ISSN: 1309 – 6575
Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi
Journal of Measurement and Evaluation in Education and Psychology
Research Article; 2024; 15(2); 94-119

The eTIMSS and TIMSS Measurement Invariance Study: Multigroup Factor Analyses and Differential Item Functioning Analyses with the 2019 Cycle

Murat Yalçınkaya* Hakan Atılgan** Selim Daşcıoğlu*** Burak Aydın****

Abstract

In this study, measurement invariance and differential item functioning (DIF) studies of the TIMSS 2019 4th and 8th-grade mathematics and science achievement tests were conducted for the country groups participating in both TIMSS and eTIMSS. The study sample consisted of 9560 responders of the first booklet of the 2019 cycle. Multiple Group Confirmatory Factor Analysis (MGCFA) was utilized to test measurement invariance, and Mantel-Haenszel (MH), Logistic Regression (LR), and SIBTEST were used for the DIF analyses. The measurement invariance results indicated strict invariance between groups for all tests which included 111 items in total. In the DIF analyses, for the 4th and 8th-grade mathematics tests, only three items showed moderate DIF with MH, and four items showed DIF with SIBTEST. For the 4th-grade science test, one item showed moderate DIF with both MH and SIBTEST. However, in the 8th-grade science test, no items showed DIF with MH and LR methods, while four items showed moderate DIF with SIBTEST. Overall, MH and SIBTEST techniques were in agreement, whereas LR method produced inconsistent results and showed disagreement with these two methods. The results of the measurement invariance analysis and the LR method were consistent and indicated equivalency of TIMSS and e-TIMSS scores.

Keywords: *Multiple Group Confirmatory Factor Analysis, Differential Item Functioning, DIF, TIMSS, Computer-Based Assessments, Paper-Pencil Assessments*

Introduction

In recent years the widespread use of technology in education and the measurement of psychometric properties have become more prevalent. The 1970s marked the first decade when tests started to be used in a computer-based environment (Drasgow, 2002). The widespread use of computers in homes and classrooms has played a significant role in improving the quality of tests and enabling the use of measurement tools in different forms. Before tests were transferred to electronic platforms, ensuring equivalence with traditional paper-pencil applications posed a significant problem. In the literature, there are numerous studies comparing computer-based systems with paper-pencil tests (Mills, Potenza, Fremer, Ward, 2002; Russel, Goldberg, O'Connor, 2003; Anakwe, 2008; Ergün, 2002; İlci, 2004; Maguire, Smith, Brallier, & Palm, 2009). However, it is observed that no such studies were conducted concerning the computer-based tests implemented in the Trends in International Mathematics and Science Study (TIMSS) 2019. During the TIMSS 2019 administration, approximately half of the participating countries chose to switch to eTIMSS, while the other half preferred paper-pencil-based administration (Mullis et al., 2020).

* MA., Ege University, Institute of Education Sciences, Measurement And Evaluation In Education, , Faculty of Education, İzmir-Türkiye, muratyalcinkaya35@gmail.com, ORCID ID: 0000-0001-8564-3096
** Prof. Dr., Ege University, Faculty of Education, İzmir-Türkiye, hakan.atilgan@ege.edu.tr, ORCID ID: 0000-0002-5562-3446
*** MA., Ege University, Institute of Education Sciences, Measurement And Evaluation In Education, , Faculty of Education, İzmir- Türkiye, selimdascioglu@gmail.com, ORCID ID: 0000-0001-6820-4585
**** Assoc. Prof., Ege University, Faculty of Education, İzmir- Türkiye, burak.aydin@ege.edu.tr, Researcher, Leuphana University, Lüneburg- Germany, burak.aydin@leuphana.de, ORCID ID:0000-0003-4462-1784

To cite this article:
Yalçınkaya, M., Atılgan, H., Daşçıoğlu S., & Aydın B. (2024). The eTIMSS and TIMSS measurement invariance study: Multigroup factor analyses and differential item functioning analyses with the 2019 cycle. *Journal of Measurement and Evaluation in Education and Psychology, 15*(2), 94-119. https://doi.org/10.21031/epod.1426739

Received: 27.01.2024
Accepted: 30.05.2024

Therefore, conducting studies that demonstrate whether computer-based and paper-pencil-based tests can be used interchangeably, and their measurement invariance and differential item functioning (DIF) is essential under these conditions (Gündoğmuş, 2017).

In general, validity, which forms the fundamental principle of this study, refers to the extent to which a measurement tool accurately measures the characteristic it intends to measure without confounding it with other attributes (Atılgan, Kan, & Aydın, 2017). It does not seem possible to refer to a more effective concept than validity in this sense (Rogers, 1995). In order to provide evidence for the construct validity of a measurement tool, studies on measurement invariance have gained prominence in the academic field. Measurement invariance is simply defined as evaluating the equality of measurement results for different groups (Moraes & Reichenheim, 2002). At the same time, measurement invariance stands out as a prerequisite in group comparisons (Meredith, 2006). Testing measurement invariance ensures that intergroup comparisons are meaningful. In cases where measurement invariance cannot be achieved, it is possible that one of the groups to be compared may have an advantage or disadvantage, leading to biased interpretations. Therefore, as in the present study, comparing countries and ranking them based on achievement scores increases the importance of measurement invariance analyses.

Furthermore, measurement invariance studies allow for interpreting data at the scale level between groups, and the determination of items showing DIF provides additional evidence for construct validity. Another positive aspect of DIF studies is that they contribute to identifying the reasons for the strengths and weaknesses of the compared groups (Klieme & Baumert, 2001). Thus, in-depth examinations at the item level in tests and subtests can provide insights into item bias and predict which group may have an advantage or disadvantage. Although different methods applied in DIF analysis generally yield similar results, they may not produce entirely consistent results due to their different matching criteria and cut-off values used for labeling items as DIF (Gök, Kelecioğlu, & Doğan, 2010). Therefore, considering all these factors, it is recommended that researchers use multiple methods in DIF analysis (Hambleton, 2006). In this study, three different DIF determination methods were utilized. While methods based on Item Response Theory (IRT) include separate structures for categorical items, this study will use MH, LR, and SIBTEST methods based fundamentally on CTT for dichotomous items. During the process of determining DIF, one group with equal ability level to the test-taking group is referred to as the reference group, while the other is referred to as the focal group (Holland & Wainer, 1993.

Purpose and Significance of the Research

This study aims to analyze and interpret the findings regarding measurement invariance and DIF between paper-pencil tests and computer-based tests administered in TIMSS 2019. For this purpose, both scale-level Confirmatory Factor Analysis (CFA) for measurement invariance and item-level DIF analyses will be conducted for country groups participating in both paper-pencil and computer-based administrations. Additionally, it is believed that the data collected will provide insights for future similar test administrations and scientific studies.

In investigating the measurement invariance between computer-based and paper-pencil tests using the data obtained from the student achievement tests of TIMSS 2019, comparing the results from models without establishing measurement invariance would not be meaningful. It is essential to determine whether the items in the computer-based version provide advantages or disadvantages to test-takers compared to the items in the paper-pencil test.

TIMSS results, being one of the leading indicators in determining country's education policies, have been applied in our country in previous years using paper-pencil tests and in the latest administration using computer-based tests. Other countries are also gradually transitioning. Therefore, the purpose of this study is to evaluate the paper-pencil administration and computer-based administration in terms of measurement invariance and to identify whether DIF exists at the item level. This will contribute to the discussion of the sustainability and feasibility of the transition to computer-based administration by examining its positive and negative aspects.

Methods

The International Association for the Evaluation of Educational Achievement (IEA) conducts TIMSS every four years. In the TIMSS 2019 administration, 580,000 students from 64 countries participated, with the inclusion of seven more countries compared to TIMSS 2015. Among these countries, 32 opted for computer-based (eTIMSS) administration, while the other 32 preferred paper-pencil-based administration see Table 1.

Table 1

Countries Participating in TIMSS 2019 Implementation

Germany *	Philippines	Japan	Sweetcorn
USA*	Finland*	Canada*	Norway*
Albania	France*	Montenegro	Pakistan
Australia	South Africa	Qatar*	Poland
Austria*	South Cyprus	Kazakhistan	Portugal*
Azerbaijan	Georgia*	South Korea*	Romania
Bahrain	Croatia*	Kosovo	Russia*
Belgium (Flemish Region)	Holland*	Kuwait	Serbia
UAE*	Hong Kong*	North Ireland	Singapore*
Bosnia and Herzegovina	England*	North Macedonia	Slovakia*
Bulgaria	Iranian	Latvia	Saudi Arabia
Czech Republic*	Ireland	Lithuania*	Chile*
Taiwan*	Spain*	Lebanon	Türkiye*
Denmark*	Israel*	Hungary*	Oman
Armenia	Sweden*	Malaysia*	Jordan
Morocco	Italy*	Malta*	New Zeland

**Countries participating in eTIMSS (MEB,2020)*

In studies involving 4th-grade students, certain countries (Albania, Bosnia and Herzegovina, Kosovo, Kuwait, Montenegro, Morocco, North Macedonia, Pakistan, Philippines, Saudi Arabia, South Africa) have preferred to use "Less Difficult Mathematics" test versions, and therefore, they were not included in this study (Mullis et al., 2020).

As a result, in this study, 29 countries participated in the paper-pencil-based administration, and 30 countries participated in the computer-based administration for the 4th-grade mathematics test. Similarly, the countries Jordan, Romania, Israel, Malaysia, Egypt did not participate in the 4th and 8th-grade mathematics and science assessments. For the 8th-grade mathematics and science tests, 17 countries participated in the paper-pencil-based administration, while 22 countries opted for computer-based administration (MEB, 2020). In the studies, only one randomly selected test booklet was examined for all grade levels and tests (Table 2). The distribution frequency of this booklet among the students was similar or very close to the frequencies observed in all other booklets (7.2%).

Table 2

Booklet Usage Rates for TIMSS 2019 Mathematics 4th-grade Test

Booklets	Frequency	Percentage	Current Percentage	Additive Percentage
Booklet 1	9560	7.2	7.2	7.2
Booklet 2	9480	7.1	7.1	14.3
Booklet 3	9505	7.1	7.1	21.4
Booklet 4	9517	7.1	7.1	28.5
Booklet 5	9543	7.2	7.2	35.7
Booklet 6	9521	7.1	7.1	42.8
Booklet 7	9586	7.2	7.2	50.0
Booklet 8	9509	7.1	7.1	57.2
Booklet 9	9506	7.1	7.1	64.3
Booklet 10	9498	7.1	7.1	71.4
Booklet 11	9517	7.1	7.1	78.6
Booklet 12	9543	7.2	7.2	85.7
Booklet 13	9514	7.1	7.1	92.9
Booklet 14	9529	7.1	7.1	100.0
Total	133328	100.0	100.0	

Derived items (Annex 13) were scored by taking the integrated answer part (TIMSS, 2019).

The integrated response part of the derived items (Appendix 13) was scored in TIMSS 2019. The extensions of the derived items were not considered, and the responses to the binary items were coded as "1" if all sub-items were answered correctly, and "0" if not. Therefore, the number of items in the 8th-grade science test, for example, was 44 for the derived items, including their sub-items, but after arranging the dependent items, 31 items were included in the analysis. The table resulting from the item matching process and the corresponding number of students are presented in Table 3.

Table 3
TIMSS 2019 Number of Items and Students

GROUP	NUMBER OF ITEMS	NUMBER OF STUDENTS
	4th GRADE	
TIMSS MATHEMATICS	24	5373
eTIMSS MATHEMATICS	24	8917
TIMSS SCIENCE	25	9284
eTIMSS SCIENCE	25	9264
	8th GRADE	
TIMSS MATHEMATICS	31	7326
eTIMSS MATHEMATICS	31	7270
TIMSS SCIENCE	31	7224
eTIMSS SCIENCE	31	7930

In all booklets, care was taken to ensure an equal distribution of item types and numbers, and to distribute the booklets to as close to an equal number of students as possible. The data for the 4th and 8th grades included in the study were downloaded and organized from the official website of the TIMSS&PIRLS International Study Center.

Analysis of Data

The evaluation of the TIMSS 2019 mathematics and science test items involved completing studies on missing data, followed by an examination of outliers. Among the main methods chosen by researchers for dealing with missing data are data deletion, estimation of missing data using imputation methods, and approximate value assignment to missing data (Büyüköztürk, Çokluk, & Şekercioğlu, 2014). Regarding the present study, due to the size of the data set and the missing data rate being less than 5% and considered random, data deletion method was selected as the most appropriate approach (Tabachnick & Fidell, 2007). During the examination of missing data, responses to items labeled as "9" in the data set, indicating that the student left the answer blank because they did not know the correct response, were coded as "0". Responses coded as "6", representing patterns where the student did not encounter the item due to technical issues or insufficient time during the exam, were removed from the data set.

