematics?
2011	BSBM14C	IM1995_01	How much do you agree that mathematics is boring?
2011	BSBM14D	IM2011_04	How much do you agree that you learn many interesting things in mathematics?

(*continued on next page*)

(*continued*)

Year	Original ID	New ID	Question/Statement
2011	BSBM14E	IM1995_03	How much do you agree that you like mathematics?
2011	BSBM16N	IM1980_02	How much do you agree that you would like a job that involves using mathematics?
2015	BSBM17A	IM1980_10	How much do you agree with these statements about learning mathematics? I enjoy learning mathematics
2015	BSBM17B	IM1980_05	How much do you agree with these statements about learning mathematics? I wish I did not have to study mathematics
2015	BSBM17C	IM1995_01	How much do you agree with these statements about learning mathematics? Mathematics is boring
2015	BSBM17D	IM2011_04	How much do you agree with these statements about learning mathematics? I learn many interesting things in mathematics
2015	BSBM17E	IM1995_03	How much do you agree with these statements about learning mathematics? I like mathematics
2015	BSBM17F	IM2015_06	How much do you agree with these statements about learning mathematics? I like any schoolwork that involves numbers
2015	BSBM17G	IM2015_07	How much do you agree with these statements about learning mathematics? I like to solve mathematics problems
2015	BSBM17H	IM2015_08	How much do you agree with these statements about learning mathematics? I look forward to mathematics class
2015	BSBM17I	IM2015_09	How much do you agree with these statements about learning mathematics? Mathematics is one of my favorite subjects
2015	BSBM20E	IM1980_02	How much do you agree with these statements about mathematics? I would like a job that involves using mathematics

Appendix C. Cronbach's Alpha reliability coefficients of the motivation scales

	1980		1995		1999		2003		2007		2011		2015
	EM (8)	IM (11)	EM (7)	IM (5)	EM (7)	IM (5)	EM (4)	IM (3)	EM (4)	IM (4)	EM (5)	IM (6)	EM (8)	IM (10)
HKG	.79	.78	.67	.74	.76	.80	.76	.77	.81	.87	.84	.89	.90	.94
HUN	.71	.83	.65	.80	.70	.82	.63	.73	.65	.85	.74	.86	.87	.93
ISR	.72	.79	.61	.78	.67	.77	.69	.73	.73	.81	.76	.87	.89	.93
JPN	.48	.47	.62	.74	.68	.76	.66	.71	.70	.85	.77	.87	.85	.93
USA	.72	.83	.68	.81	.72	.83	.72	.78	.73	.86	.79	.88	.89	.94
ENG	.73	.83	.70	.80	.70	.81	.70	.72	.72	.85	.77	.87	.87	.94

Note. EM = extrinsic motivation; IM = intrinsic motivation; (number of items)

Appendix D. MGCFA fit results

		Extrinsic Motivation Scale			Intrinsic Motivation Scale		
	Fit Indices/ Equality Constraints	Baseline	Thresholds	Thresholds and Loadings	Baseline	Thresholds	Thresholds and Loadings
SIMS 1980	χ^2	2234.983	2521.057	3509.755	5960.486	6694.920	12,866.863
	χ^2 df	100	132	160	220	268	352
	χ^2 p value	< .0001	< .0001	< .0001	< .0001	< .0001	< .0001
	RMSEA	.074	.068	.074	.082	.079	.096
	CFI	.948	.941	.918	.902	.890	.787
	TLI	.927	.938	.928	.878	.888	.833
	SRMR	.039	.040	.050	.050	.051	.065
TIMSS 1995	χ^2	5808.191	6361.405	5940.727	791.109	492.886	7678.952
	χ^2 df	70	98	122	25	45	81
	χ^2 p value	< .0001	< .0001	< .0001	< .0001	< .0001	< .0001
	RMSEA	.121	.107	.092	.074	.069	.130
	CFI	.910	.901	.908	.994	.991	.943
	TLI	.864	.894	.921	.989	.990	.965
	SRMR	.058	.059	.063	.020	.024	.060
TIMSS 1999	χ^2	6037.850	6575.345	5709.640	580.197	1000.492	6213.390
	χ^2 df	70	98	122	25	45	81
	χ^2 p value	< .0001	< .0001	< .0001	< .0001	< .0001	< .0001
	RMSEA	.133	.117	.098	.068	.067	.126
	CFI	.897	.888	.904	.997	.995	.967
	TLI	.846	.881	.917	.994	.994	.980
	SRMR	.063	.064	.066	.018	.023	.058
TIMSS 2003	χ^2	1983.096	2190.913	1888.181	.000[a]	357.732	1696.465
	χ^2 df	10	26	38	0	12	32
	χ^2 p value	< .0001	< .0001	< .0001	< .0001	< .0001	< .0001
	RMSEA	.204	.132	.101	.000	.078	.105
	CFI	.946	.934	.944	1.000	.992	.959
	TLI	.839	.924	.956	1.000	.989	.981
	SRMR	.043	.044	.045	.000	.018	.035
TIMSS 2007	χ^2	2013.168	1934.883	1665.798	121.689	247.431	3087.673
	χ^2 df	10	26	38	10	26	54
	χ^2 p value	< .0001	< .0001	< .0001	< .0001	< .0001	< .0001
	RMSEA	.214	.130	.099	.051	.044	.113
	CFI	.957	.954	.961	.999	.999	.984

