Applied Measurement in Education

ISSN: 0895-7347 (Print) 1532-4818 (Online) Journal homepage: https://www.tandfonline.com/loi/hame20

Negative Keying Effects in the Factor Structure of TIMSS 2011 Motivation Scales and Associations with Reading Achievement

Michalis P. Michaelides

To cite this article: Michalis P. Michaelides (2019) Negative Keying Effects in the Factor Structure of TIMSS 2011 Motivation Scales and Associations with Reading Achievement, Applied Measurement in Education, 32:4, 365-378, DOI: 10.1080/08957347.2019.1660349

To link to this article: https://doi.org/10.1080/08957347.2019.1660349

Published online: 26 Sep 2019.

APPLIED MEASUREMENT IN EDUCATION
2019, VOL. 32, NO. 4, 365–378
https://doi.org/10.1080/08957347.2019.1660349

Routledge
Taylor & Francis Group

Negative Keying Effects in the Factor Structure of TIMSS 2011 Motivation Scales and Associations with Reading Achievement

Michalis P. Michaelides

Department of Psychology, University of Cyprus

ABSTRACT

The Student Background survey administered along with achievement tests in studies of the International Association for the Evaluation of Educational Achievement includes scales of student motivation, competence, and attitudes toward mathematics and science. The scales consist of positively- and negatively keyed items. The current research examined the factorial structure of the 18-item motivational scales in fourth-grade mathematics in the 2011 Trends in International Mathematics and Science Study (TIMSS). Survey data from six European countries were analyzed. In comparisons of alternative models, the fit was adequate when three correlated factors were specified and negative keying was taken into account as a latent factor, or with correlated uniquenesses among negatively keyed items. Participants reading achievement scores correlated systematically to negative keying with coefficients ranging from .254 to .395 in the six samples. Unlike their higher-scoring peers, fourth-graders with lower reading achievement responded differentially to similar items depending on the direction of item keying, in such a way that their motivation scores were biased downward. Implications about the use of reverse keying in surveys for young students are discussed.

1. Introduction

Attitudes toward school subjects, self-concept and academic motivation consistently appear as predictors of academic performance in studies of school achievement. Although there is no consensus in the literature on which are the fundamental motivational constructs (Marsh, Craven, Hinkley, & Debus, 2003), the multidimensional nature for self-concept is generally acceptable (Marsh, 1993). Shavelson, Hubner, and Stanton (1976) provided one of the first descriptions of the multifaceted, hierarchical self-concept construct. That model distinguishes between general and academic self-concept, as well as between subject-specific self-concept. Domain or task specificity has been supported by various lines of research such as those based on academic self-concept (e.g. Marsh, Byrne, & Shavelson, 1988), self-efficacy (Bandura, 1997), and expectancy-value theory (e.g. Eccles & Wigfield, 2002), to name a few. However, multiple theoretical frameworks have resulted in various ways of measuring attitudes, motivation, or self-concept relevant to academic subjects. Moreover, methodological problems the measurement of these constructs have been reported (Liu & Meng, 2010), such as the use of single-item, or non-validated measures of motivation.

International large-scale assessments (ILSAs) provide a unique and invaluable opportunity to study motivational constructs and their relationship to educational outcomes. ILSAs have expanded and evolved in recent years to measure student knowledge, skills, as well as attitudes toward various school subjects at different age cohorts, and across countries. Examples of such assessment programs are the Trends in International Mathematics and Science Study (TIMSS), the Progress in

CONTACT Michalis P. Michaelides michalim@ucy.ac.cy Faculty of Social Sciences and Education, University of Cyprus, Nicosia 1678, Cyprus

International Reading Literacy Study (PIRLS), both conducted by the International Association for the Evaluation of Educational Achievement (IEA), and the Programme for International Student Assessment (PISA), administered by the Organization for Economic and Co-operative Development. In addition to achievement estimates obtained through complex test designs and data collection methods, ILSAs gather rich auxiliary information on students, teachers, schools, and student homes via background questionnaires (Rutkowski, Rutkowski, & von Davier, 2014). Adopting a research-oriented, empirical approach, ILSAs provide insights about policies and practices that foster progress in education across multiple cultural contexts.

1.1. *Motivational Scales in TIMSS across Time*

Since its first administration in 1995 and every 4 years thereafter, TIMSS has included self-report scales of student motivation and affect in mathematics and science administered along with the achievement tests in the two subjects. Attitudes toward mathematics and science, including enjoyment, perceived self-competence, value, and engagement in a subject were considered significant correlates of test performance and have featured in the Student Background questionnaires among other contextual scales (cf. Martin, Mullis, Arora, & Preuschoff, 2014). Due to its large-scale and cross-cultural nature, TIMSS data have been used to study the relationship between student attitudes toward the two subjects and achievement. This relationship has been repeatedly shown to be positive and strong (Mullis, Martin, Foy, & Arora, 2012). Unfortunately, these scales have been changing across successive cycles of the study; originally, a set of survey items comprised a general measure of attitudes toward a subject, but in more recent administrations items have been increasing in number, and grouped under distinct concepts such as self-competence, value, or affect toward a subject. However, the scales have been criticized for the lack of a solid theoretical foundation, as well as psychometric evaluation (Marsh et al., 2013).

A few studies have attempted to provide information on this lack of validity evidence on the TIMSS mathematics motivational scales. Liu and Meng (2010) examined the factor structure of the TIMSS 2003 12 questionnaire items on eighth-grade mathematics self-concept. In exploratory factor analyses for four countries, they found evidence in favor of a two-factor solution: mathematics self-concept and self-perception of mathematics importance.

A three-factor framework was adopted in TIMSS 2007: value, self-concept, and affect about mathematics, with moderate empirical support for the factorial structure (Olson, Martin, & Mullis, 2008). Two studies examined the factorial validity of the TIMSS 2007 attitudinal scales applying latent variable methodology. Selecting seven items from fourth-grade TIMSS 2007, Yang, Chen, Lo, and Turner (2012) proposed two factors termed positive affect toward mathematics and self-confidence in learning mathematics, with improved model fit when a method factor onto which negatively worded items loaded was included. Partial scalar invariance was observed between Taiwanese and US samples. In a more comprehensive investigation of motivation items with eighth-grade data from four English-speaking and four Arab countries, Marsh et al. (2013) provided confirmatory factor analysis (CFA) results in support of a four-factor solution – they singled-out an item that they called coursework. Acceptable metric invariance and partial intercept invariance across countries were found after accounting for wording effects and parallel wording. Correlated uniquenesses were allowed among items that were negatively worded and among math and science items that were similarly phrased. Moreover, they provided supporting construct-validity evidence relating the motivational factors with achievement, plans to take more coursework and long-term educational aspirations. Marsh et al. (2013) found evidence for convergent validity between self-concept and affect for mathematics with strong correlations ranging from .62 to .71 in the English-speaking samples and .77 to .88 in the Arab samples.

