International Online Journal of Educational Sciences, 2015, 7 (1), 29-44

iojes
www.iojes.net

International Online Journal of Educational Sciences

ISSN: 1309-2707

Construct Validity of TIMSS 2011 Mathematics Cognitive Domains for Turkish Students

Serkan Arıkan[1]

[1]*Muğla Sıtkı Koçman University, Faculty of Education, Department of Elementary Education, Muğla/Turkey*

ARTICLE INFO

Article History:
Received 14.06.2014
Received in revised form 23.07.2014
Accepted 12.10.2014
Available online 01.03.2015

ABSTRACT

This study investigated empirical evidence for construct validity of the TIMSS 2011 Mathematics test by using sample of fourth and eighth grade Turkish students (n=4091). The purpose of this study was to analyze construct validity of the TIMSS 2011 mathematics test by using Turkish students' data. A confirmatory factor analysis model was formed with using cognitive domain definitions for each mathematics items presented by TIMSS assessment framework. In this three-factor model, observed variables (items) were associated to three latent variables (mathematics cognitive domain), namely knowing, applying and reasoning. Fit indices of confirmatory factor analysis, conducted by LISREL, showed mixed results. For eighth grade booklets, the proposed model was supported by the data, however, for fourth grade booklets the proposed model was not supported completely. Besides, seven of the items were found to be correlated with other cognitive domains than with the defined one. Evaluation of these items showed that four of them were problematic. These results showed that for the eighth grade, the mathematics cognitive domain defined by TIMSS also works for the Turkish sample and that a scale score for each cognitive domain could be used confidently. However, for the fourth grade, a scale score for each cognitive domain should be used with caution for Turkish students.

© 2015 IOJES. All rights reserved

Keywords:
TIMSS 2011, construct validity, mathematics cognitive domains

Introduction

Student achievements in mathematics as measured by international comparison examinations such as TIMSS (Trends in International Mathematics and Science Study) have been one of the main research concerns all around the world. By using data collected in these exams, many analysis and conclusions were made, such as the evaluation of TIMSS results (Yayan, 2009); comparative analysis of questionnaires across years (Yıldırım et al., 2012); factors affecting mathematics and science achievement (Berberoglu et al., 2003; Akyüz & Berberoğlu, 2010; Demir & Kılıç, 2010; Ceylan & Berberoglu, 2007); comparison of characteristics of high achieving and low achieving countries (Incikabi et al., 2012; Ozturk & Ucar, 2010); relation of mathematics performance and motivation of test-takers (Eklöf, 2007); and opinions of students towards the Mathematics course (Bilican et al., 2011). Besides, policy makers would use these test results and implications drawn from these comparisons to improve their educational systems. However, the validity of these results could be undermined if these test results were not valid for some countries. Moreover, some concerns related to these international comparisons were provided. For instance, Eckert (2008) summarized that sampling, curriculum, common definitions and methodological issues took some criticisms. Therefore, it is necessary to provide evidence related to construct validity of these examinations for each participating country. Besides, construct validity has special importance when a test is used in an international

[1] Corresponding author's address: Muğla Sıtkı Koçman University, Faculty of Education, Kötekli, Muğla
Telephone: +902522115488
Fax:+902522111762
e-mail: serkanarikan@mu.edu.tr
DOI: http://dx.doi.org/10.15345/iojes.2015.01.003

© 2015 International Online Journal of Educational Sciences (IOJES) is a publication of Educational Researches and Publications Association (ERPA)

comparison. The reason for this special importance is that the main assumption of these assessments is the construct that is measured should be same for every participant (Baykal & Circi, 2010).

Information about TIMSS

TIMSS gives comparative achievement results for more than 60 countries. Many of these countries use TIMSS results to improve their educational systems. TIMSS started in 1995 and has been conducted in a four year cycle for both fourth and eighth grade students. A country that participates in each TIMSS cycle regularly gets information about its own developmental trends. In each administration, together with mathematics and science achievement data, detailed background information about students, teachers, school resources, quality of curriculum and instruction was collected by several questionnaires. In 2011, the fifth and most recent TIMSS was conducted (Mullis et al., 2009). TIMSS Mathematics and Science achievement test results give information about the effectiveness of a curriculum and instruction by using students' answers to test items. Therefore, countries can monitor their weaknesses and strengths in mathematics and science in an international context (Mullis et al., 2009).

Mullis et al. (2009) in TIMSS 2011 mathematics assessment framework explained important aspects of how mathematics is viewed in TIMSS. The researchers maintained that the ability to use mathematics is crucial in today's competitive world. The study also emphasized that learning mathematics for its own sake should not be the primary reason to teach mathematics. For the fourth and eighth grades, the TIMSS 2011 assessment framework emphasized two domains: content domain and cognitive domain. The content domain specifies covered subject matters, namely numbers, geometric shapes and measures and data display for the fourth grade, whereas numbers, algebra, geometry, and data and chance for the eighth grade. The cognitive domain specifies thinking processes that are measured, namely knowing, applying and reasoning for both fourth and eighth grades.

The knowing domain is related to facts, concepts and procedures that students are expected to know. TIMSS defined that the term facts refer to factual knowledge that is related to the basic language of mathematics and properties that build the mathematical thoughts. The term concepts refer to making connections using elements of knowledge. Procedures are defined as a bridge between basic knowledge and usage of mathematics and recalling sets of actions. Students are expected to be efficient and accurate in conducting computational procedures and tools in TIMSS. Knowing domain includes Recall, Recognize, Compute, Retrieve, Measure and Classify/Order sub-skills. In recall sub-skill, students are expected to recall definitions, terminology, number properties, geometric properties, and notations. In recognize sub-skill, students are expected to recognize mathematical objects (shapes, numbers, expressions, and quantities), mathematical entities that are mathematically equivalent. In compute sub-skill, students are expected to carry out algorithmic procedures for operations with whole numbers, fractions, decimals and integers; to approximate numbers to estimate computations; and to carry out routine algebraic procedures. In retrieve sub-skill, students are expected to retrieve information from graphs, tables, or other sources; to read simple scales. In measure sub-skill, students are expected to use measuring instruments; to choose appropriate units of measurement. In classify/order sub-skill, students are expected to classify objects, shapes, numbers and expressions according to common properties; to make correct decisions about class membership; to order numbers and objects by attributes (Mullis et al., 2009).