(*continued on next page*)

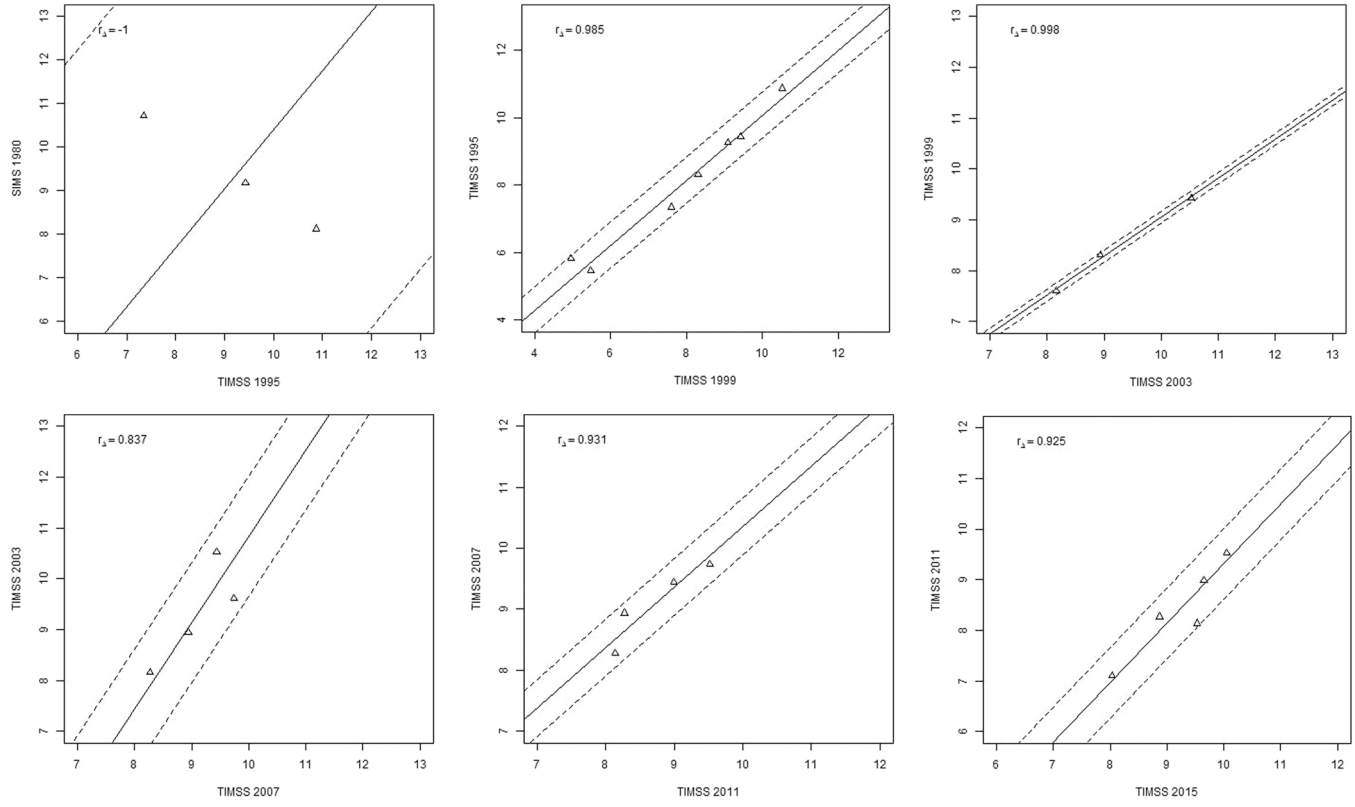
(*continued*)