Subsequently, CFA and Multiple Group Confirmatory Factor Analysis (MGCFA) were conducted. Given that the research data were categorical, the assumption of normality was not tested. Furthermore, the multicollinearity assumption was examined by assessing the tetrachoric correlation between items, and it was observed that all correlations were below .90. Additionally, Variance Inflation Factors (VIF), Tolerance Levels, and Condition Indices (CI) were examined, and it was found that CI values were below 30, VIF values were below 10, or tolerance values were above .10, indicating the absence of multicollinearity issues (Kline, 2016; Hair, Anderson, Tatham, & Black, 1998; Mertler & Vannatta, 2005; Tabachnick & Fidell, 2007). The VIF and tolerance values for each subscale are provided in Appendix 1 through Appendix 4; tetrachoric correlation coefficients are provided in Appendix 5 through Appendix 8.

The Weighted Least Squares Mean and Variance (WLSMV) method was employed as the parameter estimation method in CFA and MGCFA. It is noted in the literature that the asymptotically distribution-free estimator is used in conjunction with ordinal categorical data. WLSMV, utilized in analyses with ordinal categorical data, produces better results based on polychoric correlations, accuracy of parameter estimates, and estimated standard errors. In other words, polychoric correlations are reported to provide the best estimates of model parameters (Joreskog & Sorbom, 1981). WLSMV can be considered as an alternative method for non-normally distributed, highly skewed, or platykurtic ordinal data (Muthén, 1993). In this study, the established models were confirmed through Confirmatory Factor Analysis for the entire data set, obtaining evidence for construct validity. The learning domains specified by TIMSS were used as the sub-dimensions in the analysis (Mullis et al., 2020). CFA was conducted using the M*plus* 7.4 program with the WLSMV estimation method (Jöreskog & Sörbom, 2006).

CFA analyses were conducted to confirm the subscales specified by TIMSS. Additionally, the path diagrams of the CFA analyses performed using the M*plus* 7 program are provided in Appendix 9 through 12. Table 4 illustrates how model-data fit is assessed based on the fit indices obtained from the CFA results based on χ^2/df (Kline, 2016), CFI (Bentler, 1980), SRMR and RMSEA (Browne & Cudeck, 1993).

Table 4

Cut off values to be used in the evaluation of CFA fit indices

Fit Index	Good Fit	Acceptable Fit
χ2	p>.05	p>.05
χ2/df	0≤ χ2/df≤2	2≤ χ2/df≤8
CFI	.97≤CFI≤1.00	.95≤CFI<.97
TLI	.95≤TLI≤1.00	.90≤TLI<.95
RMSEA	0≤RMSEA≤.05	.05<RMSEA≤.08

For the 4th-grade science test, the three-factor model (life sciences, physical sciences, and earth sciences) demonstrated an acceptable fit (χ^2/df =5.955, CFI=.990, TLI=.989, and RMSEA=.016). Similarly, for the 8th-grade science test, the four-factor model (physics, chemistry, biology, and earth sciences) showed an acceptable fit (χ^2/df = 8.795, CFI=.981, TLI=.979, and RMSEA=.023). For the 4th-grade mathematics test, the three-factor model (numbers, data, measurement, and geometry) displayed a considerably lower χ2/df (37.749) statistic, indicating an acceptable fit, while the CFI (.953) indicated a good fit, and the TLI (.947) and RMSEA (.051) showed an acceptable fit. For the 8th-grade mathematics test, the four-factor model (numbers, algebra, geometry, data, and probability) exhibited an acceptable fit with a χ2/df (13.938) statistic below the acceptable limit, and a good fit based on the CFI (.981), TLI (.979), and RMSEA (.030) statistics.

MGCFA based on structural equation modeling was used to assess measurement invariance. In the literature, there are different views among researchers regarding the number of steps and the nature of operations involved in evaluating measurement invariance. In this study, a 4-step hierarchical model, encompassing configural, metric, scalar, and strict invariance, will be employed (Steenkamp & Baumgartner, 1998; Wu, Li, & Zumbo, 2007; Byrne, 2008; Meredith & Teresi, 2006).

Table 5

Parameters Used in Measurement Invariance Analysis

Invariance Model	Fixed Parameters	Tested Parameters
Configural Invariance	-	Item/Factor groups
Metric Invariance	Factor variances and covariances	Factor loadings
Scalar Invariance	+ Factor and observed variable means	Intercepts or thresholds
Strict Invariance	+ Observed Variances and Covariances	Residual variances

As shown in Table 5, in each step, one additional parameter is added and fixed at each stage to the parameters kept constant (Gregorich, 2006). Moreover, with each step, one more parameter is added and fixed in the tested parameters. In measurement invariance studies categorical variables can be forced to fit these four steps (e.g., Li, Gooden & Toland, 2016) or the number of steps can be reduced based on the number of categories (e.g., Bagdu Soyler, Aydın & Atılgan, 2021; titina et al., 2020; Raykov et al., 2018). In our analyses we preferred to use the four-step approach given that it is more common with the TIMSS analyses.

Fit Indices

MGCFA is based on Structural Equation Modeling (SEM) and allows simultaneous testing of the model in multiple groups (Tabachnick & Fidell, 2007). In the first stage of the study, which is within the scope of the MGCFA technique, CFI, TLI, and RMSEA are used to evaluate the model-data fit. In each step of the invariance testing, differences between CFI and TLI are used to provide information about the relationship between latent scores and observed scores. It is noted that CFI, TLI, and RMSEA fit indices should fall within the desired range, with $.01 \geq \Delta CFI \geq -.01$ and $.01 \geq \Delta TLI \geq -.01$ for each step of the MGCFA data sets (Cheung & Rensvold, 2002; Vandenberg & Lance, 2000). However, χ^2 statistic, being influenced by sample size, is considered in large samples like this study by taking into account other fit indices (Brown, 2006; Büyüköztürk, 2010; Tabachnick & Fidel, 2007). In the literature, it has been stated that the χ^2 difference used for measurement invariance analyses should not be used alone (Wu, Li, & Zumbo, 2007), and other findings have been reported (Cheung & Rensvold, 2002; Vandenberg & Lance, 2000). Further for the estimators appropriate for categorical data regular χ^2 tests might not be appropriate adjustments might be needed, in M*plus* this is handled with the DIFFTEST command, and its technical details are briefly studied by Kite, Johnson and Xing (2018).

After MGCFA, the derived test items were evaluated for DIF using the MH, LR, and SIBTEST procedures. While test-level CFA can be used to evaluate measurement invariance, DIF can be used for item and subtest level analyses, as observed in the literature (Cheung & Rensvold, 2002; Raju, Laffitte, & Byrne, 2002). DIF is defined as the differentiation of the probability of correctly answering a test item among different subgroups of individuals with equal abilities (Camilli & Shepard, 1994; Zumbo, 1999). DIF determination techniques based on the Classical Test Theory (CTT) are index-dependent sampling techniques (Camilli & Shepard, 1994). In the CTT-based methods, separate procedures are used for polytomous and dichotomous items. In this study, the MH, LR, and SIBTEST methods will be used for comparing the results of DIF obtained for dichotomous tests. Unlike the MGCFA, the DIF analyses were conducted separately for test dimensions. Even though it is possible to conduct multidimensional DIF (e.g., Bulut & Suh, 2017) our attempts to utilize *mirt* (Chalmers, 2012) package was unsuccessful probably due to the large sample size and relatively complex factor structure.

Mantel-Haenszel (MH)

William Haenszel and Nathan Mantel developed the DIF determination method based on the chi-square statistic in the 1950s. This technique is a method used in tests containing dichotomously scored items. The odds ratio (α) calculates the degree of performance difference between the reference and focal groups, in other words, the ratio of individuals answering correctly and incorrectly in each ability level for both reference and focal groups, taking into account the total number of respondents (Mertler and Vannatta, 2005; Agresti, 1984). To express MH more effectively, the natural logarithm is obtained, and ΔMH (delta coefficient) is determined through a logarithmic transformation. When determining DIF with the MH technique, the following interpretations are made: if $\Delta MH=0$ or $\alpha=1$, there is no DIF in the item; if $\Delta MH<0$ or $\alpha>1$, there is DIF in favor of the reference group; if $\Delta MH>0$ or $\alpha<1$, there is DIF in favor of the focal group (Camilli and Shepard, 1994; Nandakumar, 1993). Additionally, if $|\Delta MH|<1$, DIF in the item is negligible (Level A); if $1\leq|\Delta MH|<1.5$, DIF in the item is moderate (Level B); if $|\Delta MH|\geq1.5$, DIF in the item is significant (Level C) (Dorans & Holland, 1993; Zieky, 1993).

Logistic Regression (LR)

LR is a regression model used when the dependent variable is binary (1-0). In other words, LR is used when it is expected that the dependent variable will exhibit responses in a non-linear relationship with one or more independent variables (Tabachnick & Fidell, 2007). LR is a non-parametric method.

The standardized regression coefficients are considered LR effect sizes (Gierl, Jodoin & Ackerman, 2000). The standardized regression coefficients (R^2) provide the degree of DIF (Differential Item Functioning), and they are determined in three levels. If $R^2<.035$ for the difference between Model 1 and Model 3, there is no DIF or it is negligible. If $.035\leq R^2<.070$, there is moderate-level (B) DIF. If $R^2\geq .070$, there is significant-level (C) DIF. For an item to be classified as having DIF (B or C level), the chi-square value must be statistically significant at the .05 level or less, and the R^2 value must be at least .035 (Zumbo, 1999). Additionally, for items with identified DIF, the presence of non-uniform DIF is examined by checking if the difference between the R^2 values of Model 2 and Model 3 is greater than .035. If it is greater, non-uniform DIF can be considered.

SIBTEST

The SIBTEST method can statistically demonstrate whether one or more items exhibit DIF (Shealy & Stout, 1993). SIBTEST is used in DIF analyses for dichotomous data and can estimate the degree of DIF exhibited by an item. As a non-parametric method based on the IRT, SIBTEST provides a more precise synchronization of the focal and reference groups (Osterlind & Everson, 2009).

The β index primarily represents the effect size. A positive index value indicates DIF in favor of the reference group, while a negative value indicates DIF in favor of the focal group. If $\beta<|.059|$, the item is considered to have negligible DIF (Level A), if $|.059|\leq\beta<|.088|$, it has moderate DIF (Level B), and if $\beta\geq|.088|$, it has substantial DIF (Level C) (Rousses & Stout, 1996).

Results

The first stage of measurement invariance, known as configural invariance, examines whether the structure is comparable across groups. When looking at the fit indices for the 4th grade mathematics test, as shown in Table 6, all values, including RMSEA (.051), CFI (.952), and TLI (.947), fall within an acceptable range of fit. The χ^2/sd (19.720) value falls outside the specified intervals for the likelihood, as a result of biased results in large samples (Kline, 2016). Hence, as expected, all χ^2 difference tests reported in the Table 6, including the one for the 4th grade mathematics are significant. However, all other values are within the permitted minimum level intervals, confirming that the structure is similar across all groups, and the model demonstrates invariance at all stages between the TIMSS 4th-grade mathematics test using paper and pencil and computer-based methods.

Table 6

Measurement Invariance Results by TIMSS 2019 Tests Participation Pattern (eTIMSS/TIMSS)

Test	Invariance Type	χ^2/sd	$\Delta\chi^2$	RMSEA	CFI	TLI	ΔCFI	ΔTLI
4th-grade Science	Configural	3.993		0.018	0.987	0.985		
	Weak	5.636	438.32*	0.022	0.979	0.977	0.008	0.008
	Strong	6.366	621.96*	0.024	0.974	0.974	0.005	0.003
	Strict	5.269	459.12*	0.021	0.980	0.979	-0.006	-0.005
4th-grade Mathematics	Configural	19.720		0.051	0.952	0.947		
	Weak	15.716	264.95*	0.045	0.961	0.959	-0.009	-0.012
	Strong	16.177	692.81*	0.046	0.958	0.957	0.003	0.002
	Strict	19.677	312.34*	0.051	0.951	0.947	0.007	0.010
8th-grade Science	Configural	5.594		0.025	0.975	0.973		
	Weak	5.362	284.62*	0.024	0.976	0.975	-0.001	-0.002
	Strong	5.701	543.88*	0.025	0.973	0.973	0.003	0.002
	Strict	6.146	300.85*	0.026	0.972	0.970	0.001	0.003
8th-grade Mathematics	Configural	7.968		0.031	0.978	0.976		
	Weak	5.955	271.77*	0.026	0.984	0.983	-0.006	-0.007
	Strong	7.122	1235.06*	0.029	0.979	0.979	0.005	0.004
	Strict	8.920	344.67*	0.033	0.974	0.973	0.005	0.006

Note: * p<.05

Similarly, when examining the 8th-grade mathematics test, during the stage of configural invariance, all values, including RMSEA (.031), CFI (.978), and TLI (.976), fall within the good fit range. Except for the χ^2 tests, it can be observed that the structure is similar across groups, and the model demonstrates invariance at all stages based on the participation method for the 8th grade mathematics test.