The applying domain is related to students' ability to apply knowledge and conceptual understanding to solve problems. In applying domain application of mathematical tools in a specific context is essential. Similarly, routine problem solving ability is a core component of applying domain. Applying domain includes Select, Represent, Model, Implement and Solve Routine Problems sub-skills. In select sub-skill, students are expected to select a proper operation, method or strategy to solve problems where there is a known procedure, algorithm, or method of solution. In represent sub-skill, students are expected to show mathematical information and data in diagrams, tables, charts, or graphs, and generate equivalent representations for a specific mathematical relationship. In model sub-skill, students are expected to generate an appropriate model, such as an equation, geometric figure, or diagram to solve a routine problem. In implement sub-skill, students are expected to implement given mathematical instructions. In solve routine problem sub-skill, students are expected to solve routine problems (in familiar contexts or purely mathematical) similar to problems engaged in class (Mullis et al., 2009).

The reasoning domain is related to encompassing unfamiliar situations, complex contexts and multistep problems. TIMSS defined that reasoning is related to having capacity of logical and systematic thinking, ability to conduct intuitive and inductive reasoning to reach solutions to non-routine problems. Non-routine problems are defined as problems that are generally unfamiliar to students. They require higher level of cognitive skills to reach to the solution than routine problems. Reasoning domain involves Analyze, Generalize/Specialize, Integrate/Synthesize, Justıfy, and Solve Non-routine Problems. In analyze sub-skill students are expected to determine, describe or use relationships between variables or objects in mathematical situations, and to make valid inferences from given information. In generalize/specialize sub-skill, students are expected to extend their mathematical thinking and problem solving by restating results in more general and widely applicable terms. In integrate/synthesize, students are expected to make connections between forms of knowledge and different representations of it, to produce links between related mathematical ideas, to combine mathematical facts, concepts, and procedures to produce results and to use results to produce another result (Mullis et al., 2009).

Item writing process of TIMSS 2011 was achieved by many rules that were described in TIMSS 2011 Item Writing Guidelines. Item writers worked in groups of two or three and these groups were assigned specific content areas. All items were written in English and collected at the end of each day. Groups were expected to produce 10 to 12 items per day. Item writers were given a detailed guideline which emphasized matching content domain and paying particular attention to cognitive domains when item was produced. For each item, item writers were expected to consider answers to the following questions: "what knowledge does this item allow a student to show?" and "what cognitive processes does this item require a student to demonstrate?" Therefore, content and cognitive domain information of items was provided by item writers. Besides, producing items according to the best suitable item format and producing half of the items in multiple-choice and half of the items in constructed response worth 1 or 2 score point was advised. Each constructed-response items were expected to be developed with a specific scoring guide. For both item formats, separate item review checklists were given to evaluate each item. In these item checklists, "whether or not content and cognitive classification is correct" was asked explicitly. Moreover, item writers were expected to consider timing, grade appropriateness, difficulty level, potential sources of bias and ease of translation. After all items were developed, items and related definitions were reviewed by TIMSS & PIRLS International Study Center staff, the mathematics and science coordinators and consultants, the Science and Mathematics Item Review Committee and the National Research Coordinators (Mullis & Martin, 2012).

In TIMSS, each content domain includes items related to the respective cognitive domain. In TIMSS 2011, for the fourth grade, 40% of the items were related to knowing, 40% of the items were related to applying and 20% of the items were related to reasoning. Similarly, for the eighth grade, 35% of the items were related to knowing, 40% of the items were related to applying and 25% of the items were related to reasoning. These content and cognitive domains were given importance within the TIMSS structure, therefore in TIMSS 2011 scale scores of these domains were separately reported for each student. TIMSS released many items after administrations were completed. When the released items were published, the content and cognitive domain information and explanation related to that item was also presented by TIMSS. In TIMSS 2011, items of four out of fourteen mathematics booklets were released for both grade levels (Mullis et al., 2009). In Appendix A, a sample of items representing each cognitive domain for 4th and 8th grade is presented.

Turkey's standings in TIMSS

The first TIMSS was administered in 1995 with 45 participating countries; however Turkey did not participate in that administration. The following TIMSS was administered in 1999 with 38 participating countries only for grade eight. Turkey found a position as 31st among these countries at the eighth grade. The next TIMSS was administered in 2003 with 53 participating countries. Turkey did not participate in that administration. The following TIMSS was administered in 2007 with 49 participating countries for eighth grades. Turkey was placed in the 30th position among these countries. Turkey attended the TIMSS fourth grade test for the first time in the 2011 administration. Turkey positioned itself as the 35th among 53 countries in grade four. In grade eight, Turkey scored 25th place among 46 countries (TIMSS & PIRLS International Study Center, 2013). Although several reform attempts in the Turkish educational system were conducted throughout years, Turkey was still scoring positions under the average.

Construct Validity

Messick (1989) defined validity as "an integrated evaluative judgment of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of interpretations and actions based on test scores" (p.13). Therefore, in order to use TIMSS scores appropriately, empirical evidence related to its validity should be presented. It is important that the content of TIMSS should be covered by students as part of their education. Each subtopic was defined by a list of objectives that were covered by participating countries. Therefore, the appropriateness of a content domain for students in a participating country was controlled by national centers. Besides, Turkish teachers claimed that the content domain of TIMSS 2011 was very close to the national curriculum (Yücel et al., 2013). This is related to providing content validity proof for TIMSS. Content validity is related to deciding whether or not the content of an instrument is a proper sample of the domain of the content (Fraenkel & Wallen, 2005). Having the content validity is not enough for correct interpretations of scores. The question of "What does an instrument really measure?" could be answered by providing construct validity evidence (Cronbach, 1971). Construct validation is defined as a "research process by which one goes about establishing construct validity; that is, the process of collecting evidence that a test or other operational measure does indeed reflect the theoretical construct" (Arvey, 1992). Construct validity evidence is related to statistical analysis to test whether or not a test measures what it intends to measure. Confirmatory factor analysis (CFA) is used in construct validation studies. CFA gives information about what is measured by a test and what the dimensions of a test are (Zumbo, 2005). The main purpose of CFA is to test whether or not the established measurement model is supported by the collected data (Kline, 1998; Sharma, 1996). Therefore, by using CFA, whether or not the cognitive domains defined in TIMSS mathematics framework is supported by student answers should be evaluated. This evaluation is related to providing construct validity evidence for TIMSS. Especially, providing this type of evidence for a country would be helpful to show that TIMSS provides fair results in terms of measured cognitive skills for that country. Specifically, this study will focus on providing construct validity evidence for the Turkish sample in the TIMSS 2011 mathematics test.