1.2. *Effects due to Keying or Wording in the Measurement of Psychological Constructs*

Scales measuring psychological constructs have been found to be vulnerable to various response sets (Cronbach, 1946) such as acquiescence, or to inattentive responding, without much effort

during the administration of an instrument, e.g. satisficing (Krosnick, 1999). The inclusion of both positively and negatively worded items has been suggested as a way to reduce such effects (e.g. Nunnally, 1978). In the current study, the term negative keying is considered a type of method effect that encompasses negatively worded items, i.e. items that include words like "not", but also words with affixal morphemes like "un-" or "-less", or words that are antonymic to the trait measured by the scale. The assumption underlying the mixing positively and negatively keyed items is that they measure the same construct, are psychometrically interchangeable, and increase validity (Benson & Hocevar, 1985). Many validation studies for multiple scales, employing exploratory or confirmatory factor analysis or IRT methods, have not supported this assumption. Inclusion of positively and negatively worded items has been found to threaten construct validity and result in artefactual factors. Such findings suggest that the two types of wording may elicit different response processes and/or differential attention in respondents (Weems, Onwuegbuzie, Schreiber, & Eggers, 2003).

Beyond Marsh et al.'s (2013) study, method effects have been found in popular scales such as the Rosenberg Self-Esteem Scale, and the Life Orientation Test-Revised (LOT-R). Often factor analyses of these scales have failed to confirm their purported dimensionality (e.g. DiStefano & Motl, 2006; Quilty, Oakman, & Risko, 2006; Rauch, Schweizer, & Moosbrugger, 2007). For both the Rosenberg and the LOT-R scales which are supposed to be unidimensional, factor analyses typically result in two factors each consisting of either positively or negatively keyed items. When one or two method factors are introduced to capture wording direction using latent variable models, a single substantive factor is supported and models have improved fit (Michaelides, Zenger, Koutsogiorgi, Brähler, Stöbel-Richter, & Berth, 2016).

The tendency to respond differently depending on item keying, a methodological effect, can be conceptualized as a response style. In addition to being identified as a separate source of variability in responses, it has been found that method or wording factors relate to a number of personality characteristics. For example, neuroticism, avoidance motivation, depression, fear of negative evaluation and self-consciousness have a negative correlation with a negative wording factor (e.g. DiStefano & Motl, 2006; Michaelides, Koutsogiorgi, & Panayiotou, 2016; Quilty et al., 2006). Individuals high on such characteristics exhibit fewer negative wording effects, while individuals with lower scores tend to respond differently on positive vis-à-vis negative items.

Beyond the substantive interpretation for the negative factor described above, Schmitt and Stults (1985) suggested that this methodological factor emerges if some respondents – as few as 10% from the sample – fail to attend closely to the keying of the items. In this case, the individual respondent resorts to systematic, "careless" responding after having read the first few items, and continues responding in a similar manner throughout the scale, ignoring items that are keyed negatively.

As regards children, a hypothesis has been put forth by Marsh (1986, 1996) that linguistic proficiency influences their ability to respond appropriately to negatively worded items. When a negative wording effect was modeled as a factor in fifth-graders' responses on a self-description questionnaire, it was correlated to reading achievement (Marsh, 1986). Moreover, second graders' responses to positive and negative items of the questionnaire were uncorrelated, however, the two sets of items were substantially correlated with the responses of (linguistically more proficient) fifth-graders, as expected, since both item sets purport to measure similar concepts (Marsh, 1986). The same pattern was replicated with data from the National Education Longitudinal Study of 1988, where the positive and negative self-esteem factors were more highly correlated as students more vis-à-vis less proficient in reading were examined (Marsh, 1996). In a sample of German ninth-graders, Gnambs and Schroeders (2017) found higher variability due to negatively worded items of the Rosenberg scale, for students with low reading, vocabulary and cognitive abilities compared to their high performing counterparts.

The contextual framework of the TIMSS 2011 study utilizes a theoretical structure for motivation that is similar for both mathematics and science: finding the subject enjoyable, placing value on the subject, and self-confidence in learning the subject (Mullis, Martin, Ruddock, O'Sullivan, &

Preuschoff, 2009). In the publication of results for TIMSS 2011, Mullis et al. (2012) name the three scales as follows: Students Like Learning Mathematics scale, Students Value Mathematics scale, and Student Confidence with Mathematics scale. However, this structure was implemented only for grade eight; for grade four, two of the scales measuring *positive affect* ("Enjoyment," "Like"), and *self-confidence* were included, while a third one measured student *engagement* with the lesson and the teacher. We found only two studies that examined wording effects with the TIMSS 2011 administration, both using US grade eight data. Wang, Chen, and Jin (2015) used bi-factor IRT models to examine selected items from the 2011 administration and found support for moderate wording effects in both math and science, although they did not address the dimensionality of the scales. Wang, Kim, Dedrick, Ferron, and Tan (2018) applied multilevel bifactor models on the mathematics confidence scale and found evidence of both a negative and a positive wording factor.

1.3. *Aims of the Study*

The purpose of the current study was to contribute evidence on the construct validity of motivation measures from the TIMSS assessment. The combination of positively and negatively stated items in self-report scales is consequential both for corroborating their factor structure through latent variable techniques, and for evaluating their appropriateness for young populations with different levels of reading proficiency. Specific aims of this study were to (a) examine the factor structure of the three contextual scales for fourth-grade mathematics (enjoyment of, engagement with, and competence in mathematics) appearing in the TIMSS 2011 administration, allowing for the presence of keying effects, (b) evaluate the contribution of reverse keying in the variability of responses, and (c) estimate the relationship of reading ability on keying effect factors across six different European language samples. The first hypothesis was that an oblique three factor structure would fit the data well after keying effects were accounted for across all samples. The second hypothesis was that higher reading proficiency would be associated with higher disagreement on negatively keyed items, irrespective of language.

2. Method

2.1. *Data Sources and Measures*

Data were obtained from the IEA Data Repository (https://www.iea.nl/data). From the TIMSS 2011 mathematics grade four administration, responses on 18 items comprising the "Mathematics in School" section of the Student Background Questionnaire were used. The items appear in the Appendix. Responses were given on a four-point, Likert-type scale ranging from 1 = "Agree a lot" to 4 = "Disagree a lot." Six of the items were phrased negatively. The first section includes six items that measure *enjoyment*, with the exception of the last item that relates to the value students place on learning mathematics. The second section consists of five items referring to *engagement* with the teacher and the lesson. The third section includes seven items that measure perceived *competence* in mathematics.