Except for the χ^2 tests, it is observed that strict invariance is achieved in the 4th and 8th grade science test. As a result, when examining Table 6 which show the goodness-of-fit indices as well as the differences between ΔCFI and ΔTLI values considered after structural invariance at all stages of measurement invariance for both 4th and 8th-grade mathematics and science tests, it is evident that the differences are within acceptable limits, indicating the achievement of strict invariance stages.

In the context of the TIMSS and eTIMSS samples, combined data sets were analyzed using MH, SIBTEST, and LR techniques to identify items exhibiting DIF based on the participation format. α, β, and ΔR^2 coefficients were computed, and the directions and magnitudes of these coefficients were taken into account to determine the level of DIF for matched items between paper-pencil and computer-based formats. As mentioned earlier, DIF analyzes were performed separately for each sub-dimension of the tests.

Table 7

DIF Status of 4th Grade Mathematics Test Items in Booklet No. 1 in TIMSS 2019 Implementation by Country Groups Participating in TIMSS/eTIMSS

Sub-dimension	Item	MH					LR				SIBTEST			
		α	χ2	p	ΔMH	DIF Level, Direction	Δχ2	p	ΔR^2	DIF Level, Direction	β	χ2	p	DIF Level, Direction
Number	M1	.776	41.946	<.001	.595	A	101.280	<.001	<.035	A	-.054	40.733	<.001	A
	M2	1.224	18.524	<.001	-.475	A	36.838	<.001	<.035	A	.034	22.425	<.001	A
	M3	.663	64.900	<.001	.967	A	81.177	<.001	<.035	A	-.056	66.308	<.001	A
	M4	1.280	26.918	<.001	-.581	A	35.640	<.001	<.035	A	.039	27.197	<.001	A
	M5	1.625	94.205	<.001	-1.141	B-	103.062	<.001	<.035	A	.071	94.535	<.001	B-
	M6	.846	13.008	<.001	.392	A	15.188	.001	<.035	A	-.030	15.366	<.001	A
	M13	1.118	4.675	.031	-.263	A	7.710	.021	<.035	A	.016	5.494	.019	A
	M14	1.065	2.032	.154	-.148	A	3.958	.138	<.035	A	.012	2.130	.144	A
	M15	.949	.997	.318	.123	A	65.859	<.001	<.035	A	-.012	3.351	.067	A
	M16	.921	2.662	.103	.193	A	3.661	.160	<.035	A	-.007	.839	.360	A
	M17	.958	.504	.478	.102	A	24.160	<.001	<.035	A	-.011	3.060	.080	A
Measurement and Geometry	M7	1.598	126.864	<.001	-1.101	B-	129.704	<.001	<.035	A	.098	132.030	<.001	C-
	M8	.685	69.199	<.001	.890	A	70.655	<.001	<.035	A	-.065	77.081	<.001	B+
	M9	1.456	64.410	<.001	-.882	A	73.766	<.001	<.035	A	.059	62.752	<.001	B-
	M10	.782	32.395	<.001	.577	A	34.202	<.001	<.035	A	-.041	26.267	<.001	A
	M18	1.285	36.013	<.001	-.590	A	47.811	<.001	<.035	A	.051	37.031	<.001	A
	M19	.902	5.671	.017	.242	A	7.400	.025	<.035	A	-.022	6.672	.010	A
	M20	1.070	1.585	.208	-.158	A	6.253	.044	<.035	A	.006	.741	.389	A
	M21	.588	123.179	<.001	1.249	B+	131.190	<.001	<.035	A	-.086	140.335	<.001	B+
Data	M11	1.178	11.102	.001	-.386	A	13.130	.001	<.035	A	.031	10.908	.001	A
	M12	1.082	1.631	.202	-.185	A	1.621	.445	<.035	A	.010	1.541	.215	A
	M22	1.200	10.429	.001	-.428	A	11.626	.003	<.035	A	.023	8.032	.005	A
	M23	.712	44.072	<.001	.799	A	44.553	<.001	<.035	A	-.057	41.349	<.001	A
	M24	.958	.661	.416	.102	A	7.639	.022	<.035	A	-.006	0.452	.501	A

+/-: DIF favors focal/reference group.

Based on the MH results, out of the 24 items in the 4th grade mathematics test of TIMSS 2019, 21 exhibited negligible levels of DIF (Level A), while 3 items showed moderate DIF (level B). Item 21 favors students taking the paper-pencil version, whereas item 5 and 7 favor students taking the computer-based version (see Table 7). On the other hand, the LR results indicated that all items in the 4th grade mathematics test exhibited negligible levels of DIF (Level A). As for the SIBTEST results, 19 items were found to have negligible levels of DIF (level A), 4 items showed DIF at Level B, and 1 item showed DIF at Level C (see Table 7). Based on the SIBTEST analyses, items 8 and 21 favor students taking the paper-pencil version, items 5, 7 and 9 favor students taking the computer-based version.

Table 8

DIF Status of 8th Grade Mathematics Test Items in Booklet No. 1 in TIMSS 2019 Implementation by Country Groups Participating in TIMSS/eTIMSS

		MH					LR				SIBTEST			
Subest	Item	α	χ2	p	ΔMH	DIF Level, Direction	Δχ2	p	ΔR^2	DIF Level, Direction	β	χ2	p	DIF Level, Direction
Number	M1	.982	.167	.683	.043	A	2.574	.276	<.035	A	.011	1.907	.167	A
	M2	1.762	148.537	<.001	-1.331	B-	145.679	<.001	<.035	A	.098	152.008	<.001	C-
	M3	1.066	1.190	.275	-.151	A	13.552	.001	<.035	A	-.010	2.308	.129	A
	M4	1.078	3.010	.083	-.177	A	12.194	.002	<.035	A	.027	10.701	.001	A
	M5	.489	275.735	<.001	1.679	C+	309.959	<.001	<.035	A	-.118	203.741	<.001	C+
	M17	1.070	1.386	.239	-.159	A	8.871	.012	<.035	A	-.003	0.216	.643	A
	M18	1.196	17.494	<.001	-.420	A	17.702	<.001	<.035	A	.044	28.392	<.001	A
	M19	.739	25.779	<.001	.711	A	28.878	<.001	<.035	A	-.041	47.593	<.001	A
	M20	1.094	4.510	.034	-.211	A	17.759	<.001	<.035	A	.038	20.277	<.001	A
Algebra	M6	1.055	1.556	.212	-.125	A	5.014	.082	<.035	A	.010	1.490	.222	A
	M7	.860	11.903	.001	.354	A	28.760	<.001	<.035	A	-.028	12.584	<.001	A
	M8	.693	81.117	<.001	.863	A	89.486	<.001	<.035	A	-.063	56.990	<.001	B+
	M9	.432	172.488	<.001	1.974	C+	194.924	<.001	<.035	A	-.084	225.012	<.001	B+
	M10	.896	6.384	.012	.258	A	9.926	.007	<.035	A	-.012	2.340	.126	A
	M21	1.531	98.762	<.001	-1.001	B-	102.884	<.001	<.035	A	.078	94.783	<.001	B-
	M22	.895	6.846	.009	.261	A	7.176	.028	<.035	A	-.019	5.617	.018	A
	M23	1.220	16.806	<.001	-.467	A	17.532	<.001	<.035	A	.021	8.533	.004	A
	M24	1.434	40.271	<.001	-.846	A	40.972	<.001	<.035	A	.027	18.143	<.001	A
	M25	1.341	52.384	<.001	-.689	A	61.148	<.001	<.035	A	.069	69.199	<.001	B-
Geometry	M11	.592	148.233	<.001	1.232	B+	167.335	<.001	<.035	A	-.069	62.568	<.001	B+
	M12	1.379	53.325	<.001	-.755	A	47.148	<.001	<.035	A	.068	59.885	<.001	B-
	M13	.961	.764	.382	.094	A	6.522	.038	<.035	A	-.011	1.444	.230	A
	M26	.705	52.960	<.001	.822	A	70.726	<.001	<.035	A	-.080	82.549	<.001	B+
	M27	1.543	108.160	<.001	-1.019	B-	101.924	<.001	<.035	A	.104	127.927	<.001	C-
	M28	1.127	5.435	.020	-.281	A	6.991	.030	<.035	A	-.005	0.321	.571	A
Data and Probability	M14	1.080	2.497	.114	-.180	A	40.866	<.001	<.035	A	.030	11.378	.001	A
	M15	.819	17.918	<.001	.470	A	49.633	<.001	<.035	A	-.021	6.115	.013	A
	M16	.907	4.839	.028	.231	A	8.239	.016	<.035	A	-.005	0.265	.607	A
	M29	1.765	114.973	<.001	-1.335	B-	117.194	<.001	<.035	A	.065	65.862	<.001	B-
	M30	.978	.241	.623	.053	A	13.436	.001	<.035	A	.020	4.772	.029	A
	M31	.714	28.362	<.001	.792	A	32.810	<.001	<.035	A	-.042	43.260	<.001	A

+/-: DIF favors focal/reference group.

In the TIMSS 2019 8th grade mathematics test, MH results shows that 5 items have DIF at Level B, and 2 items have DIF at Level C, as reported in Table 8. Item 5, 9, and 11 favor students taking the paper-pencil version, while item 2, 22, 27 and 29 favor students taking the computer-based version. However, based on the LR results, all items showed negligible levels of DIF (Level A). As for the SIBTEST results, 8 items were found to have DIF at Level B, and 3 items exhibited DIF at Level C. Similarly, item 5, 8, 9, 11 and 26 favored students taking the paper-pencil version, while item 2, 12, 21, 25, 27 and 29 favored students taking the computer-based version in terms of DIF.

Table 9

DIF Status of 4th Grade Science Subtest Items in Booklet No. 1 in TIMSS 2019 Implementation by Country Groups Participating in eTIMSS/TIMSS

Subtest	Item	MH					LR				SIBTEST			
		α	χ2	p	ΔMH	DIF Level, Direction	Δχ2	p	ΔR^2	DIF Level, Direction	β	χ2	p	DIF Level, Direction
Life	M1	1.389	73.972	<.001	-.771	A	88.636	<.001	<.035	A	.059	76.872	<.001	B-
	M2	1.032	.734	.392	-.073	A	21.921	<.001	<.035	A	.004	.273	.602	A
	M3	1.181	13.957	<.001	-.390	A	30.329	<.001	<.035	A	.013	5.102	.024	A
	M4	.637	77.031	<.001	1.060	B+	77.856	<.001	<.035	A	-.047	92.492	<.001	A
	M5	.719	90.511	<.001	.774	A	84.115	<.001	<.035	A	-.067	86.005	<.001	B+
	M6	.941	2.598	.107	.144	A	2.194	.334	<.035	A	-.008	1.395	.238	A
	M13	.667	134.066	<.001	.952	A	152.259	<.001	<.035	A	-.082	127.134	<.001	B+
	M14	1.059	2.172	.141	-.135	A	5.944	.051	<.035	A	.009	1.868	.172	A
	M15	1.261	42.470	<.001	-.545	A	55.523	<.001	<.035	A	.040	31.547	<.001	A
	M16	1.109	7.122	.008	-.242	A	33.526	<.001	<.035	A	-.002	.049	.824	A
	M17	1.067	3.646	.056	-.153	A	22.452	<.001	<.035	A	.036	24.023	<.001	A
	M18	1.217	28.456	<.001	-.462	A	52.723	<.001	<.035	A	.023	10.757	.001	A
Physycal	M7	.944	2.694	.101	.135	A	18.744	<.001	<.035	A	.001	0.009	.926	A
	M8	.993	.029	.866	.016	A	1.981	.371	<.035	A	.001	0.035	.851	A
	M9	1.257	36.680	<.001	-.538	A	42.914	<.001	<.035	A	.026	12.489	<.001	A
	M10	.926	4.479	.034	.181	A	8.146	.017	<.035	A	-.008	1.264	.261	A
	M19	.854	21.015	<.001	.370	A	29.607	<.001	<.035	A	-.024	9.595	.002	A
	M20	1.389	83.992	<.001	-.773	A	84.483	<.001	<.035	A	.066	79.369	<.001	B-
	M21	.840	23.399	<.001	.410	A	25.570	<.001	<.035	A	-.027	13.287	<.001	A
	M22	.956	1.372	.242	.105	A	2.434	.296	<.035	A	-.020	7.441	.006	A
Earth	M11	.941	2.770	.096	.143	A	4.793	.091	<.035	A	.006	.540	.463	A
	M12	1.516	111.321	<.001	-.978	A	117.528	<.001	<.035	A	.081	98.431	<.001	B-
	M23	.654	119.245	<.001	1.000	A	192.285	<.001	<.035	A	-.096	144.527	<.001	C+
	M24	.987	.082	.775	.030	A	3.174	.205	<.035	A	-.027	12.278	.001	A
	M25	1.094	6.205	.013	-.212	A	90.601	<.001	<.035	A	.047	32.750	<.001	A

+/-: DIF favors focal/reference group.