The literature review results indicated that there was not any research related to construct validity of mathematics tests in TIMSS 2011. Therefore, studies using cognitive domain sub-scores of TIMSS would have problems in their results. In order to enlighten on this issue, the purpose of this study was to analyze construct validity of the TIMSS 2011 mathematics test by using Turkish students' data at the fourth and eighth grades. In order to achieve this goal, the cognitive domains definitions of mathematics items were used. The research problem of this study was "Is the cognitive domain defined in the TIMSS 2011 mathematics assessment framework supported by Turkish students' data for the fourth and eighth grades?"

Method

Measures and Variables

The TIMSS 2011 study administered achievement tests and questionnaires in order to collect data. The present study used released fourth and eighth grade mathematics items. In TIMSS 2011, there were fourteen mathematics booklets for the fourth grade and fourteen mathematics booklets for the eighth grade. Each item was presented in two booklets which allowed linking student responses from different booklets. The booklets were distributed randomly among students. Among these fourteen booklets, items in four booklets were released as sample items (Mullis et al., 2009). In this study, these released four booklets for the fourth and eighth grades were used because item descriptions were necessary for analysis.

In TIMSS 2011 released items document, content domain and cognitive domain explanations about all released items were given (Foy, Arora & Stanco, 2013). As Table 1 shows, the number of items in each booklet varies from 21 to 27 for the fourth grade, and 26 to 33 for the eighth grade. It is worth mentioning that for each booklet, the numbers of items related to cognitive domains differ. For instance, for the eighth grade, the number of items related to reasoning cognitive domain ranges from 5 to 13.

Table 1. General information about TIMSS 2011 released booklets

Grade	Booklet	Number of Students	Number of Items	Number of Items related to *Knowing*	Number of Items related to *Applying*	Number of Items related to *Reasoning*
4	1	531	21	7	6	8
4	2	535	26	12	9	5
4	5	537	25	11	10	4
4	6	513	27	11	12	4
8	1	503	26	7	6	13
8	2	488	32	10	13	9
8	5	490	32	15	11	6
8	6	494	33	14	14	5

Sample

The data analyzed in this study were obtained from the TIMSS 2011 study from http://timssandpirls.bc.edu/timss2011/international-database.html. The target population of TIMSS is all students enrolled at the fourth and eighth grades. A country may join in both grades or choose one of them. The sampling procedure for TIMSS 2011 was planned by TIMSS & PIRLS International Study Center in cooperation with Statistics Canada and the IEA (The International Association for the Evaluation of Educational Achievement) Data Processing and Research Center. The implementation of this procedure was achieved by the participating country's National Research Coordinator and TIMSS sampling experts. In TIMSS, the two-stage random sample design was used. In the first stage, a sample of schools was selected; in the second stage, classes of students were selected among these schools. This procedure was used in order to reach classroom based conclusions such as students' curricular and instructional experiences. With using all of these procedures, it is achieved that the average age of fourth and eighth graders is at least 9,5 and 13.5 years respectively (Joncas and Foy, 2012). The sample for the present study comprised all fourth and eighth grade Turkish students who took booklet one, two, five and six in the TIMSS 2011 study since the item descriptions were released only in these booklets. The 4th grade sample consisted of 2116 students (49.3% girl and 50.7% boy) with average age of 10.07. The 8th grade sample consisted of 1975 students (50.1% girl and 49.9% boy) with average age of 14.06. As Table 1 shows, on average 529 students answered fourth grade booklets whereas on average 494 students answered eighth grade booklets.

The Procedure for Data Analysis

The current study utilized confirmatory factor analysis to examine construct validity of TIMSS 2011 mathematics tests for Turkish students. By using the cognitive domain definitions presented by TIMSS, confirmatory factor analysis models for each released booklet were established. Specifically, in these models, items that were associated with three latent variables (knowing, applying and reasoning cognitive domain) were tested separately for each booklet.

When construct validity investigation using confirmatory factor analysis is actualized, LISREL can be used (Jöreskog & Sörbom, 1999; Byrne, 1998; Schumacker & Lomax, 2004). Therefore, in the present study, for the purpose of conducting confirmatory factor analysis LISREL was chosen. In structural equation models like confirmatory factor analysis, generally the Maximum Likelihood (ML) estimation method is used. However, Jöreskog (2004) claimed that ordinal variables should be analyzed with other estimation techniques, such as Weighted Least Square (WLS). It is also added that ordinal variables are not continuous therefore they should not be included in the analysis as they are. One should be aware of that as ordinal variables do not have origins or unit of measurement, means, variances, and covariances of ordinal variables make no sense. The only information that could be activated is number of cases in each cell of a multiway contingency table. Therefore, conducting confirmatory factor analysis with ordinal data requires some caution. The ML method requires a sample covariance matrix with a multivariate normal distribution.

However, this is not possible for ordinal data. The remedy is to use WLS method based on polychoric correlation with asymptotic covariance matrix to get correct standard errors and chi-squares. In the current study, all observed variables were converted to ordinal variables with either 0 (incorrect response) and 1 (correct response) values or 0 (incorrect response), 1 (partially correct response) and 2 (correct response) values related to maximum value of an item. Therefore, the WLS method was used for LISREL analyses as advised by Jöreskog.