The PIRLS program for the assessment of reading achievement in fourth-grade was also conducted in 2011 and coincided with the TIMSS administration. IEA provides information from both PIRLS and TIMSS assessments in joint datasets for countries that participated in both programs. The role of reading proficiency is central in the current study, so the Joint TIMSS & PIRLS 2011 Grade Four database was utilized, since the PIRLS study provides estimates of performance on a fourth-grade reading comprehension assessment. To achieve sound measurement and adequate coverage of the content of a subject matter within the constraints of a reasonable test length for students, a multiple matrix sampling design is implemented in IEA studies; five plausible values (PVs) for each student are provided based on responses to test items and background variables for precise estimation of population characteristics while reflecting the uncertainty in individual estimates (Rutkowski,

Gonzalez, von Davier, & Zhou, 2014; von Davier, Gonzalez, & Mislevy, 2009). The five PVs from PIRLS were used as indicators of individual reading achievement.

The target population in both TIMSS and PIRLS are all students in the grade of interest, fourth-grade in this case. A two-stage random sample design is employed with a selection of schools at the first stage and a selection of one or more intact classes from each school at the second stage in each country (Joncas & Foy, 2012). Data from six European countries that participated in both TIMSS and PIRLS in 2011 were analyzed: Finland, Germany, Italy, Northern Ireland, Romania, and the Russian Federation; the countries were selected to reflect diverse languages within Europe. Data on the three motivational subscales were of primary interest, and sample sizes with responses on those subscales by country were as follows: 3548 for Germany, 4095 for Italy, 3454 for Northern Ireland, 4445 for Russia, 4516 for Finland, and 4601 for Romania. Inspection of individual item distributions revealed that there were not more than 5.12% missing cases for any single item, while the average percentage of missing values across the six country datasets was 1.93%.

2.2. *Statistical Analysis*

A series of CFA models were analyzed in MPLUS 7.31 (Muthén & Muthén, 1998-2017): a unidimensional model (M1), a three correlated factors model[1] (M2), a second-order factor model where the three subscales load onto a general motivation factor (M3), three correlated factors with a negative method factor (M4), three correlated factors with correlated uniquenesses on negatively keyed items (M5), and three correlated factors with a negative and a positive method factors (M6). In Models M4 and M6, the method factors were not allowed to correlate with the motivational factors. With the exception of model M5 where correlated uniquenesses among negatively keyed items were specified, no additional covariances were allowed between error variances. Diagrams for Models M4 and M5 that allow for negative keying effects are presented in Figure 1. The robust maximum likelihood estimator (MLR) was used for the CFA models, which outperforms the conventional maximum likelihood when observed variables are ordinal[2] (Li, 2016a). Full information maximum likelihood was employed for missing data.

Provided that a method factor would be a part of a model with an acceptable fit, PIRLS reading achievement was specified as an exogenous, observed variable, correlated with the motivational subscales and the method factor. All five PVs were included in the analysis (von Davier et al., 2009) and the results were aggregated using imputation utilities in MPLUS.

Evidently, there is a linguistic element to how negatively keyed items are expressed in different languages, and how response effects relate to language proficiency; also, psychometric adaptations typically investigate whether factor structures are similar cross-culturally. Therefore, all analyses were conducted separately on each country sample to examine whether the method effect due to item keying holds across diverse language samples.

Overall model fit was evaluated with the chi-square statistic. Additional goodness-of-fit indices were examined: the Tucker Lewis Index (TLI), the Comparative Fit Index (CFI), the Root Mean Square Error of Approximation (RMSEA), and the Standardized Root Mean Square Residual (SRMR), as well as the Bayesian Information Criterion (BIC). Values above .90 for TLI and CFI, below .06 for RMSEA, and .08 for SRMR were considered as evidence of acceptable model fit; lower BIC values implied better fit. To test differences between nested models, a Satorra-Bentler scaled chi-square difference test (Satorra & Bentler, 2010) was conducted since MLR estimation was used. Invariance analysis on the best fitting model was examined across the six language samples.

[1]Multilevel analysis with the three-factor model allowed for the calculation of intraclass correlations (ICC) in the six country samples. The average ICC was 0.045. Subsequently, single-level analysis was conducted on each sample separately.

[2]A diagonally weighted least squares estimator (WLSMV in MPLUS) was also examined, as a recommended method (Li, 2016a, 2016b). However, the final model (with the plausible values included) failed to converge in all six samples. In the comparison of alternative models under WLSMV, only models M4 and M5 had acceptable fit indices (results available upon request), as was the case with the MLR estimator, hence results from the MLR analysis are presented in the paper.

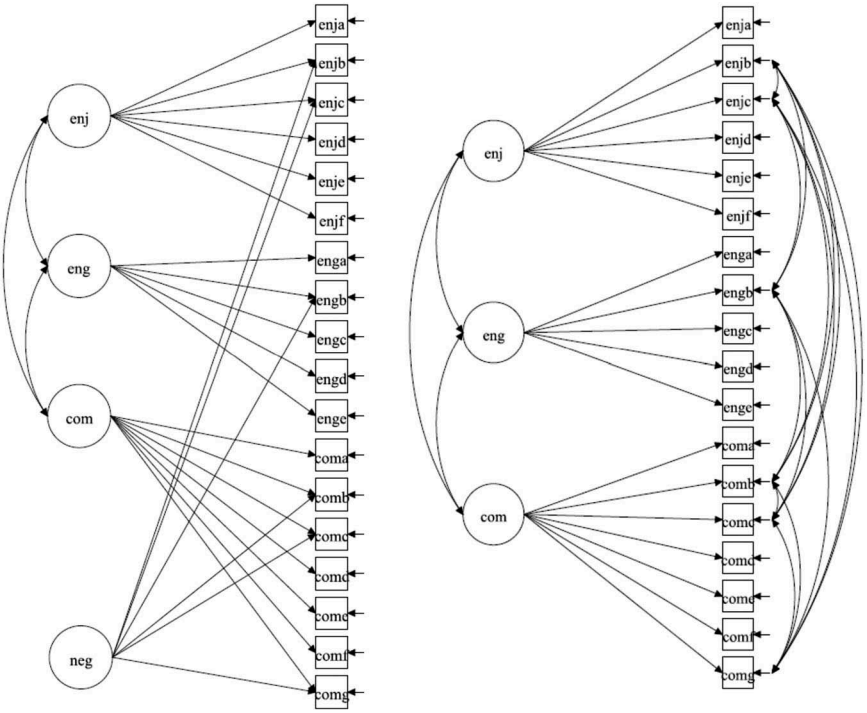

Figure 1. Models M4 (left) and M5 (right).