Based on the MH results reported in Table 9, in the TIMSS 2019 4th grade science test consisting of 25 items only 1 item exhibited DIF at Level B favors students taking the paper-pencil version, and no items showed DIF at Level C. Based on the LR results, all items showed negligible levels of DIF (Level A). For the SIBTEST results, 5 items exhibited DIF at Level B, indicating that 1 item showed DIF at this level.

Therefore, based on the SIBTEST results, item 5, 13 and 23 favored students taking the paper-pencil version, while item 1 and 12 favored students taking the computer-based version in terms of DIF.

Table 10

DIF Status of 8th-grade Science Subtest Items in Booklet No. 1 in TIMSS 2019 Implementation by Country Groups Participating in eTIMSS/TIMSS

Subtest	Item	MH					LR				SIBTEST			
		α	χ2	p	ΔMH	DIF Level, Direction	Δχ2	p	ΔR^2	DIF Level, Direction	β	χ2	p	DIF Level, Direction
Biology	M1	1.280	37.819	<.001	-.581	A	40.502	<.001	<.035	A	.048	39.996	<.001	A
	M2	.926	4.270	.039	.180	A	9.260	.010	<.035	A	.003	.164	.686	A
	M3	1.497	109.744	<.001	-.948	A	109.631	<.001	<.035	A	.084	113.852	<.001	B-
	M4	.896	5.957	.015	.257	A	10.021	.007	<.035	A	-.029	16.962	<.001	A
	M5	.847	18.905	<.001	.390	A	65.766	<.001	<.035	A	-.030	13.562	<.001	A
	M15	1.043	1.162	.281	-.099	A	1.832	.400	<.035	A	.014	2.950	.086	A
	M16	1.041	1.157	.282	-.094	A	9.258	.010	<.035	A	.024	8.795	.003	A
	M17	.898	6.551	.011	.253	A	7.557	.023	<.035	A	-.034	2.173	<.001	A
	M18	.761	53.260	<.001	.642	A	54.437	<.001	<.035	A	-.047	33.027	<.001	A
	M19	1.206	16.830	<.001	-.440	A	20.574	<.001	<.035	A	.017	6.203	.013	A
	M20	.793	19.784	<.001	.545	A	24.984	<.001	<.035	A	-.031	27.736	<.001	A
Chemistry	M6	1.066	2.176	.140	-.149	A	14.593	.001	<.035	A	.022	5.370	.021	A
	M21	.814	23.915	<.001	.484	A	35.221	<.001	<.035	A	-.043	19.908	<.001	A
	M22	1.006	.012	.914	-.014	A	.666	.717	<.035	A	.004	.255	.614	A
	M23	1.136	7.513	.006	-.300	A	7.936	.019	<.035	A	.018	4.708	.030	A
	M24	1.038	.701	.403	-.088	A	2.655	.265	<.035	A	.009	1.045	.307	A
Physics	M7	.995	.011	.918	.011	A	.427	.808	<.035	A	.010	1.499	.221	A
	M8	.733	61.356	<.001	.731	A	67.089	<.001	<.035	A	-.062	57.906	<.001	B+
	M9	.695	80.003	<.001	.855	A	90.701	<.001	<.035	A	-.073	83.041	<.001	B+
	M10	.906	6.808	.009	.232	A	9.634	.008	<.035	A	-.002	.034	.854	A
	M11	1.351	59.731	<.001	-.707	A	69.915	<.001	<.035	A	.072	78.148	<.001	B-
	M25	1.160	14.258	<.001	-.349	A	21.779	<.001	<.035	A	.029	13.055	<.001	A
	M26	1.187	14.025	<.001	-.403	A	18.518	<.001	<.035	A	.009	1.613	.204	A
	M27	.928	2.768	.096	.175	A	16.139	<.001	<.035	A	-.022	8.622	.003	A
	M28	1.323	43.694	<.001	-.657	A	44.481	<.001	<.035	A	.046	36.452	<.001	A
Earth	M12	.852	15.190	<.001	.376	A	19.419	<.001	<.035	A	-.013	2.103	.147	A
	M13	.932	2.176	.140	.165	A	4.815	.090	<.035	A	-.034	18.700	<.001	A
	M14	1.081	3.039	.081	-.182	A	11.566	.003	<.035	A	-.008	1.011	.315	A
	M29	1.067	2.688	.101	-.153	A	4.258	.119	<.035	A	.046	26.647	<.001	A
	M30	.997	.002	.966	.006	A	12.598	.002	<.035	A	-.019	5.042	.025	A
	M31	1.075	3.385	.066	-.169	A	2.336	.311	<.035	A	.035	14.760	<.001	A

+/-: DIF favors focal/reference group.

Based on the MH and LR results reported in Table 10, in the TIMSS 2019 8th grade science test consisting of 31 items, all items showed negligible levels of DIF (Level A). However, according to the SIBTEST results, 4 items exhibited DIF at Level B. Item 8 and 9 favored students taking the paper-pencil version, while items 3 and 11 favored students taking the computer-based version in terms of DIF, see Table 10.

Discussion

In this study, measurement invariance based on the participation format in paper-pencil TIMSS and computer-based eTIMSS mathematics and science achievement tests in TIMSS 2019 is examined, along with whether the items exhibit DIF. The stages of measurement invariance are tested hierarchically. Following the findings from the stages of measurement invariance, DIF analyses are conducted using three different approaches, namely MH, LR, and SIBTEST, to determine the items exhibiting DIF for mathematics and science subtests between paper-pencil and computer-based groups. These analyses also indicate whether DIF favors the focal or reference groups.

The results of the analyses indicate that in TIMSS 2019, at both 4th and 8th grade levels, the stages of measurement invariance, including configural, metric, scalar, and strict invariance, are established for all subtests in mathematics and science based on the ΔCFI and ΔTLI. But χ^2 difference tests indicated lack of invariance, as expected with large sample sizes. The variables in the mathematics and science achievement test models, including item and factor loadings, item intercepts, and error variances, are considered to be invariant across paper-pencil and computer-based groups for all subtests and grade levels, indicating measurement invariance. In other words, the observed differences between paper-pencil and computer-based groups for all subtests seem to stem from genuine ability differences between the groups. Consequently, it can be concluded that the computer-based eTIMSS and paper-pencil TIMSS assessments conducted for the first time in 2019 are comparable across all four subtests. This finding is considered to be particularly significant, and it is suggested that countries participating in the paper-pencil administration should expedite the transition to computer-based assessment procedures once they complete the necessary infrastructure work.

Most of the measurement invariance studies conducted for large-scale exams in the literature involve the hierarchical stages and results reached through MGCFA analyses for variables such as gender, school environment, and achievement vary and their outcomes differ (Arim & Ercikan, 2014; Gündoğmuş, 2017; Wruster, 2022). In line with this research, Wu, Li, and Zumbo (2007) present the results of binary comparisons of 21 countries selected for TIMSS 1999 mathematics and science tests. The results obtained for all tests included in our study are consistent with the conclusion of measurement invariance at the level of strong invariance. Ercikan and Koh (2009) find strong invariance in three out of eight test booklets for TIMSS 2003 cycle science and mathematics tests between Canada-England and France. In contrast, similar uniformity is not observed in the others. In this sense, it can be said that the results are consistent. Similarly, in Akyıldız's (2009) study, the MGCFA comparisons of 35 countries in the PIRLS 2001 achievement tests provide evidence of strong invariance, which is consistent with the results obtained for all tests included in this study. In Eriştiren's (2021) study, the measurement invariance achieved at all stages in the analyses conducted with binary categorical data for the Turkish language achievement test in the LGS 2018, inclusive of 3000 students, is in line with this study.

The MGCFA results at the scale level were also evaluated in terms of DIF at the item level. The results of the analyses conducted with three different methods for item-level analysis and MGCFA at the scale level were compared and evaluated in line with the examples in the literature. The items in the mathematics and science subtests at the 4th and 8th grade levels were analyzed using the MH, LR, and SIBTEST methods, depending on the mode of test administration (paper-pencil/computer-based).

For the 4th grade mathematics subtest, based on the MH method, a total of three items showed DIF at the B level, while the SIBTEST method showed five items with DIF, and the LR method did not reveal any DIF items. When comparing the MH and SIBTEST methods, three similar items with DIF were found in both methods, and two items showed DIF in the SIBTEST method but not in the MH method. Among the three DIF items identified in both the MH and SIBTEST methods, two items favored students taking the paper-pencil test (focus group), and two items favored students taking the computer-based test (reference group). These findings support Yörü and Atar's (2019) recommendation to use at least two methods to identify DIF, as the results obtained from the three DIF methods in the 4th grade mathematics test were qualitatively different. Additionally, in the study by Eriştiren (2021), it was observed that MH and SIBTEST techniques showed consistency, but LR method did not exhibit the same level of consistency, which aligns with the current study's results.

Regarding the 8th grade mathematics subtest, based on the MH method, seven items showed DIF, while the LR method did not reveal any DIF items, and the SIBTEST method showed 11 items with DIF. Among the DIF items in the SIBTEST method, four items were not present in the MH method. Four items among the DIF items in both the MH and SIBTEST methods favored the focus group, and three items favored the reference group. However, of the four other items marked DIF by SIBTEST, two favor focal and two favor reference group.

In the 4th grade science subtest, the MH method revealed one item with DIF, the LR method showed no DIF items, and the SIBTEST method showed six items with DIF. Among the DIF items, item 4 showed DIF only in the MH method and favored the focal group at the B level. The SIBTEST method flagged three items favor focal and the rest favor reference group. These results align with previous studies by Gök, Kelecioğlu, and Doğan (2010) and Ercikan and Koch (2009), indicating a low level of agreement between the MH and LR methods for DIF detection. Furthermore, similar findings were observed between this study and Eriştiren's (2021) study on measurement invariance using the results from the entrance exam for secondary education.

When examining the DIF results of the 8th grade science subtest, no items showed DIF in the MH and LR methods, while four items exhibited DIF in the SIBTEST method. Among the DIF items identified in the SIBTEST method, two favored the focal group, and two favored the reference group. However, the SIBTEST method revealed DIF in four items, indicating its lack of alignment with the other two methods. Overall, the DIF analyses conducted in this study suggest that using multiple methods, such as MH, LR, and SIBTEST, can enhance the accuracy of identifying DIF in educational assessments.

In terms of the DIF analyses conducted using the MH and SIBTEST techniques showed some agreement, for the disagreements SIBTEST flagged more items than the MH method. However, the LR approach did not agree with SIBTEST and MH, and did not flag any B or C level DIF in our analysis. In other words, no set of items was consistently advantageous or disadvantageous to either the reference or focus group across all subtest results based on the LR approach.

Overall, the MGCFA conclusions based on the ΔCFI and ΔTLI are in agreement with the LR approach, and they provide evidence for the measurement invariance. The MGCFA conclusions based on the χ^2 difference tests are in agreement with the SIBTEST and MH conclusions and they can arguably be considered as concerns about the invariance. These findings are inconsistent with some literature (Çepni, 2011; Wiberg, 2009) while being consistent with others (Doğan, 2008; Gök, 2010). Similarly, Eriştiren's (2021) study on measurement invariance and DIF in entrance exams to secondary education also presents similar findings to this study. While measurement invariance was largely achieved across all stages in the tests, discrepancies in DIF were observed, particularly concerning achievement levels based on school type, where the MH and SIBTEST analyses showed converging results, but the LR method exhibited incongruent results. Additionally, the discrepancies observed in the results of the study by Özdemir (2003) comparing two-category and partial credit scoring methods for multiple-choice items in a Turkish reading comprehension test support the outcomes of this study.

It should be noted that MGCFA analyses took into account the factor structure while the DIF analyses were conducted separately for each dimension. Despite our efforts to conduct multidimensional DIF our attempts to utilize R was unsuccessful probably due to the large sample size and relatively complex factor structure.

Our final attempt was to run DIF analyses for the entire test, assuming unidimensionality; with this assumption the number of flagged items were less compared to what we reported in this paper. To be on the conservative side, we reported the DIF analyses that conducted separately for each dimension. Future studies are needed to address this limitation.

Declarations

Conflict of Interest: No potential conflict of interest was reported by the authors.

Ethical Approval: I declare that all ethical guidelines for the author have been followed. This study does not require any ethics committee approval as it includes open-access data.

References

Agresti, A. (1984). *Analysis of ordinal categorical data*. New York: John Wiley & Sons.