As LISREL needs covariance matrixes and asymptotic covariance matrixes to run WLS estimation method, these matrices were generated by using the PRELIS program. Confirmatory factor analysis requires an assessment to establish whether or not the proposed model is a good one. A good model is a model in which the difference between covariance matrix obtained from student data and covariance matrix implied by the hypothesized model is minimum (Ullman, 2001). This difference is evaluated by using several fit indices. Sun (2005) advised to report Standardized Root Mean Square Residual (SRMR), Non-Normed Fit Index (NNFI), Comparative Fit Index (CFI) and Root Mean Square Error of Approximation (RMSEA) fit indices to assess goodness of fit of confirmatory factor analysis in construct validation research. In this study, criteria for a good model were determined as values below 0.08 for SRMR (Hu & Bentler, 1999), 0.95 or higher for NNFI (Hu & Bentler, 1999), 0.95 or higher for CFI (Ullman, 2001), and values below 0.05 for RMSEA (Byrne, 1998; Hu & Bentler, 1999; Ullman, 2001). In addition to these fit indices, frequently used fit indices of Goodness of Fit Index (GFI) and Adjusted Goodness of Fit Index (AGFI) also reported to get wider evaluation. For these indices, criteria for a good model were values above 0.95 for GFI (Schumacker & Lomax, 2004), and values above 0.90 for AGFI (McDonald & Moon-Ho, 2002). Sun (2005) reported that SRMR and RMSEA have a minimum value of zero; CFI, GFI and AGFI have upper bound of 1. It is also added that although NNFI gets value between "0" to "1", it is not always constrained to zero-to-1 range.

Results

The Cronbach's alpha reliability coefficients calculated by SPSS 21 for fourth and eighth grade booklets were given in Table 2. These values indicated that fourth grade booklets had "good" reliability as Cronbach's alpha reliability coefficients of these booklets were between 0.800 and 0.900, and eighth grade booklets had "excellent" reliability as Cronbach's alpha reliability coefficients of these booklets were above 0.900 (George & Mallery, 2001).

Figure 1 shows a conceptual diagram of the confirmatory factor model proposed by mathematics cognitive domain definitions in TIMSS 2011. As seen in this figure, each item was proposed to be associated with its own cognitive domain. It was expected that this proposed model would be supported by student data. Moreover, it was necessary to investigate whether or not any item was also correlated with another cognitive domain by evaluating suggested modification indices and cross-loadings values. In the analysis, error terms were freed to correlate.

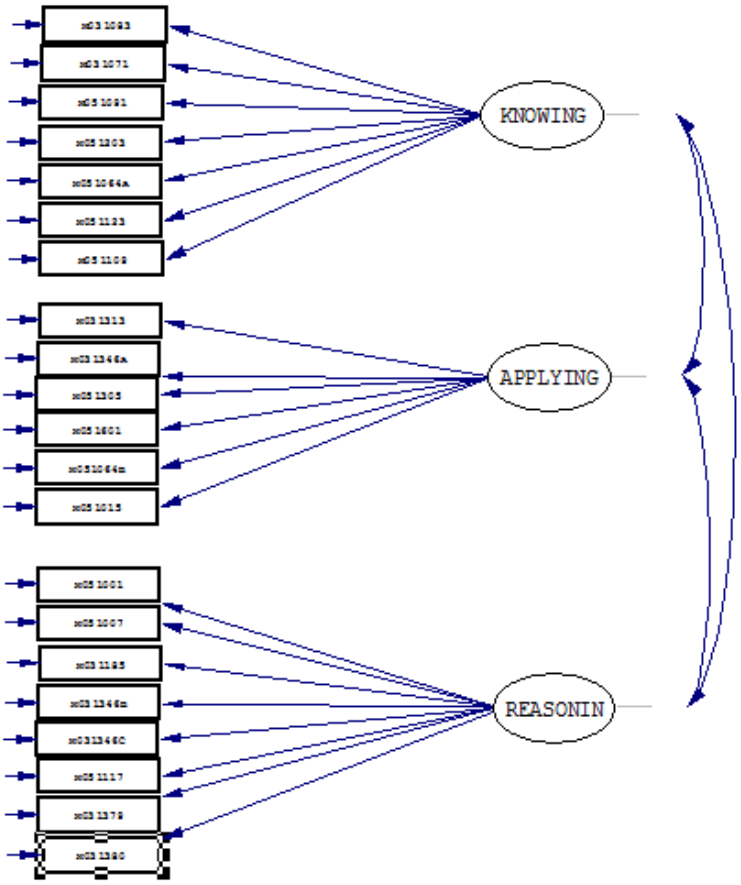


Figure1. Conceptual diagram for confirmatory factor analysis of TIMSS 2011 Turkish sample

Confirmatory factor analyses results for fourth grade booklets by using LISREL 8.70 were presented in Table 2. The results showed that SRMR, RMSEA, NNFI and CFI values indicated a good fit of the data to the model. However, GFI and AGFI values for all fourth grade booklets were below the expected values of 0.95 and 0.90 respectively. These mixed results required to investigate possible problems. Therefore, modification indices suggested by LISREL were investigated. For booklet 1, item M031346A was also suggested to be associated with knowing and reasoning cognitive domains; item M031346C was also suggested to be associated with the applying cognitive domain; item M051601 was also suggested to be associated with the knowing cognitive domain; and item M051064B was also suggested to be associated with the knowing cognitive domain. For booklet 5, item M031183 was also suggested to be associated with the reasoning cognitive domain and item M041160A was also suggested to be associated with the applying cognitive domain. For booklet 6, item M041160A was also suggested to be associated with the applying cognitive domain and item M031079B was also suggested to be associated with the reasoning cognitive domain. As modification indices suggested that there might be mismatch among some items and their cognitive domain definitions, cross-loadings of these items with suggested cognitive domains were calculated and provided in Table 3. Among these seven items (M041160A was problematic for two different booklets), four of them had higher cross-loading values with CFA suggested cognitive domains than TIMSS suggested cognitive domains. In Appendix B, items that exhibited problems are presented.

According to the LISREL 8.70 analysis results given in Table 2 for the eighth grade, GFI, AGFI, SRMR, RMSEA, NNFI and CFI values were within acceptable ranges. Thus, these results indicated that the proposed model was supported by the collected eighth grade data as all indices showed a very good fit. Besides that, for all booklets, there were not any modification indices suggestions. Therefore, it was concluded that the cognitive domain defined in the TIMSS eighth grade was valid for Turkish students' data.