Notes. enj: enjoyment factor, eng: engagement factor, com: competence factor, neg: negative factor. Observed variables represent the 18 items of the motivational scales.

3. Results

The comparison of models revealed that the unidimensional model (M1) had unacceptable fit indices for all samples (Table 1). The three-correlated factors model (M2) had a significantly better fit compared to M1 according to the Satorra-Bentler scaled chi-square difference test (all $\Delta\chi^2 > 700$ for 3 df), and improved but still unacceptable fit indices. The second-order factor model (M3) performed similarly to the three-correlated factors model for three samples, but resulted in inadmissible solutions for the other three samples. The model with three substantive and two method factors (M6) did not converge for any of the samples.

The models that accounted for negative keying, either via a latent method factor (M4), or via correlated uniquenesses (M5), had adequate fit indices: TLI values ranged from .904 to .944, CFI from .922 to .957, RMSEA .044 to .056, and SRMR from .037 to .050; robust chi-square values for all models were significant. BIC values were also lower for M4 and M5 than for the other models. The Satorra-Bentler scaled chi-square difference tests comparing these models with their respective three-factor M2 models were all significant implying improved fit when negative keying was accounted for (M4 compared to M2: all $\Delta\chi^2 > 872$ for 6 df; M5 compared to M2: all $\Delta\chi^2 > 1168$ for 15 df). Although not excellent, the fit indices for M4 and M5 provide reasonable support for the a priori hypothesized factor structure of three trait factors plus negative method effects. Post hoc searches for major sources of lack-of-fit revealed residual correlations among items within one factor that were similarly worded, as well as crossloadings; e.g., item 1d "I learn many interesting things in mathematics" from the enjoyment scale was related to the engagement factor in some of the samples. There were some similarities in the suggested parameters with large modification indices for some, but not all samples. No changes were pursued to further improve model fit.

Table 1. Fit indices for the alternative models by country.

Model	χ^2	TLI	CFI	RMSEA	SRMR	BIC	SB $\Delta\chi^2$ (Δdf)
M1 – Unidimensional (df = 135)							
FI	9847.489*	.667	.706	.126	.099	187,913	
GE	7123.600*	.626	.670	.121	.100	142,146	
IT	4971.446*	.703	.738	.094	.078	164,488	
NI	6512.024*	.638	.680	.117	.096	143,530	
RO	5814.677*	.670	.709	.096	.088	176,526	
RU	6182.384*	.685	.722	.100	.083	168,829	
M2 – Three correlated factors (df = 132)							
FI	3047.473*	.898	.912	.070	.054	178,956	3336 (3)
GE	2375.136*	.877	.894	.069	.055	135,562	2479 (3)
IT	2764.398*	.835	.858	.070	.062	161,159	1107 (3)
NI	2851.942*	.842	.864	.077	.069	138,485	1800 (3)
RO	4232.300*	.756	.790	.082	.081	173,645	702 (3)
RU	3081.274*	.843	.865	.071	.060	164,049	1342 (3)
M3 – Second order factor							
Same fit indices as for M2 for IT, RO, and RU; inadmissible solutions for FI, GE, and NI due to a negative error variance estimate for a disturbance term for the Enjoyment factor							
M4 – Three correlated factors with a negative latent method factor (df = 126)							
FI	1788.908*	.939	.950	.054	.044	177,358	921 (6)
GE	1198.478*	.939	.949	.049	.042	133,996	873 (6)
IT	1572.637*	.905	.922	.053	.047	159,461	934 (6)
NI	1472.850*	.918	.932	.056	.050	136,674	1052 (6)
RO	1523.248*	.913	.928	.049	.043	169,256	2367 (6)
RU	1406.148*	.929	.941	.048	.040	161,629	1278 (6)
M5 – Three correlated factors with correlated uniquenesses among negative items (df = 117)							
FI	1539.853*	.944	.957	.052	.041	177,124	1345 (15)
GE	1055.446*	.942	.956	.048	.040	133,879	1168 (15)
IT	1474.422*	.904	.927	.053	.047	159,400	1191 (15)
NI	1222.107*	.928	.945	.052	.047	136,430	1525 (15)
RO	1334.368*	.918	.938	.048	.042	169,046	2940 (15)
RU	1113.817*	.940	.954	.044	.037	161,300	1874 (15)
M6 – Three correlated factors and two method factors (df = 114)							
All	No convergence						
M4 with Reading PVs as a correlate of the latent factors (df = 140)							
FI	1951.390*	.935	.947	.053	.044	228,054	
GE	1278.397*	.937	.949	.048	.041	174,327	
IT	1664.108*	.903	.921	.051	.046	206,409	
NI	1648.506*	.911	.927	.056	.050	176,803	
RO	1654.149*	.911	.927	.048	.043	223,650	
RU	1537.045*	.925	.939	.047	.039	211,812	

Note: *$p < .001$ FI = Finland, GE = Germany, IT = Italy, NI = Northern Ireland, RO = Romania, RU = Russia, PV = Plausible Value, TLI = Tucker Lewis Index, CFI = Comparative Fit Index, RMSEA = Root Mean Square Error of Approximation, SRMR = Standardized Root Mean Square Residual, BIC = Bayesian Information Criterion, SB $\Delta\chi^2$ (Δdf) = Satorra-Bentler scaled chi-square difference test (difference in degrees of freedom)

Estimated parameters in the two best fitting models were reviewed. The 18 items loaded significantly on their respective factors and in the expected direction in both M4 and M5: loadings were strong on the enjoyment and confidence factors, and moderate to strong on the engagement factor. All negatively keyed items had, in addition to their primary factor loading, a moderate but significant loading on the latent method factor (M4) or significant correlated uniquenesses (M5) in all six country samples (see Supplementary Material for standardized loadings by factor for the revised M4 with PVs as a correlate). In agreement with the first hypothesis, this finding suggests that there was systematic covariation among reversely worded items in student responses which, if accounted for, improves model fit.

M5 with correlated uniquenesses among pairs of negatively keyed items does not allow for further investigation of psychometric characteristics of the response tendency due to negative keying (e.g. reliability, associations with reading achievement); M4 that specifies this tendency as a latent factor

was selected for subsequent analysis. A multi-group invariance study of the M4 model examined the extent of equivalence across the six samples. The results for configural invariance, $\chi^2(756) = 8951.885$, $p < .001$, CFI = .938, TLI = .925, RMSEA = .051, SRMR = .044, suggested acceptable fit of the factor structure – same number of factors and correspondence of items to factors – across all six groups. However, difference in fit statistics did not support metric invariance, the equivalence of factor loadings: $\Delta\chi^2(100) = 1921.544$, $p < .001$, ΔCFI = .014, ΔRMSEA = .002.