Akyıldız, M. (2009). Pırls 2001 testinin yapı geçerliliğinin ülkelerarası karşılaştırılması. *Yüzüncü Yıl Üniversitesi Eğitim Fakültesi Dergisi*, *6*(1)

Anakwe, B. (2008). Comparison of student performance in paper-based versus ccomputer-based testing. *Journal of Education for Business.* September-October, 13-17.

Arım, G, R., Ercikan, K. (2014). Comparability between the American and Turkish versions of the TIMSS mathematics test results*. Eğitim ve Bilim*. *39*(172), 33- 48.

Atılgan, H., Kan, A., Aydın, B. (2017). *Eğitimde ölçme ve değerlendirme.* Onuncu Baskı. Ankara: Anı Yayıncılık.

Bağdu Söyler, P., Aydın, B., & Atılgan, H. (2021). PISA 2015 Reading Test Item Parameters Across Language Groups: A measurement Invariance Study with Binary Variables. *Journal of Measurement and Evaluation in Education and Psychology*, *12*(2), 112-128. https://doi.org/10.21031/epod.800697

Büyüköztürk, Ş. (2010). *Sosyal bilimler için veri analizi el kitabı*. Pegem Akademi:Ankara.

Büyüköztürk, Ş., Çokluk, Ö., & Şekercioğlu, G. (2014). Sosyal bilimler için çok değişkenli istatistik SPSS ve LISREL uygulamaları. Ankara: Pegem Akademi.

Camilli G. Shepard L. A. (1994). *Methods for Identifying Biased Test Items. Volume 4.* California: SAGE Publications. Inc.

Chalmers, R. P. (2012). mirt: A multidimensional item response theory package for the R environment. *Journal of statistical Software*, *48*, 1-29.

Cheung, G. W., Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural Equation Modeling, 9(*2), 233-255.

Çepni, Z. (2011). *Değişen madde fonksiyonlarının SIBTEST, Mantel-Haenzsel,lojistik regresyon ve madde tepki kuramı yöntemleriyle incelenmesi* (doktora tezi). Hacettepe Üniversitesi, Ankara

Doğan, N; Öğretmen, T. (2008). *Değişen madde fonksiyonunu belirlemede Mantel - Haenszel, ki-kare ve lojistik regresyon tekniklerinin karşılaştırılması.* Eğitim ve Bilim Dergisi. 33(148).

Dorans, N. J., & Holland, P. W. (1993). *DIF detection and description: Mantel-Haenszel and standardization.* In P. W. Holland & H. Wainer (Eds.), Differential item functioning (pp. 35–66). Lawrence Erlbaum Associates, Inc.

Drasgow, F (2002). The work ahead: a psychometric infrastructure for computerized adaptive tests. In C.N. Mills, M.T. Potenza, J.J. Fremer, & W.C. Ward (Eds.), Computer-based testing: Building the foundation for future assessments (pp. 67–88). Hillsdale, NJ: Lawrence Erlbaum.

Ercikan, K; Koh, K. (2009). *Examining the construct comparability of the English and French versions of TIMSS.* Internatıonal Journal Of Testıng, 5(1), 23–35

Ergün, E. (2002). *Üniversite öğrencilerinin bilgisayar destekli ölçmeden elde ettikleri eaşarının kalem-kâğıt testi başarısı, bilgisayar kaygısı ve bilgisayar tecrübeleri açısından* i*ncelenmesi*. Yayımlanmamış yüksek lisans tezi. Anadolu Üniversitesi Eğitim Bilimleri Enstitüsü, Eskişehir.

Eriştiren, İ. (2021). *Ortaöğretime Geçiş Sınavlarında ölçme değişmezliği ve DIF'nin incelenmesi* (Yüksek Lisans Tezi). Haccettepe Üniversitesi, Ankara.

Gök, B., Kelecioğlu, H. ve Doğan, N. (2010). Değişen madde fonksiyonunu belirlemede Mantel- Haenzsel ve lojistik regresyon tekniklerinin karşılaştırılması. *Eğitim ve Bilim*, *35*, 3-16.

Gierl, M. J., Jodoin, M. G., & Ackerman, T. A. (2000). American Educational Research Association (AERA) New Orleans, Louisiana, USA April 24-27, 2000.

Gündoğmuş, İ. (2017). *Kâğıt-kalem, bilgisayar ve tablet ortamında gerçekleştirilen sınavlar için ölçme değişmezliğinin ve öğrenci görüşlerinin incelenmesi*. Hacettepe Üniversitesi, Ankara

Gregorich, S. E. (2006). Do self-report instruments allow meaningful comparisons across diverse population groups?: Testing measurement invariance using the confirmatory factor analysis framework. *Medical Care*, *44*(1),

Hair, J.F. Jr., Anderson, R.E., Tatham, R.L., and Black, W.C. (1998). *Multivariate data analysis*, (5th Edition). Upper Saddle River, NJ: Prentice Hall.

Hambleton, R. K. (2006). Good practices for identifying differential item functioning. Medical Care, 44, 182-188.

İlci, B. (2004). *Geleneksel kâğıt-kalem yöntemi ile ve bilgisayarda online uygulanan çoktan seçmeli sayısal yetenek ve sözel yetenek testlerine ait madde ve test* i*statistiklerinin karşılaştırılması.* Yüksek lisans tezi. Hacettepe Üniversitesi Sosyal Bilimler Enstitüsü, Ankara.

Jöreskog, K. G. ve Sörbom, D. (2006). LISREL (Version 8.8) [computer software]. Chicago: Scientific Softare International Inc.

Kite, B. A., Johnson, P. E., & Xing, C. (2018, January 28). Replicating the Mplus DIFFTEST Procedure. https://pj.freefaculty.org/guides/crmda_guides/44.difftest/44.difftest.html

Klieme E., Baumert J. (2001). Identifying national cultures of mathematics education: Analysis of cognitive demands and differential item functioning in TIMSS. *European Journal of Psychology of Education, 16:3, 385-402.*

Kline, R. B. (2016). *Principles and practice of structural equation modeling* (4th ed.b.). New York & London: The Guilford Press.

Li, Z., Gooden, C. J., & Toland, M. D. (2016). Measurement invariance with categorical indicators. *Applied Psychometric Strategies Lab, Applied Quantitative and Psychometric Series. Presentation conducted at the University of Kentucky, Lexington, KY. Retrieved from https://education. uky. edu/edp/apslab/events*.

MEB. (2020). *TIMSS 2019 ulusal matematik ve fen bilimleri ön raporu: 4. ve 8. sınıflar.* Ankara.

Meredith, W., & Teresi, J. A. (2006). *An essay on measurement and factorial invariance.* Medical care, 44(11), 69-S77.

Mertler, C. A. & Vannatta, R. A. (2005). *Advanced and multivariate statistical methods: Practical application and interpretation* (3rd ed.). Los Angeles: Pyrczak.

Mills, C. N., Potenza, M.T., Fremer, J.J., Ward, W.C. (2001). *Computer Based Testing: Building the Foundation for Future Assessment.* Lawrance Erlbaum Associates, Publishers: Londra

Moraes, C.L & Reichenheim, M.E. (2002). Cross-cultural measurement equivalence of the revised conflict tactics scales (cts2) portuguese version used to identify violence within couples. Cad. Saúde Pública, 18 (3).

Muthén, B. O. (1993). Goodness of fit with categorical and other non-normal variables. In K. A. Bollen & J. S. Long (Eds.), Testing structural equation models (pp. 205–243). Newbury Park, CA: Sage.

Mullis, I. V. S., Martin, M. O., Foy, P., Kelly, D., & Fishbein, B. (2020). *TIMSS 2019 international results in mathematics and science*. Boston College, TIMSS & PIRLS International Study Center.

Nandakumar, R. (1993). A fortran 77 program for detecting differential item functioning through the mantel-haenszel statistic. *Educational and Psychological Measurement, 53*, 679–684.

Osterlind S. J. Everson H. T. (2009). *Differential Item Functioning: Second Edition.* California: SAGE Publications. Inc.

Özdemir, D. (2003). Çoktan seçmeli testlerde iki kategorili ve önsel ağırlıklı puanlamanın değişen madde fonksiyonuna etkisi ile ilgili bir araştırma. *Eğitim ve Bilim, 28(*129), 37-43.

Raju, N. S., Laffitte, L. J., & Byrne, B. M. (2002). Measurement equivalence: A comparison of methods based on confirmatory factor analysis and item response theory. *Journal of Applied Psychology, 87*(3), 517-529.

Raykov, T., Dimitrov, D. M., Marcoulides, G. A., Li, T., & Menold, N. (2018). Examining measurement invariance and differential item functioning with discrete latent construct indicators: A note on a multiple testing procedure. *Educational and Psychological Measurement*, *78*(2), 343-352.

Rogers, T. B. (1995). *The psychological testing enterprise: An introduction.* Pasific Grove, California: Brooks/Cole.

Russel, M., Goldberg, A., O'Connor, K. (2003). Computer based test and validity: A look back into the future. *Assessment in Education.* 10, 279- 293.

Shealy, R. and Stout, W. (1993). A model-based standardization approach that separates true bias/DIF from group ability differences and detect test bias/DTF as well as item bias/DIF. *Psychometrika*, *58*, 159-194. doi: 10.1007/BF02294572

Steenkamp, B., E., M. and Baumgartner, H. (1998). Assessing measurement invariance in cross-national consumer research. journal of consumer research, 25(1),78-107.

Tabachnick, B. G. & Fidell, L. S. (2007). *Using multivariate statistics*. Boston: Pearson Education.

Wiberg, M. (2009). Differential item functioning in mastery tests: A comparison of three methods using real data. International *Journal of Testing*, *9*, 41–59

Vandenberg, R. J., Lance, C. E. (2000). A review and synthesis of the measurement invariance literature: Suggestions, practices, and recommendations for organizational research. *Organizational Research Methods,* 4, 4-70

Wu, A. D., Li, Z. and Zumbo, B. D. (2007). Decoding the meaning of factorial invariance and updating the practice of multigroup confirmatory factor analysis: a demonstration with TIMSS data. *Practical Assessment, Research and Evaluation,* 12, 1-26.

Zumbo, B. D. (1999). *A handbook on the theory and methods of differential item functioning (DIF) logistic regression modeling as a unitary framework for binary and Likert-type (ordinal) item scores*. Canada: Ottowa, Directorate of Human Resources Research and Evaluation National Defense Headquarters: Author.

Appendix

Appendix 1. 4th Grade Science VIF Analysis Results

ITEMS	Tolerance	VIF
NEW KIND OF MAMMAL DISCOVERED (A)	0,810	1,234
COVER YOUR MOUTH THOUGH NOT SICK (1)	0,862	1,160
HAMAD'S GARDEN: WHICH SURVIVE (1)	0,873	1,146
HAMAD'S GARDEN: PLANT STRUCTURE (1)	0,915	1,092
TWO THINGS ANIMALS NEED (1)	0,893	1,120
CELERY STALK LEAVES TURN RED (B)	0,821	1,217
WOODEN AND METAL CUBES ON BALANCE (B)	0,902	1,109
TWO METAL BARS (C)	0,858	1,166
DROPS OF WAX ON A METAL FRAME (1)	0,722	1,385
OBJECT INSIDE A WOODEN BOX (C)	0,897	1,115
AMOUNT OF WATER AND LAND ON EARTH (D)	0,892	1,121
WHAT MAKES UP SOLAR SYSTEM (C)	0,809	1,236
LIVING AND NON-LIVING THINGS IN A DESERT (1)	0,863	1,159
HUMAN ORGAN WITH SAME FUNCTION AS GILLS (B)	0,789	1,267
CHARACTERISTICS OF LIVING AND TOY DUCK (DERIVED) (1)	0,811	1,233
EXPLAIN DECREASE IN INSECT POPULATION (1)	0,727	1,376
WHAT MAKES VENUS FLYTRAP DIFFERENT FROM OTHER PLANTS (B)	0,904	1,107
WHY GROUND SQUIRREL HOLDS TAIL OVERHEAD (1)	0,763	1,311
CHANGE WHERE MATERIALS IN OBJECTS STAY THE SAME (A)	0,911	1,098
CAUSE OF SKYDIVER'S FALL (C)	0,822	1,217
ENERGY CHANGE IN A FLASHLIGHT (A)	0,889	1,125
WHY MARY'S BOX IS EASIER TO MOVE (D)	0,817	1,225
ADVANTAGES TO FARMING NEAR A RIVER (1)	0,843	1,186
DISADVANTAGES TO FARMING NEAR A RIVER (1)	0,809	1,236
POSITION OF THE EARTH WHEN IT IS SUMMER IN CITY A (C)	0,920	1,087