Table 2. Cronbach's alpha reliability coefficient and confirmatory factor analysis results for TIMSS 2011 Turkish sample

Grade	Booklet	Alpha	GFI	AGFI	SRMR	RMSEA	NNFI	CFI
4	1	0.846	0.74	0.68	0.079	0.044	0.99	0.99
4	2	0.860	0.80	0.77	0.060	0.024	1.00	1.00
4	5	0.874	0.78	0.73	0.060	0.028	0.99	1.00
4	6	0.883	0.75	0.70	0.066	0.034	0.99	0.99
8	1	0.909	0.95	0.94	0.033	0.048	0.96	0.96
8	2	0.919	0.95	0.94	0.032	0.040	0.97	0.97
8	5	0.916	0.94	0.93	0.035	0.000	1.04	1.00
8	6	0.913	0.92	0.91	0.038	0.000	1.04	1.00

Table 3. Cross-loadings of problematic items with suggested cognitive domain

Grade	Booklet	Item	Knowing	Applying	Reasoning
4	1	M031346A	1.24**	-1.61*	1.08**
4	1	M031346C		-0.64**	1.77*
4	1	M051601	1.51**	-0.79*	
4	1	M051064B	0.48**	0.30*	
4	5	M031183		-1.84*	2.30**
4	5	M041160A	1.40*	-0.61**	
4	6	M041160A	1.19*	-0.41**	
4	6	M031079B		1.48*	-0.65**

* Suggested by TIMSS item developers

** Suggested by CFA analysis results

Conclusion and Discussion

In the present study, the main aim was to analyze whether the cognitive domain defined in TIMSS 2011 mathematics assessment framework was supported by the fourth and eighth grade Turkish students' data. In order to achieve this aim, a three-factor confirmatory model was formed for eight released booklets. In the confirmatory factor analysis model, each item was associated with its defined cognitive domain, and the fit was tested. The results indicated that the overall structure was supported by Turkish students' data and none of the items was more correlated with another cognitive domain for eight graders. These findings imply that the cognitive structure of TIMSS 2011 was also valid for eighth grade Turkish students. However, for the fourth grade Turkish students' data mixed results were obtained. SRMR, RMSEA, NNFI and CFI values were within acceptable range, whereas GFI and AGFI values were below the expected values which implied the proposed model was not adequately supported. Especially, seven items in fourth grade booklets were suggested to be associated with other cognitive domains than the defined cognitive domains. Evaluation of these modifications showed that four of these suggestions produced higher cross-loading values.

Modification indices suggestions and cross-loading values implied that items M031346A, M051601, M051064B, and M031183 were problematic and malfunctioned for Turkish sample (Appendix B). Evaluation of these items in terms of cognitive domain definitions would make it possible to detect problems. For the item M031346A, TIMSS item developers claimed that it measures applying cognitive skill whereas analysis results suggested knowing and reasoning cognitive skills. In this item, trading cards according to given rule

is asked. There are four more items (M031346B, M031346C, M031379, and M031380) similar to this item in which their cognitive domains were claimed as reasoning by TIMSS. All of these items are related to content of proportion. As proportion topic is not covered by Turkish students at the fourth grade (MONE, 2009), item M031346A should be considered as non-routine problem which is associated with reasoning cognitive domain (Mullis et al., 2009). Another item considered to be problematic is M051601. This item is claimed to be related to applying cognitive skill, whereas CFA results indicated it as knowing cognitive skill. When this item is investigated, it is seen that the next figure in a pattern is asked. The patterns concept is given special importance in Turkish curriculum. It starts from the first grade and expands in a spiral structure. Especially at the third grade, expansion of patterns and the next figures in patterns are one of the objectives of curriculum (MONE, 2009). Turkish students are familiar with these basic patterns. Therefore, the cognitive process in this item is related to recalling the pattern which is associated with knowing cognitive domain (Mullis et al., 2009). For the item M051064B, TIMSS reported that this item measures applying cognitive skill, however CFA results indicated knowing cognitive skill. In this item, locating a point according to a given example is asked. There is one other item (M051064A), which is very similar to this item. However, that item is reported to measure knowing cognitive skill by TIMSS. According to TIMSS 2011 Assessment Framework, item M051064B is related to retrieve information which is associated with knowing cognitive domain (Mullis et al., 2009). The last item considered to be problematic is M031183. This item is claimed to be related to applying cognitive skill, whereas CFA results indicated reasoning cognitive skill. In this item, dividing a rational number into two is asked. As this topic is not covered at 4th grade in Turkish educational system (MONE, 2009), Turkish students could only reach solution to this non-routine problem by activating their intuitive reasoning ability (Mullis et al., 2009). Therefore, these four items were decided to be problematic not only by statistical analysis, but also according to evaluation of Turkish educational program and TIMSS Assessment framework. These findings imply that the TIMSS 2011 fourth grade booklets might not be adequately valid for Turkish students in terms of the defined cognitive domain as it was for eighth grade booklets. Therefore, separately reported cognitive domain scores for fourth grade students should be used with caution.

The construct validity evidence for the eighth grade implied that the TIMSS 2011 mathematics assessment was fair for Turkish students in terms of defined cognitive domain. Items in eighth grade booklets adequately represent knowing, applying and reasoning cognitive domains. Therefore, separately estimated subscores of each cognitive domain could be used to conduct research inquiring reasons for low achievement of Turkish students confidently. Prospective detailed analysis by using these subscores is expected to produce reliable and valid results.

The general structure of items at the fourth grade was not adequately supported in the Turkish context. It is considered that "construct-irrelevant variance" (Hubley & Zumbo, 2011) was occurred for fourth grade booklets. The main reason was considered to be due to some items that were associated with more than one cognitive domain. It would require another research topic on whether or not these items also contain DIF. In addition, there is a need to investigate if this low fit occurred only for the Turkish data or if it is general for all TIMSS 2011 fourth grade data. Although the actual reason could be investigated, it is important to note that the result obtained for the fourth grade implied that separately estimated subscores of each cognitive domain could not be used for Turkish students confidently. Any study that would use the knowing, applying and reasoning subscores would be biased for fourth grade Turkish students. Besides that Cronbach's alpha reliability coefficients for fourth grade booklets were lower than eighth grade booklets. This finding implies that some of the fourth grade items could have lower item-total correlation and that students would have difficulty understanding these items.