Negatively keyed statements are usually longer or more complex than positively keyed ones. Therefore, what has been conceptualized as a negative factor, may be confounded with the linguistic complexity of the items. Utilizing only the English-speaking sample, an additional model for the Northern Ireland data was tested: a three-factor model with a specific "complexity" factor composed of items 1B, 2A, 2B, 2D, 3B, and 3G.[3] However, the fit for this alternative model was worse than the fit of M4 for Northern Ireland. The loadings of the positively keyed items on the "complexity" factor were very low and modification indices suggested that the negatively keyed items which were not judged as "complex" enough should also be allowed to load on the specific factor, suggesting that the direction an item was phrased was more important than linguistic complexity for forming a specific methodological factor.

A final alternative model was also tested only on the English-speaking sample: the negatively keyed items were grouped into two "negative method factors" depending on whether (a) the word "not" was present in the statement (negatively worded), or (b) if the negation was formed based on the content of a word (e.g. boring, harder). The analysis led to an inadmissible solution with a correlation between the two "negative method factors" above 1, which suggests that all negatively keyed items should comprise a single latent factor instead of two factors with two different ways of forming negatively keyed items.

Omega indices were calculated on the unstandardized estimates of the model (Reise, 2012) with a latent factor for negative items (model M4) to estimate the proportion of variance attributable to the negative method factor vis-à-vis the three substantive factors. Overall omega reliability for the 18-item scale ranged from .957 (Romania), .962 (Italy, Russia), .965 (Germany), .966 (Northern Ireland), and .968 (Finland) in the six samples. Importantly, the average omega subscale index for the negative factor was .244, and varied from .196 (Italy), .201 (Northern Ireland), .259 (Germany), .235 (Romania), .285 (Russia), and .290 (Finland). The percentage of reliable variance attributable to the negative factor, i.e. omega subscale for the negative factor divided by the overall omega, was not negligible: it was on average 26.5%, and ranged from 20.4% (Italy) to 30.0% (Finland).

The model with a latent factor for negative items (M4) was subsequently revised to include reading achievement as a correlate of the negative keying factor. In all six samples, the fit indices were adequate (see Table 1, M4 with Reading PVs as a correlate of the latent factors). The standardized factor loadings from this model from all samples can be seen in the Supplementary Material. The correlation coefficients between reading achievement and negative keying are presented in Table 2 and were positive and significant ranging from 0.254 (Finland) to 0.395 (Romania), in accordance with the second hypothesis. Controlling for motivation level, students with high reading achievement tended to disagree more strongly[4] with negatively keyed items compared to their low achieving peers.

As evidence of discriminant validity, reading achievement was unrelated to enjoyment of mathematics in most cases; it was weakly correlated with mathematics competence with higher scores associated with stronger agreement to statements expressing competence. A similar but even weaker pattern was found for reading achievement with engagement with mathematics. Had the negative keying factor not been included in the model, the bias in the correlation coefficients of reading

[3]The six items were ranked as more complex than the rest according to their syntactic structure. This measure of structural complexity may be judged as subjective, but it took into consideration grammatical properties such as embedded questions ("I know what my teacher expects me to do"), reduced relative clauses ("I think of things not related to the lesson"), and comparatives ("Mathematics is harder for me than any other subject"). Four of the six complex items were negatively keyed.

[4]Higher scores on the TIMSS motivational scales imply more disagreement.

Table 2. Correlation coefficients between wording, achievement and motivations factors in mathematics.

	Correlations of Reading Achievement with				Motivation factors intercorrelations		
Country	Negative keying	Enjoyment	Engagement	Competence	Enjoyment-Engagement	Enjoyment-Competence	Engagement-Competence
Finland	.254***	.022	−.068***	−.138***	.720***	.673***	.468***
Germany	.266***	.020	−.040	−.218***	.628***	.670***	.345***
Italy	.312***	−.001	−.068**	−.146***	.663***	.726***	.515***
N. Ireland	.323***	.009	.010	−.160***	.624***	.689***	.394***
Romania	.395***	−.149***	−.260***	−.307***	.740***	.759***	.624***
Russia	.357***	−.035*	−.068***	−.175***	.696***	.719***	.515***

*** $p < .001$, ** $p < .01$, * $p < .05$

Table 3. Competence item means by wording type and reading level – Romanian sample.

	Sample		
Item	Overall (N = 4643)	Low reading achievement (N = 2322)	High reading achievement (N = 2321)
Positively-keyed			
A	1.53	1.69	1.38
D	1.52	1.64	1.40
E	2.02	2.18	1.88
F	1.86	2.04	1.70
Negatively-keyed (after reverse-scoring)			
B	2.05	2.46	1.66
C	1.76	2.07	1.46
G	1.92	2.23	1.63
Overall mean of positively-keyed items	1.74	1.89	1.59
Overall mean of negatively-keyed items	1.92	2.27	1.59

achievement with the motivation factors would range from −0.04 to 0.09. Convergent validity was supported by the correlations among the motivation latent factors which were positive and significant as anticipated: strong between enjoyment with engagement and with competence, and moderate for engagement with competence.

Looking at mean item responses is helpful for further clarifying the relationship between reading achievement and the negative keying factor. Results from the competence subscale for the Romanian sample are presented in Table 3 to elucidate the above findings. Looking at the overall sample column, the item means are largely overlapping, but for the positively worded items are slightly lower (1.52–2.02) than for the negatively keyed[5] ones (1.76–2.05), indicating that overall, students express slightly higher agreement with statements suggesting competence in mathematics when they are phrased positively, rather than negatively. Then, the sample was split in two equal halves based on median reading achievement according to the PIRLS first PV. For the high reading achievement students, means for the positively worded items (1.38–1.88) are similar to those of the negatively keyed ones (1.46–1.66). In fact, the averages of all positively keyed versus all negatively keyed items are identical (1.59), and are interpreted as relatively strong agreement to competence items irrespective of item keying. For low achieving students, means for the positively worded items (1.64–2.18) indicate some agreement to competence items, while means for the negatively keyed ones (2.07–2.46) show little agreement and approach the middle of the 1-to-4 response scale. Stated differently, students with low reading achievement who agree with positively worded items are reluctant to disagree as strongly with negatively keyed items even though the items purport to measure the

[5]Negatively worded items have been reverse-coded only for this analysis (Table 3), such that low values indicated strong disagreement with those negatively worded items, i.e. high competence.

same construct. Therefore, it appears that reading comprehension relates to the ability of low achieving students to respond consistently to negative items as they do with the positive ones. Considering only positively keyed items, these students report slightly lower perceived competence in mathematics than their higher achieving counterparts (1.89 versus 1.59, the latter signifying more agreement to competence items). If negatively stated items are also considered, then their perceived competence level is biased downward even further (2.27 versus 1.59, the latter signifying more disagreement with negative competence items).