Appendix 2. 4th Grade Mathematics VIF Analysis Results

ITEMS	Tolerance	VIF
NUMBERS WITH 6 AS A FACTOR (DERIVED) (1)	0,898	1,114
FIGURE WITH THREE QUARTERS SHADED (A)	0,856	1,168
WHO PAID LESS FOR EACH BOTTLE (1)	0,756	1,323
FRACTION WATERED ON MONDAY (1)	0,404	2,475
FRACTION WATERED ON TUESDAY (1)	0,373	2,682
NEXT 2 NUMBERS IN THE PATTERN (DERIVED) (1)	0,686	1,458
STREET PARALLEL TO GREEN STREET (A)	0,839	1,192
PERPENDICULAR TO APPLE STREET (B)	0,940	1,064
NUMBER OF TRIANGLES NEEDED (B)	0,908	1,101
SHAPE THAT FOLDS INTO A BOX (D)	0,940	1,064
MOST FREQUENT SCORE ON QUIZ (1)	0,818	1,223
SCORE OF 4 OR MORE ON QUIZ (1)	0,728	1,374
NUMBER WITH 7 HUNDREDS AND 6 ONES (C)	0,876	1,141
DISTANCE TRAVELED EACH DAY ON BICYCLE (B)	0,756	1,323
FRACTIONS GREATER THAN 1/2 (DERIVED) (1)	0,726	1,378
EXPRESSION FOR STICKERS GIVEN TO EACH FRIEND (D)	0,745	1,343
COST BANANAS AND PLUMS (DERIVED) (2)	0,828	1,208
UNITS FOR MEASUREMENTS (DERIVED) (1)	0,882	1,134
WEIGHT OF 1 PEAR (C)	0,807	1,240
NUMBER OF SHAPES TO COVER SQUARE (DERIVED) (2)	0,763	1,311
COMPLETE FIGURE WITH LINE OF SYMMETRY (1)	0,867	1,154
WATER LEVEL IN DAM - WEEK 8 (1)	0,811	1,233
PICTOGRAPH OF ANIMAL WEIGHTS (DERIVED) (1)	0,738	1,355
BAR GRAPH OF CARS EACH MORNING (DERIVED) (1)	0,669	1,495

Appendix 3. 8th Grade Science VIF Analysis Results

ITEMS	Tolerance	VIF
PENGUIN BEHAVIOR AND SURVIVAL (2)	0,859	1,164
ORGANISM WITH CELL WALLS (C)	0,898	1,114
HOW DECOMPOSERS GET ENERGY (B)	0,821	1,217
ORGANISM THAT COMPETES WITH HUMANS (1)	0,760	1,317
GARDEN WITH BIRD FEEDER (DERIVED) (1)	0,869	1,151
WHY SOLUTION 2 IS PALER THAN 1 (1)	0,796	1,256
WHICH IS A PHYSICAL CHANGE (D)	0,896	1,116
MODEL FLASHLIGHT: BULB WON'T LIGHT (1)	0,840	1,190
MODEL FLASHLIGHT: 2 PARALLEL BULBS (1)	0,814	1,229
MODEL FLASHLIGHTS: COMPARISON (C)	0,923	1,083
TWO BAR MAGNETS REPELLING (A)	0,818	1,223
PLANETS: SHORTEST DAY LENGTH (D)	0,887	1,128
PLANETS: DISTANCE FROM SUN (1)	0,759	1,318
TEMPERATURE OUTSIDE AN AIRPLANE (A)	0,769	1,300
RELATIONSHIP BETWEEN INSECTS AND FLOWERING PLANTS (D)	0,827	1,210
WHERE IN A CELL DNA REPLICATION OCCURS (B)	0,902	1,108
INCREASE GREEN SPACE AS CARBON DIOXIDE INCREASES (1)	0,689	1,451
WHY LEAVES' MASSES DECREASED (C)	0,901	1,110
CLASSIFY ANIMALS BASED ON A SINGLE CHARACTERISTIC (1)	0,762	1,312
IDENTIFY THE CHARACTERISTIC USED TO CLASSIFY ANIMALS (1)	0,863	1,158
LOCATION OF SUBATOMIC PARTICLES (1)	0,831	1,203
ORDER ELEMENTS FROM SMALLEST TO LARGEST ATOMIC NUM (1)	0,804	1,244
ACIDIC, BASIC, OR NEUTRAL SOLUTION (DERIVED) (1)	0,814	1,229
MIXING AN ACID AND BASE SOLUTION (D)	0,837	1,195
GAS MOLECULES IN AN EXPANDING BALLOON (A)	0,850	1,177
THINGS TOM SHOULD DO (DERIVED) (1)	0,612	1,633
VEHICLE WITH DIFFERENT WEIGHTS ON DIFFERENT PLANETS (D)	0,747	1,338
CELL PHONE IN A VACUUM (1)	0,743	1,346
WHY BALLOON GETS BIGGER AS IT RISES (B)	0,923	1,083
EVIDENCE OF GLOBAL WARMING (A)	0,749	1,335
NATURAL RESOURCE FORMATION SHOWN IN DIAGRAMS (C)	0,866	1,154

Appendix 4. 8th Grade Mathematics VIF Analysis Results

ITEMS	Tolerance	VIF
OCTAGON WITH EQUIVALENT SHADING (B)	0,740	1,352
TIME WHEN PAT FINISHES LAST LAP (1)	0,677	1,476
PERCENTAGE OF LAPS FINISHED (1)	0,633	1,581
MULTIPLES OF 3 (D)	0,745	1,342
CONVERT DECIMAL TO A FRACTION (1)	0,725	1,378
EXPRESSION FOR AREA OF RECTANGLE (C)	0,738	1,355
EXPRESSION WITH EXPONENTS OF Y (B)	0,725	1,380
NUMBER OF MATCHES FOR FIGURE 10 (1)	0,768	1,303
RULE FOR NUMBER OF MATCHES (1)	0,652	1,534
GRAPH OF Y = 2X (A)	0,884	1,132
ROTATION AND REFLECTION (D)	0,921	1,086
SURFACE AREA OF THE PRISM (C)	0,805	1,242
VALUE OF ANGLE X OUTSIDE TRIANGLE (C)	0,740	1,351
NUMBER OF BALLS IN A BAG (B)	0,753	1,327
LIV'S SMARTPHONE USE (D)	0,720	1,389
SMARTPHONE USE LISTENING TO MUSIC (A)	0,769	1,300
STATEMENTS FOR ALL VALUES OF INTEGER A (DERIVED) (2)	0,752	1,329
ARROW TO SHOW 5/12 ON NUMBER LINE (B)	0,743	1,345
VALUE OF FRACTION X IN SQUARE (1)	0,681	1,469
NUMBER OF BLUE BEADS ON BRACELET (1)	0,762	1,312
VALUE OF 2(6X - 3Y) WHEN X = 3 AND Y = 2 (C)	0,752	1,329
EXPRESSION EQUIVALENT TO 2Y + 6XY2 (A)	0,761	1,315
FORMULA FOR STOPPING DISTANCE (1)	0,624	1,601
VALUE OF X GIVEN PERIMETER OF TRIANGLE ABC (1)	0,542	1,844
ADDITIONAL POINT ON A STRAIGHT LINE (D)	0,776	1,288
VALUE OF ANGLE X IN A QUADRILATERAL (1)	0,634	1,578
METHODS OF FOLDING PAPER (DERIVED) (1)	0,846	1,182
COORDINATES TO COMPLETE KLMN (DERIVED) (1)	0,623	1,606
MEAN TEMPERATURE FOR 5 DAYS (1)	0,587	1,704
BEST GRAPH FOR TOWN INFORMATION (DERIVED) (1)	0,774	1,292
BAR GRAPH OF NEWSPAPER SALES (1)	0,764	1,309

Appendix 5. 4th Grade Science Tetrachoric Correlation Analysis Results

	M1	M2	M3	M4	M5	M6	M7	M8	M9	M10	M11	M12	M13	M14	M15	M16	M17	M18	M19	M20	M21	M22	M23	M24	M25
M1	1																								
M2	0.34	1																							
M3	0.35	0.31	1																						
M4	0.3	0.23	0.32	1																					
M5	0.26	0.25	0.24	0.25	1																				
M6	0.36	0.34	0.32	0.27	0.27	1																			
M7	0.22	0.2	0.24	0.21	0.18	0.24	1																		
M8	0.29	0.25	0.26	0.23	0.21	0.31	0.25	1																	
M9	0.44	0.4	0.41	0.36	0.3	0.43	0.33	0.36	1																
M10	0.26	0.2	0.23	0.22	0.2	0.27	0.21	0.27	0.32	1															
M11	0.24	0.19	0.21	0.23	0.2	0.21	0.21	0.24	0.29	0.19	1														
M12	0.34	0.26	0.31	0.27	0.22	0.34	0.26	0.31	0.43	0.25	0.32	1													
M13	0.27	0.27	0.29	0.29	0.29	0.27	0.19	0.24	0.34	0.21	0.25	0.29	1												
M14	0.4	0.29	0.33	0.29	0.28	0.39	0.25	0.3	0.44	0.28	0.27	0.4	0.31	1											
M15	0.38	0.27	0.29	0.28	0.25	0.34	0.26	0.28	0.42	0.26	0.24	0.36	0.29	0.42	1										
M16	0.47	0.33	0.4	0.36	0.29	0.41	0.27	0.36	0.49	0.3	0.3	0.41	0.34	0.45	0.39	1									
M17	0.26	0.19	0.24	0.2	0.17	0.23	0.2	0.21	0.29	0.16	0.17	0.24	0.2	0.27	0.27	0.3	1								
M18	0.4	0.33	0.36	0.33	0.3	0.37	0.23	0.31	0.45	0.29	0.28	0.36	0.3	0.41	0.36	0.5	0.29	1							
M19	0.24	0.2	0.21	0.22	0.16	0.22	0.18	0.21	0.28	0.16	0.21	0.23	0.18	0.23	0.25	0.28	0.16	0.24	1						
M20	0.32	0.3	0.3	0.21	0.23	0.35	0.2	0.27	0.38	0.26	0.25	0.38	0.27	0.4	0.35	0.39	0.2	0.37	0.21	1					
M21	0.3	0.18	0.21	0.2	0.19	0.25	0.18	0.23	0.31	0.21	0.21	0.24	0.17	0.29	0.28	0.31	0.18	0.3	0.25	0.26	1				
M22	0.31	0.31	0.32	0.25	0.25	0.35	0.24	0.32	0.42	0.27	0.26	0.34	0.27	0.37	0.35	0.41	0.25	0.39	0.25	0.36	0.3	1			
M23	0.29	0.28	0.28	0.28	0.22	0.26	0.17	0.24	0.34	0.2	0.19	0.27	0.29	0.31	0.28	0.36	0.21	0.35	0.18	0.29	0.22	0.3	1		
M24	0.33	0.28	0.33	0.29	0.24	0.32	0.26	0.29	0.4	0.28	0.24	0.3	0.29	0.37	0.31	0.41	0.25	0.4	0.24	0.34	0.27	0.35	0.47	1	
M25	0.22	0.17	0.2	0.17	0.13	0.22	0.19	0.2	0.27	0.18	0.19	0.27	0.17	0.23	0.23	0.27	0.19	0.25	0.14	0.25	0.16	0.24	0.15	0.23	1

Appendix 6. 4th Grade Mathematics Tetrachoric Correlation Analysis Results

	M1	M2	M3	M4	M5	M6	M7	M8	M9	M10	M11	M12	M13	M14	M15	M16	M17	M18	M19	M20	M21	M22	M23	M24
M1	1																							
M2	0.1	1																						
M3	0.33	0.29	1																					
M4	0.19	0.49	0.43	1																				
M5	0.2	0.54	0.46	0.93	1																			
M6	0.32	0.37	0.54	0.5	0.53	1																		
M7	0.17	0.29	0.33	0.38	0.43	0.36	1																	
M8	0.14	0.17	0.26	0.25	0.27	0.26	0.21	1																
M9	0.14	0.24	0.26	0.32	0.35	0.32	0.24	0.15	1															
M10	0.13	0.16	0.24	0.22	0.23	0.25	0.15	0.12	0.17	1														
M11	0.17	0.26	0.35	0.37	0.4	0.39	0.28	0.17	0.26	0.17	1													
M12	0.23	0.35	0.46	0.47	0.52	0.5	0.4	0.24	0.31	0.2	0.69	1												
M13	0.25	0.29	0.32	0.36	0.38	0.41	0.24	0.2	0.27	0.17	0.28	0.34	1											
M14	0.31	0.23	0.55	0.39	0.4	0.5	0.31	0.25	0.25	0.17	0.31	0.41	0.33	1										
M15	0.24	0.47	0.5	0.55	0.62	0.52	0.45	0.26	0.3	0.21	0.37	0.5	0.35	0.43	1									
M16	0.3	0.31	0.54	0.46	0.5	0.51	0.35	0.24	0.29	0.2	0.33	0.46	0.39	0.54	0.46	1								
M17	0.28	0.28	0.5	0.4	0.42	0.48	0.32	0.25	0.25	0.23	0.31	0.38	0.31	0.44	0.5	0.45	1							
M18	0.26	0.16	0.31	0.25	0.27	0.32	0.25	0.14	0.19	0.12	0.2	0.28	0.21	0.33	0.27	0.35	0.29	1						
M19	0.23	0.27	0.43	0.37	0.41	0.44	0.29	0.2	0.24	0.21	0.29	0.38	0.29	0.39	0.44	0.4	0.43	0.29	1					
M20	0.28	0.33	0.47	0.45	0.49	0.5	0.38	0.25	0.3	0.29	0.33	0.44	0.35	0.43	0.48	0.45	0.47	0.35	0.44	1				
M21	0.22	0.25	0.34	0.34	0.39	0.38	0.24	0.19	0.26	0.23	0.26	0.29	0.27	0.26	0.39	0.34	0.34	0.24	0.28	0.38	1			
M22	0.22	0.34	0.37	0.45	0.48	0.44	0.27	0.23	0.31	0.17	0.37	0.42	0.34	0.36	0.36	0.45	0.32	0.3	0.34	0.38	0.36	1		
M23	0.32	0.33	0.49	0.44	0.48	0.52	0.34	0.22	0.29	0.22	0.34	0.46	0.38	0.45	0.5	0.49	0.45	0.35	0.43	0.48	0.4	0.47	1	
M24	0.25	0.43	0.51	0.53	0.57	0.55	0.42	0.28	0.34	0.27	0.4	0.52	0.37	0.43	0.57	0.48	0.49	0.31	0.46	0.55	0.42	0.52	0.56 1.00	1