		Extrinsic Motivation Scale			Intrinsic Motivation Scale		
	Fit Indices/ Equality Constraints	Baseline	Thresholds	Thresholds and Loadings	Baseline	Thresholds	Thresholds and Loadings
	TLI	.870	.947	.969	.998	.999	.991
	SRMR	.042	.040	.041	.006	.011	.040
TIMSS 2011	χ^2	3488.355	3278.904	2897.775	4433.711	4862.247	3905.620
	χ^2 df	25	45	61	45	69	113
	χ^2 p value	< .0001	< .0001	< .0001	< .0001	< .0001	< .0001
	RMSEA	.158	.113	.091	.132	.112	.078
	CFI	.957	.955	.961	.988	.987	.990
	TLI	.915	.950	.968	.980	.986	.993
	SRMR	.040	.040	.042	.027	.029	.037
TIMSS 2015	χ^2	10,016.033	10,224.567	8882.919	14,752.416	6824.035	7809.566
	χ^2 df	100	132	160	175	275	291
	χ^2 p value	< .0001	< .0001	< .0001	< .0001	< .0001	< .0001
	RMSEA	.130	.115	.097	.120	.064	.067
	CFI	.959	.958	.964	.984	.993	.992
	TLI	.942	.955	.968	.979	.994	.994
	SRMR	.043	.044	.044	.028	.035	.036

[a]This model is just-identified with zero degrees of freedom. Model fit cannot be assessed in this case.