4. Discussion

The current study provided evidence on the factorial validity of the motivation scales used in TIMSS 2011. Unlike previous studies that have made selective use of the available items (Wang et al., 2015; Yang et al., 2012), all items of the enjoyment, engagement, and competence in mathematics scales in fourth-grade TIMSS 2011 were utilized for a comprehensive psychometric evaluation of the three motivation subscales as operationalized in the assessment program. Moreover, the implementation of latent variable methodology has allowed for a response style conceptualization of the reaction to negatively keyed items modeled as systematic covariation among those items. Findings revealed that this response style could be detected in linguistically diverse samples and was associated with estimates of reading achievement.

Although the scales are not identical in the administration rounds of the TIMSS assessment across years, they are used in a variety of contexts for substantive purposes, such as the relationship between motivation, subjective beliefs, and attitudes about a school subject, and achievement, as well as for cross-country comparisons. As has been argued by Marsh et al. (2013), theoretical and psychometric support regarding the dimensional structure of the scales is required. The alternative CFA models presented above indicate that the intended oblique three-factor structure could not be confidently supported; in all samples examined, there were non-negligible effects due to negatively keyed items that need to be accounted for to achieve acceptable model fit, as hypothesized. The effects appeared under both the latent method factor and the correlated uniquenesses approach, and were generalizable across samples from six linguistically diverse countries. Model equivalence was supported only up to the configural level suggesting invariant factor structure across groups. Non-invariance at the metric level implies that not all item loadings are equivalent across groups; students from different countries do not attach the same meaning to the same item. This finding is not surprising because multiple group invariance, especially in the context of ILSAs with many groups and large sample sizes, has been characterized as "too cumbersome to be practical due to the many possible violations of invariance" (Asparouhov & Muthén, 2014, p.495).

Even though negative keying is commonly encountered in attitudinal scales in order to reduce acquiescent response style by requiring more attentive responding (Podsakoff, MacKenzie, Lee, & Podsakoff, 2003), this study has found that it complicates the factor structure of the instrument. The evidence suggests that response patterns to items with opposite wording are different (Wang et al., 2015, 2018). In this particular case, students more likely to agree with positively worded items were less likely to disagree with the negatively keyed ones as would have been expected had the two types of items been interchangeable. The above findings are not unique for the TIMSS motivation scale, since similar results have been reported for other scales as well (DiStefano & Motl, 2006; Marsh et al., 2013; Michaelides, Koutsogiorgi, et al., 2016; Rauch et al., 2007). It is though the first illustration of a consistent keying effect in a large-scale survey with fourth-graders across multiple languages.

Another original contribution of this study has been the examination of alternative specifications for a nuisance factor due to the direction of wording: analysis on the English-speaking sample from Northern Ireland showed that the methodological factor was comprised neither by the linguistically more complex items, nor by items with specific ways expressing negation – including the word "not" versus including other words with negative connotations. It was the set of all negatively keyed items that introduced systematic trait-irrelevant variability in responses.

Research designed to address item linguistic complexity and its interaction with item keying could provide further support for this new finding.

The use of the latent variable approach allowed for the specification of response due to negative keying as a factor orthogonal to the trait factors. Omega indices showed that negative keying contributes to the overall reliability of the instrument. About a quarter of the reliable variance could be attributed to the method factor. Moreover, the factor correlated systematically with reading achievement scores across samples in accordance with the second hypothesis. Thus, the evidence lends support to the notion of response due to wording as a style factor (Horan, DiStefano, & Motl, 2003; Michaelides, Koutsogiorgi, & Panayiotou, 2017).

Reading proficiency was related to the extent student responses were influenced by negatively keyed items. As has been shown, fourth-graders who were linguistically less proficient, responded differentially to positive and negative items, in a manner that their total scores were biased downward. Difficulty in processing negatively keyed items may potentially explain this finding. This result concurs with Marsh's (1986, 1996) findings that method effects decrease for more verbally proficient and older students. Associations of wording effects with cognitive abilities can be traced even in adolescents (Gnambs & Schroeders, 2017). It may thus be seen as a cognitive developmental phenomenon. As a result, the appropriateness of administering negatively keyed items may be questioned in studies with young and linguistically less proficient children. Failure to recognize the difference of negatively keyed from positively keyed items as individuals respond carelessly, by responding similarly to all items irrespective of keying (Schmitt & Stults, 1985) might also explain this finding. Such a claim would require evidence that participants become more attentive with higher proficiency and as they grow older. Moreover, in the TIMSS scales examined, negatively keyed items appear early in the "Mathematics in School" scales (positions 2 and 3; see Appendix), so it is unlikely that students have already developed a response tendency from the first positively keyed item and retained a systematic and careless response pattern on all items, ignoring that some items are phrased negatively.

Explaining the variation in the relationship found between reading proficiency and negative keying across the six language samples can be tentative at best. How negation is expressed in different languages, and therefore how effortlessly young students can handle negative statements might be a reason why the correlation was stronger in the Romanian or Russian rather than the Finnish or German samples. Alternatively, this variation might be attributed to cultural differences in responding to questionnaires. Response tendencies appear to differ across ethnic groups and relate to cultural values and affluence (e.g. He, Dominguez Espinosa, Poortinga, & Van de Vijver, 2014). There is also evidence that responses to positively worded items are highly consistent to responses to negatively keyed items in some countries but not in others, and this variation was related to human development indices (Schmitt & Allik, 2005).

It is worth-noting some substantive findings on the construct validity of the TIMSS 2011 motivation scales. The trait factors of self-reported enjoyment of, engagement with, and competence for mathematics were moderately-to-strongly inter-correlated. The three constructs appear to be distinct, in agreement with the multidimensional conceptualization of academic motivation (Shavelson et al., 1976). The strongest relationships in most samples were found between enjoyment and competence in mathematics, and the levels of association were similar to those reported by Marsh et al. (2013), slightly higher than English-speaking samples, and lower than Arab samples they had examined. Mathematics motivational factors with PIRLS reading achievement were only weakly correlated, if at all; this was not unexpected and is consistent with the notion of domain-specificity in the self-concept literature. The patterns of associations revealed that competence was consistently associated with achievement, unlike engagement or enjoyment (Marsh et al., 2013), even if reading – and not math – achievement was used in the analysis. Finally, the overall high omega reliabilities with the majority of the variance attributable to the motivational traits is a psychometrically positive finding for these contextual scales.