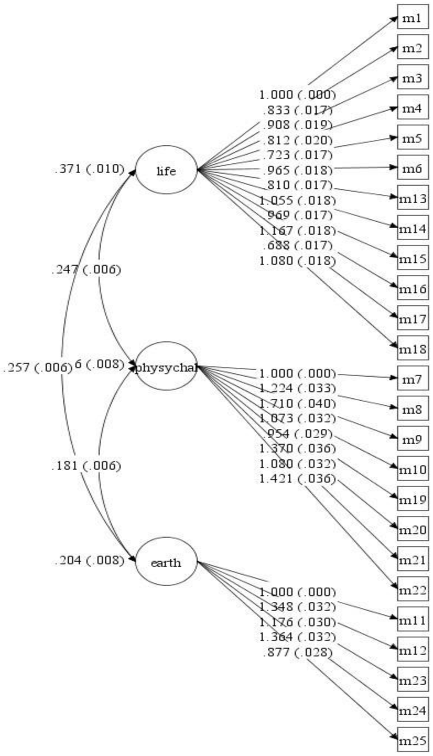
Appendix 7. 8th Grade Science Tetrachoric Correlation Analysis Results

	M1	M2	M3	M4	M5	M6	M7	M8	M9	M10	M11	M12	M13	M14	M15	M16	M17	M18	M19	M20	M21	M22	M23	M24	M25	M26	M27	M28	M29	M30	M31
M1	1																														
M2	0.21	1																													
M3	0.3	0.26	1																												
M4	0.33	0.25	0.38	1																											
M5	0.25	0.14	0.25	0.34	1																										
M6	0.36	0.2	0.29	0.37	0.31	1																									
M7	0.21	0.22	0.26	0.31	0.19	0.22	1																								
M8	0.25	0.18	0.27	0.34	0.24	0.34	0.21	1																							
M9	0.26	0.24	0.29	0.38	0.26	0.36	0.23	0.39	1																						
M10	0.2	0.14	0.19	0.21	0.18	0.24	0.14	0.2	0.2	1																					
M11	0.25	0.18	0.31	0.38	0.25	0.32	0.23	0.28	0.31	0.2	1																				
M12	0.23	0.2	0.26	0.25	0.22	0.27	0.17	0.2	0.23	0.16	0.27	1																			
M13	0.36	0.24	0.36	0.43	0.29	0.41	0.26	0.34	0.38	0.27	0.38	0.32	1																		
M14	0.34	0.19	0.34	0.4	0.33	0.42	0.23	0.27	0.3	0.24	0.35	0.32	0.44	1																	
M15	0.25	0.13	0.28	0.41	0.29	0.28	0.23	0.26	0.25	0.19	0.34	0.17	0.35	0.36	1																
M16	0.18	0.21	0.23	0.28	0.16	0.2	0.22	0.18	0.21	0.14	0.23	0.15	0.25	0.21	0.23	1															
M17	0.37	0.26	0.41	0.49	0.33	0.45	0.31	0.41	0.42	0.24	0.39	0.3	0.48	0.45	0.36	0.31	1														
M18	0.18	0.13	0.21	0.26	0.23	0.23	0.17	0.24	0.23	0.15	0.25	0.17	0.26	0.25	0.26	0.16	0.3	1													
M19	0.29	0.25	0.35	0.44	0.31	0.33	0.26	0.29	0.29	0.22	0.37	0.27	0.4	0.38	0.41	0.29	0.44	0.3	1												
M20	0.21	0.19	0.27	0.37	0.26	0.27	0.19	0.25	0.28	0.13	0.29	0.18	0.35	0.29	0.33	0.24	0.37	0.24	0.51	1											
M21	0.24	0.22	0.27	0.33	0.23	0.31	0.26	0.29	0.31	0.18	0.28	0.23	0.35	0.28	0.28	0.23	0.43	0.2	0.32	0.3	1										
M22	0.28	0.27	0.32	0.39	0.24	0.33	0.27	0.29	0.33	0.21	0.33	0.29	0.39	0.36	0.22	0.27	0.41	0.23	0.38	0.31	0.37	1									
M23	0.22	0.27	0.31	0.33	0.16	0.24	0.24	0.28	0.29	0.12	0.3	0.2	0.32	0.27	0.27	0.23	0.39	0.24	0.33	0.3	0.32	0.34	1								
M24	0.21	0.25	0.27	0.3	0.16	0.24	0.21	0.25	0.25	0.14	0.26	0.19	0.28	0.26	0.22	0.22	0.34	0.18	0.28	0.26	0.27	0.31	0.48	1							
M25	0.25	0.22	0.25	0.31	0.23	0.27	0.25	0.24	0.29	0.16	0.29	0.21	0.33	0.31	0.28	0.2	0.37	0.22	0.3	0.25	0.32	0.34	0.31	0.26	1						
M26	0.42	0.34	0.43	0.51	0.36	0.5	0.31	0.38	0.43	0.34	0.44	0.38	0.54	0.53	0.38	0.31	0.56	0.33	0.49	0.38	0.4	0.5	0.4	0.37	0.41	1					
M27	0.35	0.25	0.36	0.42	0.34	0.42	0.27	0.34	0.4	0.22	0.34	0.33	0.46	0.43	0.33	0.21	0.48	0.28	0.36	0.32	0.36	0.39	0.34	0.36	0.36	0.62	1				
M28	0.31	0.27	0.36	0.44	0.27	0.36	0.27	0.35	0.37	0.21	0.39	0.28	0.44	0.43	0.35	0.27	0.52	0.28	0.45	0.35	0.34	0.37	0.38	0.34	0.32	0.52	0.46	1			
M29	0.15	0.09	0.2	0.23	0.17	0.18	0.13	0.15	0.15	0.13	0.22	0.17	0.24	0.28	0.23	0.14	0.22	0.12	0.26	0.19	0.15	0.25	0.18	0.16	0.19	0.27	0.19	0.26	1		
M30	0.35	0.24	0.38	0.39	0.3	0.37	0.22	0.31	0.3	0.25	0.36	0.29	0.45	0.46	0.34	0.21	0.45	0.25	0.43	0.32	0.31	0.39	0.3	0.28	0.3	0.58	0.45	0.44	0.25	1	
M31	0.25	0.26	0.32	0.33	0.17	0.24	0.21	0.2	0.25	0.16	0.25	0.22	0.31	0.27	0.22	0.2	0.32	0.17	0.29	0.22	0.24	0.29	0.3	0.29	0.22	0.36	0.32	0.33	0.13	0.28	1

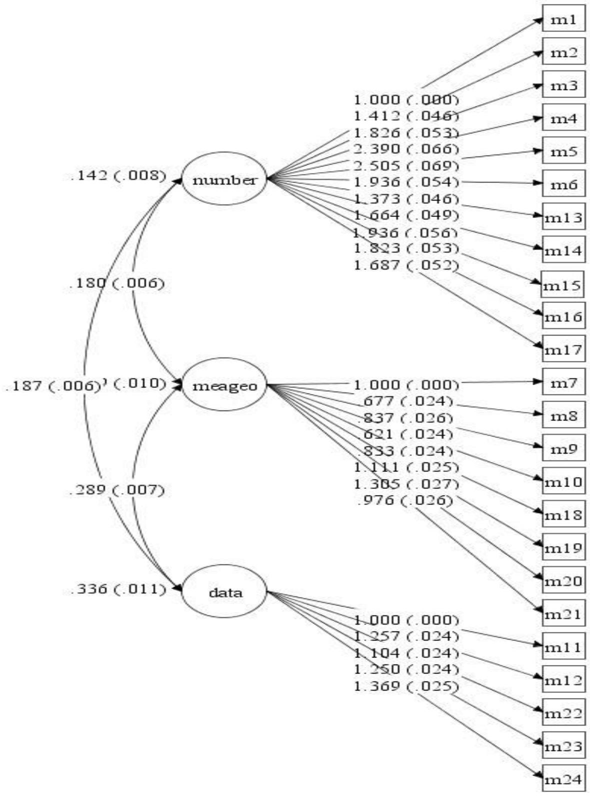
Appendix 8. 8th Grade Mathematics Tetrachoric Correlation Analysis Results

	M1	M2	M3	M4	M5	M6	M7	M8	M9	M10	M11	M12	M13	M14	M15	M16	M17	M18	M19	M20	M21	M22	M23	M24	M25	M26	M27	M28	M29	M30	M31
M1	1																														
M2	0.46	1																													
M3	0.61	0.71	1																												
M4	0.41	0.46	0.51	1																											
M5	0.38	0.45	0.52	0.41	1																										
M6	0.36	0.38	0.45	0.44	0.48	1																									
M7	0.41	0.47	0.5	0.45	0.49	0.48	1																								
M8	0.35	0.42	0.45	0.39	0.4	0.36	0.38	1																							
M9	0.51	0.51	0.6	0.52	0.56	0.52	0.52	0.72	1																						
M10	0.27	0.26	0.37	0.3	0.29	0.28	0.29	0.24	0.42	1																					
M11	0.26	0.19	0.31	0.25	0.2	0.19	0.2	0.22	0.34	0.2	1																				
M12	0.41	0.36	0.48	0.35	0.33	0.36	0.33	0.31	0.46	0.27	0.24	1																			
M13	0.4	0.48	0.5	0.4	0.4	0.37	0.42	0.35	0.47	0.26	0.21	0.36	1																		
M14	0.44	0.54	0.58	0.39	0.38	0.39	0.43	0.38	0.46	0.23	0.2	0.36	0.41	1																	
M15	0.46	0.46	0.59	0.43	0.4	0.38	0.41	0.34	0.51	0.28	0.28	0.39	0.4	0.41	1																
M16	0.42	0.43	0.54	0.39	0.35	0.32	0.36	0.33	0.48	0.25	0.23	0.35	0.35	0.37	0.52	1															
M17	0.42	0.44	0.52	0.48	0.51	0.52	0.47	0.37	0.56	0.38	0.25	0.42	0.43	0.39	0.48	0.41	1														
M18	0.46	0.46	0.56	0.4	0.39	0.38	0.41	0.32	0.49	0.27	0.21	0.41	0.38	0.44	0.47	0.4	0.46	1													
M19	0.51	0.5	0.62	0.51	0.53	0.47	0.46	0.45	0.65	0.38	0.32	0.45	0.45	0.45	0.5	0.48	0.54	0.52	1												
M20	0.41	0.51	0.55	0.38	0.36	0.31	0.37	0.38	0.48	0.23	0.2	0.35	0.38	0.42	0.4	0.39	0.39	0.41	0.52	1											
M21	0.37	0.45	0.49	0.42	0.45	0.43	0.47	0.34	0.49	0.26	0.18	0.34	0.38	0.41	0.39	0.36	0.45	0.4	0.48	0.36	1										
M22	0.32	0.36	0.44	0.41	0.42	0.51	0.46	0.32	0.49	0.26	0.17	0.31	0.35	0.34	0.38	0.31	0.49	0.37	0.45	0.32	0.41	1									
M23	0.47	0.52	0.58	0.5	0.54	0.5	0.53	0.43	0.58	0.34	0.25	0.4	0.44	0.47	0.5	0.45	0.52	0.49	0.59	0.47	0.58	0.5	1								
M24	0.55	0.6	0.67	0.56	0.6	0.58	0.58	0.48	0.67	0.39	0.31	0.51	0.54	0.53	0.58	0.54	0.62	0.57	0.68	0.57	0.6	0.55	0.71	1							
M25	0.37	0.4	0.49	0.38	0.36	0.34	0.36	0.3	0.46	0.28	0.26	0.35	0.34	0.36	0.4	0.37	0.42	0.39	0.47	0.36	0.36	0.35	0.47	0.53	1						
M26	0.45	0.57	0.57	0.46	0.52	0.45	0.53	0.43	0.56	0.24	0.21	0.36	0.62	0.49	0.44	0.39	0.45	0.45	0.53	0.47	0.48	0.45	0.58	0.66	0.42	1					
M27	0.31	0.38	0.38	0.34	0.31	0.28	0.32	0.29	0.39	0.21	0.17	0.31	0.3	0.32	0.32	0.31	0.32	0.31	0.37	0.32	0.32	0.28	0.38	0.45	0.32	0.38	1				
M28	0.49	0.53	0.56	0.5	0.52	0.48	0.5	0.46	0.6	0.39	0.3	0.46	0.47	0.5	0.47	0.43	0.54	0.48	0.59	0.48	0.53	0.48	0.62	0.67	0.57	0.59	0.44	1			
M29	0.55	0.59	0.63	0.53	0.53	0.5	0.53	0.45	0.6	0.36	0.28	0.48	0.52	0.57	0.55	0.51	0.52	0.55	0.61	0.51	0.55	0.48	0.63	0.7	0.5	0.64	0.47	0.66	1		
M30	0.38	0.48	0.49	0.39	0.36	0.31	0.38	0.35	0.44	0.21	0.23	0.31	0.37	0.44	0.37	0.37	0.35	0.36	0.43	0.41	0.39	0.34	0.46	0.48	0.37	0.45	0.31	0.51	0.52	1	
M31	0.46	0.51	0.56	0.48	0.46	0.43	0.48	0.4	0.55	0.34	0.29	0.42	0.42	0.49	0.5	0.47	0.48	0.49	0.53	0.44	0.43	0.44	0.53	0.59	0.44	0.5	0.34	0.54	0.56	0.45	1