The findings of this study propose important implications for developers of TIMSS. Although detailed information and many results were given about each cognitive domain measured in TIMSS (Mullis et al., 2009 & Mullis et al., 2012), a statistical explanation about construct validity for each country was not present. Similarly, National Turkish TIMSS 2011 reports (Yücel et al., 2013) gave explanations about the results related to these cognitive domains, however, no information about validity of these cognitive domains was given. Whether or not the defined cognitive domain structure was supported by data should also be reported in TIMSS technical reports. This study is expected to be a starting point in validating these cognitive domains.

These findings are very important for Turkish policy makers. They should place more emphasis on the scientific reasons of low achievement. Confirmed structure of TIMSS mathematics assessment at the eighth grade implies that when constructing the mathematics curriculum, instructions on how cognitive domains would be taught should be clearly defined in order to be more successful in TIMSS and other international examinations. Turkey had an average scale score of 452 for total: 441 for knowing, 459 for applying and 465 for reasoning at the eighth grade (Mullis et al., 2012). Having the lowest score on the knowing cognitive domain implies that Turkish students had problems even with basic mathematical knowledge. Therefore, placing an importance on not only the content domain but also the cognitive domain starting from knowing to higher order thinking skills like applying and reasoning would increase student achievement. Similarly, Olkun and Aydoğdu (2003) proposed, related to low TIMSS 1999 mathematics achievement, that teaching mathematics content dimension like geometry with only memorization techniques without using higher order thinking skills might be one of the reason of low achievement of Turkish students.

The results of this study could be used for shaping teaching practices. As Yücel et al. (2013) explained, Turkish teachers claimed that the content domain of TIMSS 2011 was congruent with the Turkish national curriculum. Turkish curriculum got the 11th position in this congruence at the fourth grade and 3rd at the eighth grade. Therefore, as Turkey had a similar curriculum with TIMSS, the content domain probably was not the reason of low achievement. Teachers should place more emphasis on cognitive domains when they teach mathematics. After transferring basic knowledge and facts, activities related to routine and non-routine problems should be used more frequently.

References

Akyüz, G. & Berberoğlu, G. (2010). Teacher and classroom characteristics and their relations to mathematics achievement of the students in the TIMSS. *New Horizons in Education, 58*(1), 77-95.

Arvey, R. D. (1992). Constructs and construct validation: Definitions and issues. *Human Performance*, *5*(1&2), 59-69.http://dx.doi.org/10.1080/08959285.1992.9667924

Baykal, A. & Circi, R. (2010). Item revision to improve construct validity: A study on released science items in Turkish PISA 2006. *Procedia Social and Behavioral Sciences*, 2, 1931-1935.http://dx.doi.org/10.1016/j.sbspro.2010.03.1012

Berberoglu, G., Celebi, O., Ozdemir, E., Uysal, E., Yayan, B. (2003). Factors effecting achievement level of Turkish students in the third international mathematics and science study (TIMSS). *Eğitim Bilimleri ve Uygulama, 2* (3), 3-14.

Bilican, S., Demirtaşlı, R. N., Kilmen, S. (2011). The attitudes and opinions of the students towards mathematics course: The comparison of TIMSS 1999 and TIMSS 2007. *Educational Sciences: Theory & Practice*, 11(3), 1277-1283.

Byrne, B. M. (1998). *Structural equation modeling with LISREL, PRELIS, and SIMPLIS: Basic concepts, applications, and programming.* New Jersey: Lawrence Erlbaum Associates, Inc.

Ceylan, E. & Berberoğlu, G. (2007). Factors related with students' science achievement: A modeling study. *Education and Science*, 32 (144), 36-48.

Cronbach, L. J. (1971). Test validation. In R. L. Thorndike (Ed.), *Educational measurement* (2nd ed., pp. 443-507). Washington, DC: American Council on Education.

Demir, İ. & Kılıç, S. (2010). Using PISA 2003, examining the factors affecting students' mathematics achievement. *Hacettepe Üniversitesi Eğitim Fakültesi Dergisi (H. U. Journal of Education),* 38, 44-54.

Eckert, J. M. (2008). Trends in mathematics and science study (TIMSS) international accountability and implications for science instruction. *Research in Comparative and International Education,* 3 (2), 202-210.http://dx.doi.org/10.2304/rcie.2008.3.2.202

Eklöf, H. (2007). Test-taking motivation and mathematics performance in TIMSS 2003. *International Journal of Testing*, 7(3), 311-326.http://dx.doi.org/10.1080/15305050701438074

Foy, P., Arora, A. and Stanco, G. M. (2013). *TIMSS 2011 user guide for the international database*. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.

Fraenkel, J. R. & Wallen, N. E. (2005). *How to design and evaluate research in education.* New York: McGraw-Hill.

George, D., & Mallery, P. (2003). SPSS *for Windows step by step: A simple guide and reference. 11.0 update* (4th ed.). Boston: Allyn & Bacon.

Hu, L.-T. & Bentler, P.M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: conventional criteria versus new alternatives, *Structural Equation Modeling,* 6, 1–55. http://dx.doi.org/10.1080/10705519909540118

Hubley, A. M., & Zumbo, B. D. (2011). Validity and the consequences of test interpretation and use. *Social Indicators Research*, *103*(2), 219-230.http://dx.doi.org/10.1007/s11205-011-9843-4

Incikabi, L., Ozgelen S., Hartono, T. (2012). A comparative analysis of numbers and biology content domains between Turkey and the USA. *International Journal of Environmental & Science Education.* 7 (4), 523-536.

Joncas, M. & Foy, P. (2012). Sample design in TIMSS and PIRLS. Retrieved January 10, 2014, from http://pirls.bc.edu/methods/t-sample-design.html.

Jöreskog, K. G., & Sörbom, D. (1999). *LISREL 8.30.* Chicago: Scientific Software International.

Jöreskog, K. G. (2004). *Structural Equation modeling with ordinal variables using LISREL.* Retrieved June 30, 2013, from http://202.129.48.214/CD4007/LISREL88S/PDFfiles/ordinal.pdf.

Kline, R.B. (1998). *Principles and practice of structural equation modeling*, New York: The Guilford Press.