A potential implication of this study is to reconsider the use of negative keying in future revisions of the background questionnaires in ILSAs, particularly with younger study participants. Avoiding negative statements is an option, as is rewording them as positive; however, the purported benefit to reduce response sets would be lost. Alternatively, they could be retained in the instruments without counting them for scoring purposes (Marsh, 1996). In practice, researchers studying substantive questions may choose to omit those items from their analysis (e.g., Sherer & Nilsen, 2016). However, while this practice may circumvent construct-irrelevant variance due to item keying, another validity threat is introduced: the content captured by the "curtailed" scales is limited with consequences on the under-representation of the construct measured. Test developers probably introduce negatively keyed items with the intention to capture the content of the construct the scale purports to measure, and not just for reducing acquiescence or promoting more careful responding. Essentially, reverse keyed items could be ignored following explicit guidelines from scale developers or scoring rubrics that those items function as unscored, "filler" items with no bearing on construct coverage.

At a minimum, if the responses to negative items are scored, analysis should account for keying effects with appropriate statistical techniques, to reduce biases in the estimation of relationships with other variables. As demonstrated, latent variable models are useful in examining the presence of keying effects and separating substantive interpretations from methodological artifacts. CFA approaches can be easily implemented, especially with large samples such as those encountered in ILSA studies.

Disclosure statement

No potential conflict of interest was reported by the author.

Funding

This work was supported by the University of Cyprus [Starting grant to MM].

References

Asparouhov, T., & Muthén, B. (2014). Multiple-group factor analysis alignment. *Structural Equation Modeling: A Multidisciplinary Journal*, *21*, 495–508. doi:10.1080/10705511.2014.919210

Bandura, A. (1997). *Self-efficacy: The exercise of control*. New York, NY: W.H. Freeman and Company.

Benson, J., & Hocevar, D. (1985). The impact of item phrasing on the validity of attitude scales for elementary school children. *Journal of Educational Measurement*, *22*, 231–240. doi:10.1111/jedm.1985.22.issue-3

Cronbach, L. J. (1946). Response sets and test validity. *Educational and Psychological Measurement*, *6*, 475–494. doi:10.1177/001316444600600405

DiStefano, C., & Motl, R. W. (2006). Further investigating method effects associated with negatively worded items on self-report surveys. *Structural Equation Modeling*, *13*, 440–464. doi:10.1207/s15328007sem1303_6

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

Gnambs, T., & Schroeders, U. (2017). Cognitive abilities explain wording effects in the Rosenberg Self-Esteem Scale. *Assessment*. Advance online publication. doi:10.1177/1073191117746503

He, J., Dominguez Espinosa, A., Poortinga, Y. H., & Van de Vijver, F. J. R. (2014). Acquiescent and socially desirable response styles in cross-cultural value surveys. In L. T. B. Jackson, D. Meiring, F. J. R. Van de Vijver, E. Idemudia, & W. K. Gabrenya Jr. (Eds.), *Toward sustainable development through nurturing diversity* (pp. 98–111). Melbourne, FL: International Association for Cross-Cultural Psychology. Retrieved from www.iaccp.org

Horan, P. M., DiStefano, C., & Motl, R. W. (2003). Wording effects in self-esteem scales: Methodological artifact or response style? *Structural Equation Modeling*, *10*, 444–455. doi:10.1207/S15328007SEM1003_6

Joncas, M., & Foy, P. (2012). Sample design in TIMSS and PIRLS. In M. O. Martin & I. V. S. Mullis (Eds.), *Methods and procedures in TIMSS and PIRLS 2011*(pp.1-21). Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.

Krosnick, J. A. (1999). Survey research. *Annual Review of Psychology*, *50*, 537–567. doi:10.1146/annurev.psych.50.1.537

Li, C.-H. (2016a). The performance of ML, DWLS, and ULS estimation with robust corrections in structural equation models with ordinal variables. *Psychological Methods*, *21*, 369–387. doi:10.1037/met0000093

Li, C.-H. (2016b). Confirmatory factor analysis with ordinal data: Comparing robust maximum likelihood and diagonally weighted least squares. *Behavior Research Methods*, *48*, 936–949. doi:10.3758/s13428-015-0619-7

Liu, S., & Meng, L. (2010). Re-examining factor structure of the attitudinal items from TIMSS 2003 in cross-cultural study of mathematics self-concept. *Educational Psychology*, *30*, 699–712. doi:10.1080/01443410.2010.501102

Marsh, H. W. (1986). Negative item bias in ratings scales for preadolescent children: A cognitive-developmental phenomenon. *Developmental Psychology*, *22*, 37–49. doi:10.1037/0012-1649.22.1.37

Marsh, H. W. (1993). Academic self-concept: Theory, measurement, and research. In J. Suls (Ed.), *Psychological perspectives on the self* (pp. 59–98). New York, NY: Lawrence Erlbaum Associates Inc.

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

Marsh, H. W., Abduljabbar, A. S., Abu-Hilal, M. M., Morin, A. J., Abdelfattah, F., Leung, K. C., … Parker, P. (2013). Factorial, convergent, and discriminant validity of TIMSS math and science motivation measures: A comparison of Arab and Anglo-Saxon countries. *Journal of Educational Psychology*, *105*, 108–128. doi:10.1037/a0029907

Marsh, H. W., Byrne, B. M., & Shavelson, R. J. (1988). A multifaceted academic self-concept: Its hierarchical structure and its relation to academic achievement. *Journal of Educational Psychology*, *80*, 366–380. doi:10.1037/0022-0663.80.3.366

Marsh, H. W., Craven, R. G., Hinkley, J. W., & Debus, R. L. (2003). Evaluation of the Big-Two-Factor Theory of academic motivation orientations: An evaluation of jingle-jangle fallacies. *Multivariate Behavioral Research*, *38*, 189–224. doi:10.1207/S15327906MBR3802_3

Martin, M. O., Mullis, I. V. S., Arora, A., & Preuschoff, C. (2014). Context questionnaire scales in TIMSS and PIRLS 2011. In L. Rutkowski, M. von Davier, & D. Rutkowski (Eds.), *Handbook of international large-scale assessment* (pp. 299–316). Boca Raton, FL: CRC Press.