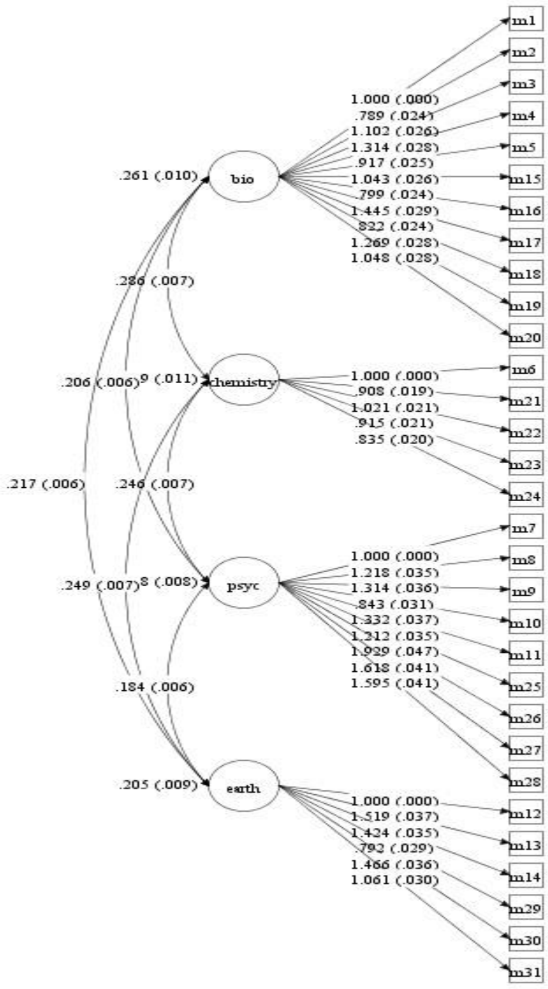
Appendix 9. 4th Grade Science CFA Path Diagram



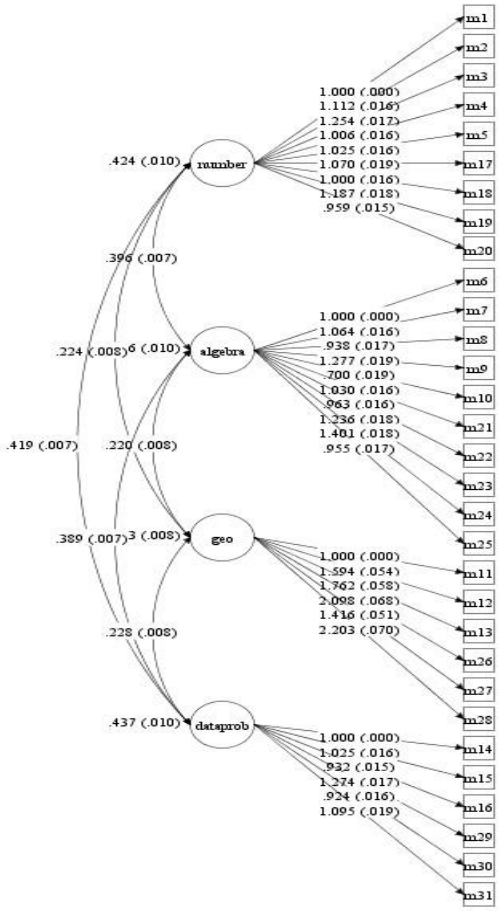
Appendix 10. 4th Grade Mathematics CFA Path Diagram



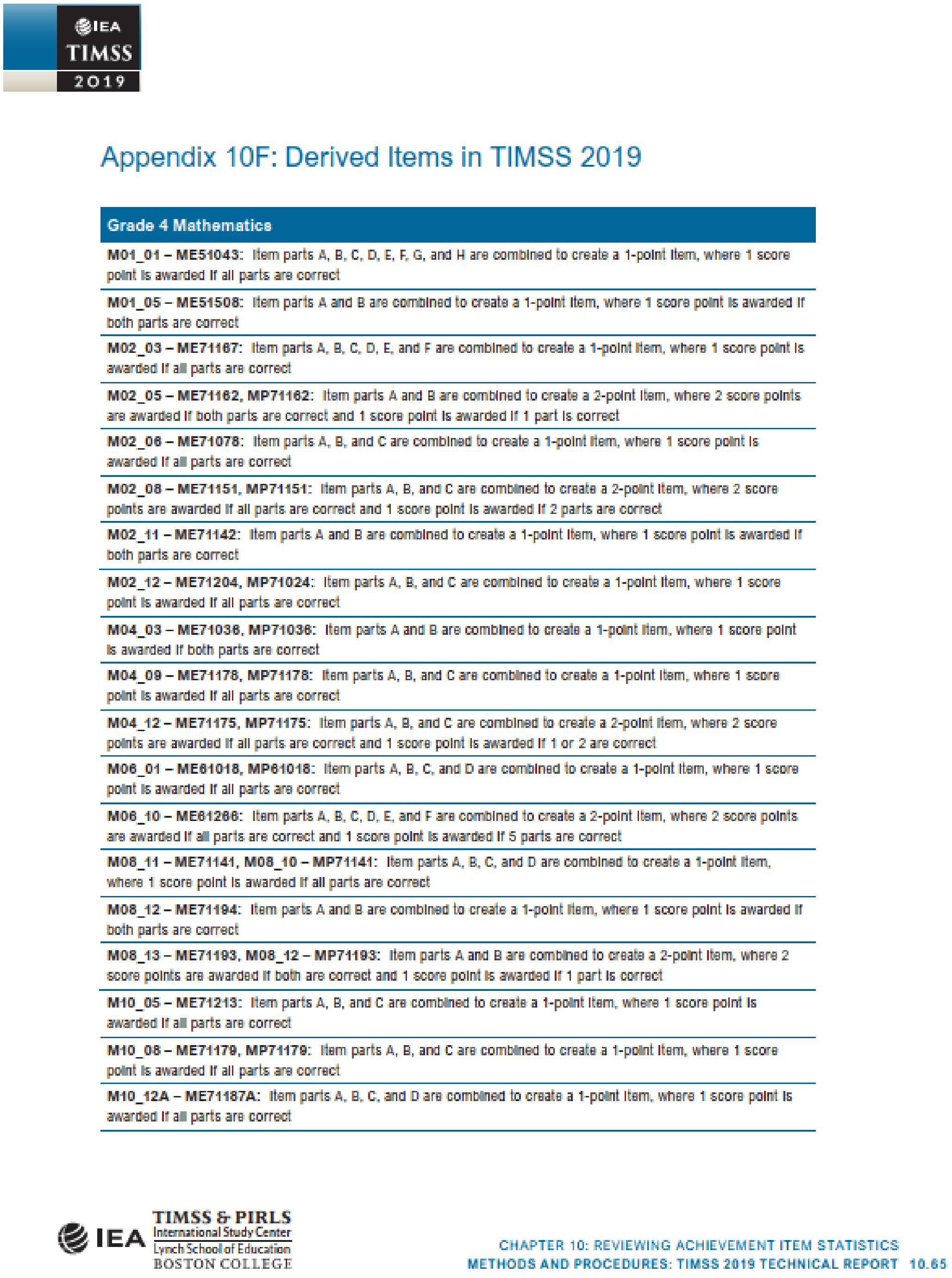
Appendix 11. 8th Grade Science CFA Path Diagram



Appendix 12. 8th Grade Mathematics CFA Path Diagram



Appendix 13. *Derived Items in TIMSS 2019*



IEA TIMSS 2019

Appendix 10F: Derived Items in TIMSS 2019

Grade 4 Mathematics
M01_01 – ME51043: Item parts A, B, C, D, E, F, G, and H are combined to create a 1-point item, where 1 score point is awarded if all parts are correct
M01_05 – ME51508: Item parts A and B are combined to create a 1-point item, where 1 score point is awarded if both parts are correct
M02_03 – ME71167: Item parts A, B, C, D, E, and F are combined to create a 1-point item, where 1 score point is awarded if all parts are correct
M02_05 – ME71162, MP71162: Item parts A and B are combined to create a 2-point item, where 2 score points are awarded if both parts are correct and 1 score point is awarded if 1 part is correct
M02_06 – ME71078: Item parts A, B, and C are combined to create a 1-point item, where 1 score point is awarded if all parts are correct
M02_08 – ME71151, MP71151: Item parts A, B, and C are combined to create a 2-point item, where 2 score points are awarded if all parts are correct and 1 score point is awarded if 2 parts are correct
M02_11 – ME71142: Item parts A and B are combined to create a 1-point item, where 1 score point is awarded if both parts are correct
M02_12 – ME71204, MP71024: Item parts A, B, and C are combined to create a 1-point item, where 1 score point is awarded if all parts are correct
M04_03 – ME71036, MP71036: Item parts A and B are combined to create a 1-point item, where 1 score point is awarded if both parts are correct
M04_09 – ME71178, MP71178: Item parts A, B, and C are combined to create a 1-point item, where 1 score point is awarded if all parts are correct
M04_12 – ME71175, MP71175: Item parts A, B, and C are combined to create a 2-point item, where 2 score points are awarded if all parts are correct and 1 score point is awarded if 1 or 2 are correct
M06_01 – ME61018, MP61018: Item parts A, B, C, and D are combined to create a 1-point item, where 1 score point is awarded if all parts are correct
M06_10 – ME61266: Item parts A, B, C, D, E, and F are combined to create a 2-point item, where 2 score points are awarded if all parts are correct and 1 score point is awarded if 5 parts are correct
M08_11 – ME71141, M08_10 – MP71141: Item parts A, B, C, and D are combined to create a 1-point item, where 1 score point is awarded if all parts are correct
M08_12 – ME71194: Item parts A and B are combined to create a 1-point item, where 1 score point is awarded if both parts are correct
M08_13 – ME71193, M08_12 – MP71193: Item parts A and B are combined to create a 2-point item, where 2 score points are awarded if both are correct and 1 score point is awarded if 1 part is correct
M10_05 – ME71213: Item parts A, B, and C are combined to create a 1-point item, where 1 score point is awarded if all parts are correct
M10_08 – ME71179, MP71179: Item parts A, B, and C are combined to create a 1-point item, where 1 score point is awarded if all parts are correct
M10_12A – ME71187A: Item parts A, B, C, and D are combined to create a 1-point item, where 1 score point is awarded if all parts are correct

IEA TIMSS & PIRLS
International Study Center
Lynch School of Education
BOSTON COLLEGE

CHAPTER 10: REVIEWING ACHIEVEMENT ITEM STATISTICS
METHODS AND PROCEDURES: TIMSS 2019 TECHNICAL REPORT 10.65