McDonald, R. P., & Moon-Ho, R. H. (2002). Principles and practice in reporting structural equation analyses. *Psychological Methods, 7*(1), 64-82.http://dx.doi.org/10.1037//1082-989X.7.1.64

Messick, S. (1989). Validity. In R. L. Linn (Ed.), Educational measurement (pp. 13–103). New York, NY: Macmillan

MONE (2009). *İlköğretim matematik dersi 1-5. sınıflar öğretim programı*. Ankara: Talim ve Terbiye Kurulu Başkanlığı.

Mullis, I. V., Martin, M. O., Ruddock, G. J., O'Sullivan, C. Y., & Preuschoff, C. (2009). *TIMSS 2011 assessment frameworks*. Amsterdam: International Association for the Evaluation of Educational Achievement.

Mullis, I. V. & Martin, M. O. (2012). *TIMSS 2011 Item writing guidelines.* Retrieved January 10, 2014, from http://pirls.bc.edu/methods/pdf/T11_Item_writing_guidelines.pdf.

Mullis, I. V., Martin, M. O., Foy, P. Arora, A. (2012). *TIMSS 2011 International Results in Mathematics*. Amsterdam: International Association for the Evaluation of Educational Achievement.

Olkun, S. & Aydoğdu, T. (2003). Üçüncü uluslararası matematik ve fen araştırması (TIMSS) nedir? Neyi sorgular? Örnek geometri soruları ve etkinlikler. *İlköğretim Online*, 2 (1), 28-35.

Ozturk, D., & Ucar, S. (2010). TIMSS verileri kullanarak Tayvan ve Türkiye' deki 8.sınıf öğrencilerinin fen başarısına etki eden faktörlerin belirlenmesi ve karşılaştırılması. *Cukurova University Journal of Social Sciences, 19*(3), 241-256.

Schumacker, R. E. & Lomax, R. G. (2004). *A beginners guide to structural equation modeling*. New Jersey: Lawrence Erlbaum Associates, Inc.

Sharma, S. (1996). *Applied multivariate techniques*. New York: J. Wiley.

Sun, J. (2005). Assessing goodness of fit in confirmatory factor analysis. *Measurement and Evaluation in Counseling and Development*, 37, 240-256.

TIMSS & PIRLS International Study Center (2013). Retrieved July 31, 2014, from http://timss.bc.edu/

Ullman, J. B. (2001). Structural equation modeling. In B. Tabachnick & L. S. Fidell (Eds.), *Using multivariate statistics* (4th ed., pp.653-771). Boston: Allyn & Bacon.

Yayan, B. (2009). Uluslararası matematik ve fen çalışması (TIMSS 2007) ve Türk öğrencilerin TIMSS 2007'deki matematik performanslarının değerlendirilmesi. *Cito Eğitim: Kuram ve Uygulama Dergisi*, 3, 39-52.

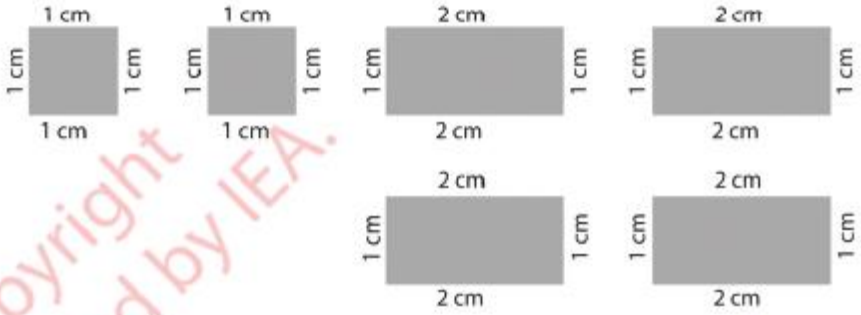
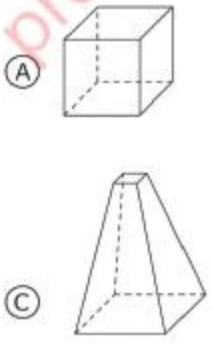
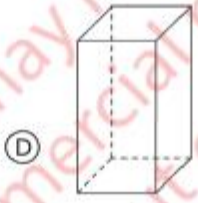
Yıldırım, Ö., Demirtaşlı, Ç. N., Akbaş, U. (2012). The opinions of mathematics teachers on homework and ın-class assessment: TIMSS 1999 and TIMSS 2007 periods. *Eğitim ve Bilim*, 37 (163), 126-142.

Yücel, C., Karadağ, E., & Turan, S. (2013). *TIMSS 2011 ulusal ön değerlendirme raporu*. (Eğitimde Politika Analizi Raporlar Serisi I). Eskişehir: Osmangazi Üniversitesi Eğitim Fakültesi.

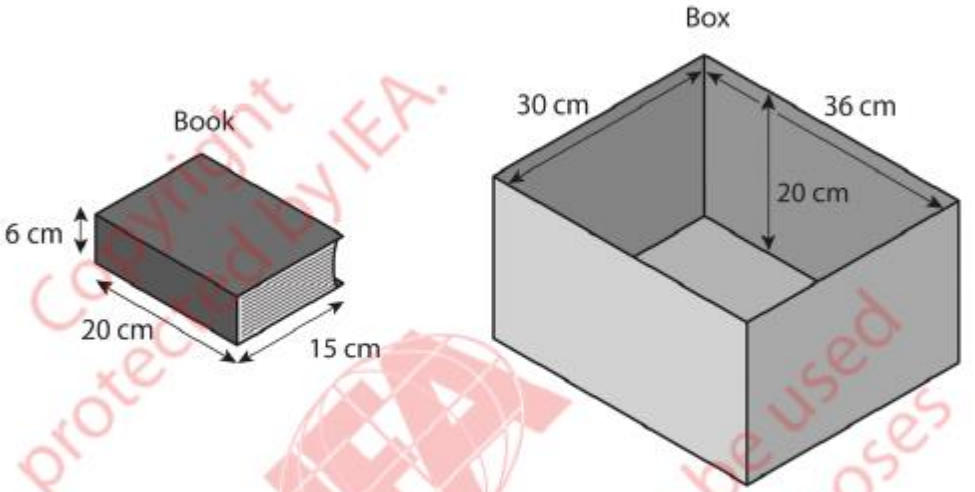
Zumbo, B. D. (2005). Structural equation modeling and test validation. In B. S. Everitt & D. C. Howell (Eds), *Encyclopedia of statistics in behavioral science* (pp. 1951-1958). Hoboken, N.J.: John Wiley & Sons.