Michaelides, M. P., Koutsogiorgi, C., & Panayiotou, G. (2016). Method effects on an adaptation of the Rosenberg Self-Esteem Scale in Greek and the role of personality traits. *Journal of Personality Assessment*, *98*, 178–188. doi:10.1080/00223891.2015.1089248

Michaelides, M. P., Koutsogiorgi, C., & Panayiotou, G. (2017). Method/group factors: Inconsequential but meaningful. A comment on Donnellan, Ackerman and Brecheen (2016). *Journal of Personality Assessment*, *99*, 334–335. doi:10.1080/00223891.2016.1233560

Michaelides, M. P., Zenger, M., Koutsogiorgi, C., Brähler, E., Stöbel-Richter, Y., & Berth, H. (2016). Personality correlates and gender invariance of wording effects in the German version of the Rosenberg Self-Esteem Scale. *Personality and Individual Differences*, *97*, 13–18. doi:10.1016/j.paid.2016.03.011

Mullis, I. V., Martin, M. O., Foy, P., & Arora, A. (2012). *TIMSS 2011 international results in mathematics*. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Lynch School of Education, Boston College, and Amsterdam: International Association for the Evaluation of Educational Achievement.

Mullis, I. V. S., Martin, M. O., Ruddock, G. J., O'Sullivan, C. Y., & Preuschoff, C. (2009). *TIMSS 2011 assessment frameworks*. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Lynch School of Education, Boston College, and Amsterdam: International Association for the Evaluation of Educational Achievement.

Muthén, L. K., & Muthén, B. O. (1998-2017). *Mplus user's guide* (8th ed.). Los Angeles, CA: Muthén & Muthén.

Nunnally, J. M. (1978). *Psychometric theory*. New York, NY: McGraw-Hill.

Olson, J. F., Martin, M. O., & Mullis, I. V. S. (Eds.). (2008). *TIMSS 2007 technical report*. Chestnut Hill, MA: TIMSS & PIRLS International Study Center.

Podsakoff, P. M., MacKenzie, S. B., Lee, J. Y., & Podsakoff, N. P. (2003). Common method biases in behavioral research: a critical review of the literature and recommended remedies. *The Journal of Applied Psychology*, *88*, 879–903. doi:10.1037/0021-9010.88.5.879

Quilty, L. C., Oakman, J. M., & Risko, E. (2006). Correlates of the Rosenberg self-esteem scale method effects. *Structural Equation Modeling*, *13*, 99–117. doi:10.1207/s15328007sem1301_5

Rauch, W. A., Schweizer, K., & Moosbrugger, H. (2007). Method effects due to social desirability as a parsimonious explanation of the deviation from unidimensionality in LOT-R scores. *Personality and Individual Differences*, *42*, 1597–1607. doi:10.1016/j.paid.2006.10.035

Reise, S. P. (2012). The rediscovery of bifactor measurement models. *Multivariate Behavioral Research*, *47*, 667–696. doi:10.1080/00273171.2012.715555

Rutkowski, D., Rutkowski, L., & von Davier, M. (2014). A brief introduction to modern international large-scale assessment. In L. Rutkowski, M. von Davier, & D. Rutkowski (Eds.), *Handbook of international large-scale assessment* (pp. 3–9). Boca Raton, FL: CRC Press.

Rutkowski, L., Gonzalez, E., von Davier, M., & Zhou, Y. (2014). Assessment design for international large-scale assessments. In L. Rutkowski, M. von Davier, & D. Rutkowski (Eds.), *Handbook of international large-scale assessment* (pp. 75–95). Boca Raton, FL: CRC Press.

Satorra, A., & Bentler, P. M. (2010). Ensuring positiveness of the scaled difference chi-square test statistic. *Psychometrika*, *75*, 243–248. doi:10.1007/s11336-009-9135-y

Schmitt, D. P., & Allik, J. (2005). Simultaneous administration of the Rosenberg Self-Esteem Scale in 53 nations: Exploring the universal and culture-specific features of global self-esteem. *Journal of Personality and Social Psychology*, *89*, 623–642. doi:10.1037/0022-3514.89.4.623

Schmitt, N., & Stults, D. M. (1985). Factors defined by negatively keyed items: The result of careless respondents? *Applied Psychological Measurement*, *9*, 367–373. doi:10.1177/014662168500900405

Shavelson, R. J., Hubner, J. J., & Stanton, G. C. (1976). Self-concept: Validation of construct interpretations. *Review of Educational Research*, *46*, 407–441. doi:10.3102/00346543046003407

Sherer, R., & Nilsen, T. (2016). The relations among school climate, instructional quality, and achievement motivation in mathematics. In T. Nilsen & J.-E. Gustafsson (Eds.), *Teacher quality, instructional quality and student outcomes* (pp. 51–80). Basel, Switzerland: IEA and Springer Open.

von Davier, M., Gonzalez, E., & Mislevy, R. (2009). What are plausible values and why are they useful?. In M. von Davier & D. Hastedt (Eds.), *IERI Monograph Series:Issues and methodologies in large-scale assessments* (pp. 9–36). Hamburg, Germany: IEA-ETS Research Institute.

Wang, W. C., Chen, H. F., & Jin, K. Y. (2015). Item response theory models for wording effects in mixed-format scales. *Educational and Psychological Measurement*, *75*, 157–178. doi:10.1177/0013164414528209

Wang, Y., Kim, E. S., Dedrick, R. F., Ferron, J. M., & Tan, T. (2018). A multilevel bifactor approach to construct validation of mixed-format scales. *Educational and Psychological Measurement*, *78*, 253–271. doi:10.1177/0013164417690858

Weems, G. H., Onwuegbuzie, A. J., Schreiber, J. B., & Eggers, S. J. (2003). Characteristics of respondents who respond differently to positively and negatively worded items on rating scales. *Assessment & Evaluation in Higher Education*, *28*, 587–606. doi:10.1080/0260293032000130234

Yang, Y., Chen, Y. H., Lo, W. J., & Turner, J. E. (2012). Cross-cultural evaluation of item wording effects on an attitudinal scale. *Journal of Psychoeducational Assessment*, *30*, 509–519. doi:10.1177/0734282911435461

Appendix

The 18 items arranged by section in the Student Background Questionnaire

1. How much do you agree with these statements about learning mathematics?
   a) I enjoy learning mathematics
   b) I wish I did not have to study mathematics *
   c) Mathematics is boring *
   d) I learn many interesting things in mathematics
   e) I like mathematics
   f) It is important to do well in mathematics
2. How much do you agree with these statements about your mathematics lessons?
   a) I know what my teacher expects me to do
   b) I think of things not related to the lesson *
   c) My teacher is easy to understand
   d) I am interested in what my teacher says
   e) My teacher gives me interesting things to do
3. How much do you agree with these statements about mathematics?
   a) I usually do well in mathematics
   b) Mathematics is harder for me than for many of my classmates *
   c) I am just not good at mathematics *
   d) I learn things quickly in mathematics
   e) I am good at working out difficult mathematics problems
   f) My teacher tells me I am good at mathematics
   g) Mathematics is harder for me than any other subject *

The response scale for all items was: 1 = Agree a lot, 2 = Agree a little, 3 = Disagree a little, 4 = Disagree a lot. * Negatively keyed items.