Appendix E. Delta plots of the extrinsic motivation scale bridges

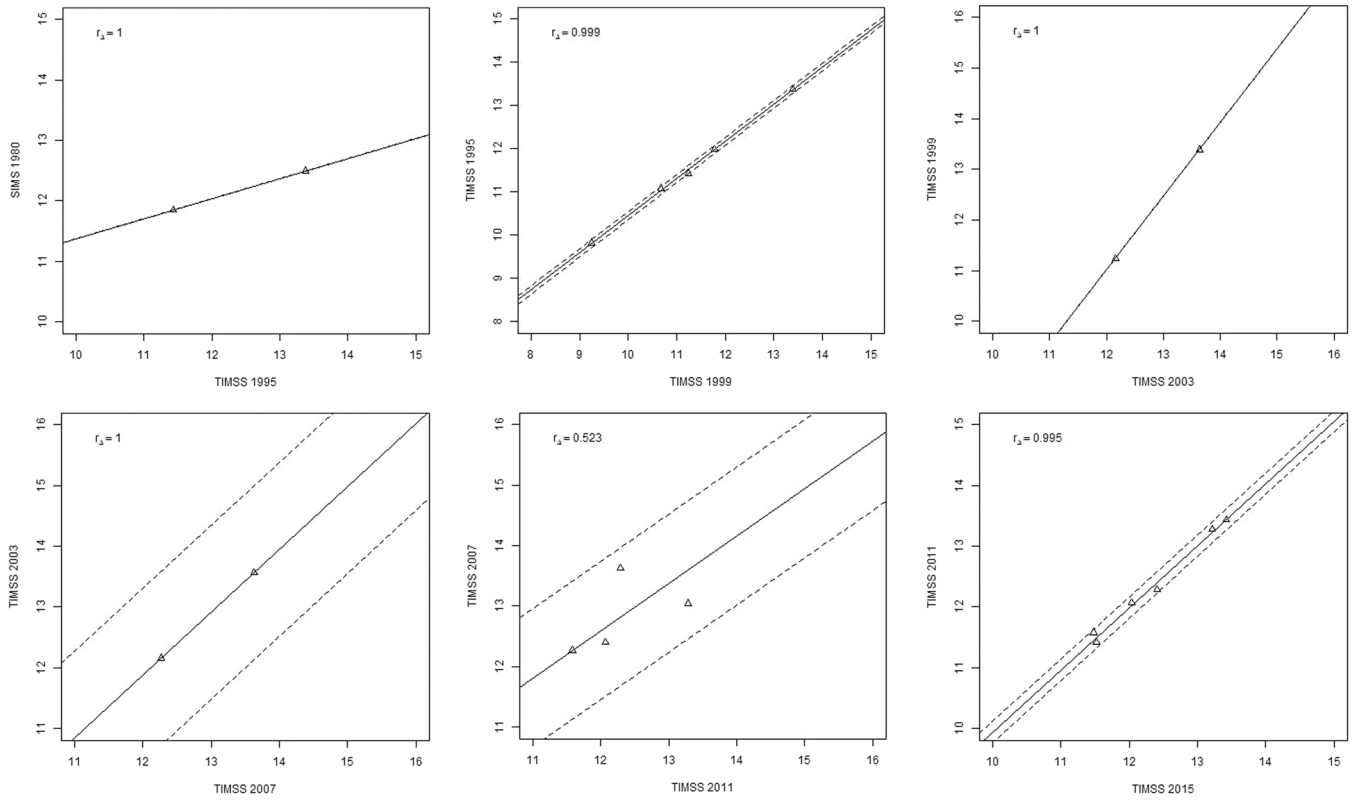
See Appendix section here



.

Appendix F. Delta plots of the intrinsic motivation scale bridges

See Appendix section here



.

References

Afrassa, T. M. (2005). Monitoring mathematics achievement over time: A secondary analysis of FIMS, SIMS and TIMS: A Rasch analysis. In R. Maclean, S. Alagumalai, R. Baker, Y. C. Boediono, D. D. Cheng, W. Curtis, et al. (Eds.), *Education in the Asia-Pacific Region: Vol. 4. Applied Rasch measurement: A book of exemplars: Papers in honour of John P. Keeves* (pp. 61–77). Springer.

Altinok, N., Angrist, N., & Patrinos, H., 2018, Global data set on education quality (1965–2015). ⟨http://documents.worldbank.org/curated/en/706141516721172989/Global-data-set-on-education-quality-1965–2015⟩ (Policy research working paper; no. WPS 8314).

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control, 19*(6), 716–723. https://doi.org/10.1109/TAC.1974.1100705.

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

Avvisati, F., Le Donné, N., & Paccagnella, M. (2018). Conclusion: An OECD conference on the cross-cultural comparability of questionnaire measures in large-scale assessments. In F. J. R. van de Vijver, F. Avvisati, E. Davidov, M. Eid, J.-P. Fox, N. Le Donné, K. Lek, B. Meuleman, M. Paccagnella, & R. van de Schoot (Eds.), OECD Education Working Papers: Vol. 201. Invariance analyses in large-scale studies.

Beaton, A. E., & Zwick, R. (1990). The effect of changes in the national assessment: Disentangling the NAEP 1985-86 reading anomaly. *Revised. (NAEP Report No. 17–TR–21)*. Educational Testing Service,.

Beauducel, A., & Herzberg, P. Y. (2006). On the performance of maximum likelihood versus means and variance adjusted weighted least squares estimation in CFA. *Structural Equation Modeling: A Multidisciplinary Journal, 13*(2), 186–203. https://doi.org/10.1207/s15328007sem1302_2

Brown, T. A. (2015). Confirmatory factor analysis for applied research. *Methodology in the Social Sciences* (Second ed..,). 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

Cortina, J. M. (1993). What is coefficient alpha? An examination of theory and applications. *The Journal of Applied Psychology, 78*(1), 98–104. https://doi.org/10.1037/0021-9010.78.1.98

Davidov, E., Meuleman, B., Cieciuch, J., Schmidt, P., & Billiet, J. (2014). Measurement equivalence in cross-national research. *Annual Review of Sociology, 40*(1), 55–75. https://doi.org/10.1146/annurev-soc-071913-043137

Dedrick, R. F., Marfo, K., & Harris, D. M. (2007). Experimental analysis of question wording in an instrument measuring teachers' attitudes toward inclusive education. *Educational and Psychological Measurement, 67*(1), 116–131. https://doi.org/10.1177/0013164406292034

D'Urso, E. D., Roover, K., de, Vermunt, J. K., & Tijmstra, J. (2020). Scale length does matter:: Recommendations for measurement invariance testing with categorical factor analysis and item response theory approaches. *Advance Online Publication*. https://doi.org/10.31234/osf.io/udbna

Eccles, J. S., & Wigfield, A. (2002). Motivational beliefs, values, and goals. *Annual Review of Psychology, 53*, 109–132. https://doi.org/10.1146/annurev.psych.53.100901.135153