Appendix A.

TIMSS 2011 Sample Released Items

Sample Item 1: M031083 Grade: 4th Cognitive Domain: Knowing	 Susan has the 6 pieces of cardboard shown above. Which of the following shapes could Susan make using all 6 of these pieces without cutting them?   
Sample Item 2: M051305 Grade: 4th Cognitive Domain: Applying	Duncan first traveled 4.8 km in a car and then he traveled 1.5 km in a bus. How far did Duncan travel? Ⓐ 6.3 km Ⓑ 5.8 km Ⓒ 5.13 km Ⓓ 4.95 km
Sample Item 3: M051001 Grade: 4th Cognitive Domain: Reasoning	In a soccer tournament, teams get: 3 points for a win 1 point for a tie 0 points for a loss Zedland has 11 points. What is the **smallest** number of games Zedland could have played? Answer: _____

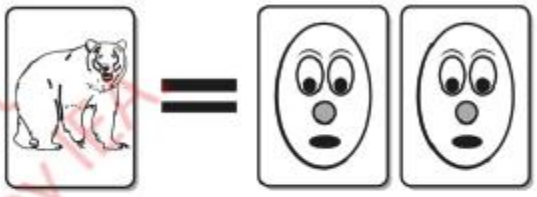
Copyright protected by IEA. This item may not be used for commercial purposes without express permission from IEA.

Sample Item 4: M052216 Grade: 8th Cognitive Domain: Knowing	Which number is equal to $\frac{3}{5}$? Ⓐ 0.8 Ⓑ 0.6 Ⓒ 0.53 Ⓓ 0.35
Sample Item 5: M052061 Grade: 8th Cognitive Domain: Applying	Kim is packing eggs into boxes. Each box holds 6 eggs. She has 94 eggs. What is the smallest number of boxes she needs to pack all the eggs? Answer: _____ boxes
Sample Item 6: M052206 Grade: 8th Cognitive Domain: Reasoning	Ryan is packing books into a rectangular box. All the books are the same size.  What is the largest number of books that will fit inside the box? Answer: _____

Copyright protected by IEA. May not be used for commercial purposes without permission of IEA

Appendix B.

TIMSS 2011 Problematic Items

Problematic Item 1: M031346A Grade: 4th Cognitive Domain: Applying According to CFA: Reasoning	The town fair had a booth where people could trade cards.  1 animal card is worth 2 cartoon cards.  2 animal cards are worth 3 sports cards. Some children went to the booth to trade cards. A. Becky had 5 animal cards to trade for cartoon cards. How many cartoon cards would she get? Answer: _____ cartoon cards
Problematic Item 2: M051601 Grade: 4th Cognitive Domain: Applying According to CFA: Knowing	Cooney has to form figures 1 to 4 with matches. Figures 1, 2, and 3 are shown below. He needs four matches to form figure 1, seven matches to form figure 2, and ten matches to form figure 3. He uses the same rule each time to make the next figure in the pattern.  1 2 3 How many matches will he need to form figure 4? Answer: _____

Copyright protected by IEA. May not be used for commercial purposes without express permission.

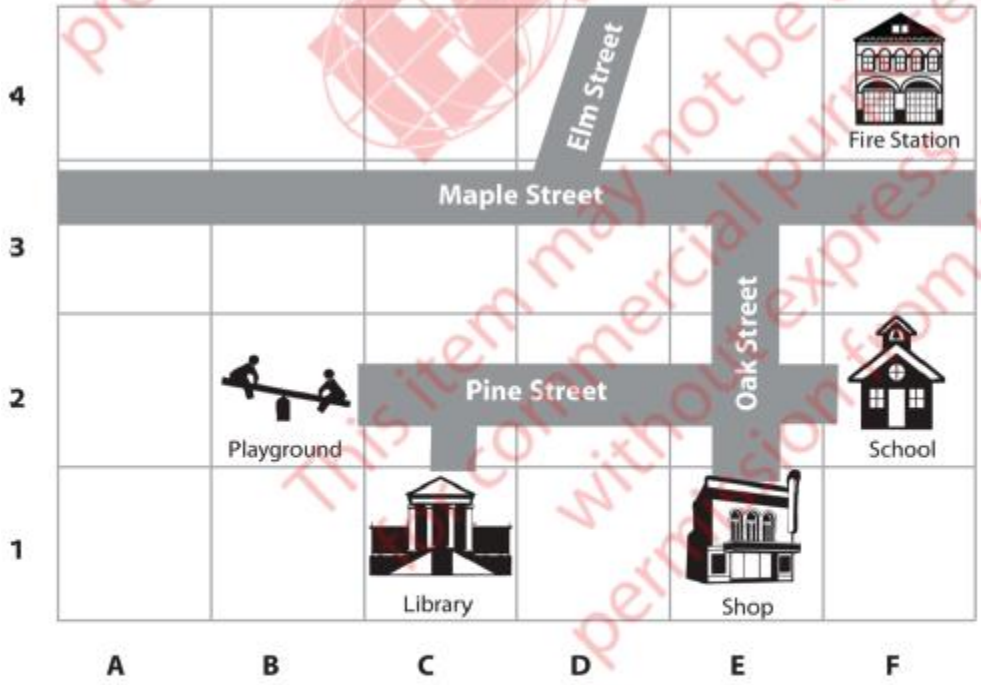
Problematic Item 3:

M051064B

Grade: 4th

Cognitive Domain: Applying

According to CFA: Knowing



B. Troy lives in a house in square C4. Put an X in the square to show where Troy lives.

Problematic Item 4:

M031183

Grade: 4th

Cognitive Domain: Applying

According to CFA: Reasoning

Ingredients	
Eggs	4
Flour	8 cups
Milk	$\frac{1}{2}$ cup

The above ingredients are used to make a recipe for 6 people. Sam wants to make this recipe for only 3 people.

Complete the table below to show what Sam needs to make the recipe for 3 people. The number of eggs he needs is shown.

Ingredients	
Eggs	2
Flour	___ cups
Milk	___ cups

Copyright protected by IEA.