Hanushek, E., & Wößmann, A. (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

Kim, E. S., & Yoon, M. (2011). Testing measurement invariance: A comparison of multiple-group categorical CFA and IRT. *Structural Equation Modeling: A Multidisciplinary Journal, 18*(2), 212–228. https://doi.org/10.1080/10705511.2011.557337

Kolen, M. J., & Brennan, R. L. (2014). *Test equating, scaling, and linking: Methods and practices (3. ed.). Statistics for social and behavioral sciences.* 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., Rosén, M., & Gustafsson, J.-E., 2020, Four decades of measuring attitude towards mathematics. Electronic board session cancelled. 2020 Annual Meeting of the National Council on Measurement in Education, San Francisco, CA, United States.

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

Marco, G. L. (1977). Item characteristic curve solutions to three intractable testing problems. *Journal of Educational Measurement, 14*(2), 139–160. ⟨https://www.jstor.org/stable/1434012⟩.

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., 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. ⟨https://timssandpirls.bc.edu/publications/timss/2015-methods/chapter-15.html⟩.

Masters, G. N. (1982). A rasch model for partial credit scoring. *Psychometrika, 47*(2), 149–174. https://doi.org/10.1007/BF02296272

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

Mullis, I., Martin, M. O., & Loveless, T. (2016). *20 Years of TIMSS: International trends in mathematics and science achievement, curriculum, and instruction.* Boston College,.

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement, 16*(2), 159–176. https://doi.org/10.1177/014662169201600206

Organisation for Economic Co-operation and Development (Ed.), 2019, TALIS 2018 technical report. ⟨http://www.oecd.org/education/talis/TALIS_2018_Technical_Report.pdf⟩.

Muthén, L. K., & Muthén, B. O. (1998-2017). *Mplus User's guide* (8th ed.). Muthén & Muthén.

Putnick, D. L., & Bornstein, M. H. (2016). Measurement invariance conventions and reporting: The state of the art and future directions for psychological research. *Developmental Review: DR, 41*, 71–90. https://doi.org/10.1016/j.dr.2016.06.004

Schuman, H., & Presser, S., 1996, Questions and answers in attitude surveys: Experiments on question form, wording, and context. Sage.

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. https://doi.org/10.3102/0013189X10363170.

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

Steiger, J.H., & Lind, J.C., 1984, Statistically based tests for the number of common factors. Paper presentation. Annual Meeting of the Psychometric Society, Iowa City, IA.

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., & 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

Svetina, D., Rutkowski, L., & Rutkowski, D. (2020). Multiple-Group Invariance with Categorical Outcomes Using Updated Guidelines: An Illustration Using M plus and the lavaan/semTools Packages. *Structural Equation Modeling: A Multidisciplinary Journal, 27*(1), 111–130. https://doi.org/10.1080/10705511.2019.1602776

van de Vijver, F. J. R. (2015). Methodological aspects of cross-cultural research. In M. J. Gelfand, C. Chiu, & Y. Hong (Eds.), *Advances in Culture and Psychology: Vol. 5. Handbook of Advances in Culture and Psychology* (pp. 101–160). Oxford University Press. https://doi.org/10.1093/acprof:oso/9780190218966.003.0003.

van de Vijver, F. J. R. (2018). Towards an integrated framework of bias in noncognitive assessment in international large-scale studies: Challenges and prospects. *Educational Measurement: Issues and Practice, 37*(4), 49–56. https://doi.org/10.1111/emip.12227

van de Vijver, F. J. R., Avvisati, F., Davidov, E [E.], Eid, M., Fox, J.-P., Le Donné, N., Lek, K., Meuleman, B [B.], Paccagnella, M., & van de Schoot, R. (Eds.), 2018, *OECD Education Working Papers: Vol. 201 . Invariance analyses in large-scale studies.* https://doi.org/10.1787/254738dd-en.

van de Vijver, F. J. R., & Leung, K. (1997). *Methods and Data Analysis for Cross-cultural Research. Cross-cultural Psychology Series: Vol. 1*. Sage,. http://www.loc.gov/catdir/enhancements/fy0655/96051274-d.html.

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: TIMSS 2019 technical teport.* von Davier, & I. V. S. Mullis (Eds.), *Methods and procedures: TIMSS 2019 technical teport* . ⟨https://timssandpirls.bc.edu/timss2019/methods/chapter-16.html⟩.

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), Article 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