

Factors Contributing to Mathematics Achievement Differences of Turkish and Australian Students in TIMSS 2007 and 2011

Serkan Arikan

Mugla Sitki Kocman University, TURKEY

Fons J. R. van de Vijver & Kutlay Yagmur

Tilburg University, the NETHERLANDS

•Received 26 June 2015•Revised 7 November 2015 •Accepted 21 January 2016

Large-scale studies, such as the Trends in International Mathematics and Science Study (TIMSS), provide data to understand cross-national differences and similarities. In this study, we aimed to identify factors predicting mathematics achievement of Turkish students by comparing to Australian students. First, construct equivalence and item bias were evaluated to check the comparability. Second, factors predicting mathematics achievement of Turkish and Australian students were identified. Then, propensity score matching on background variables was conducted to identify the remaining achievement differences. Results indicated that mathematics skills were free of construct bias in these groups. After removal of some biased items, we obtained an item bias free booklet. Additionally, students' self-confidence and educational resources at home were significant predictors of achievement. Propensity score analysis indicated that educational resources and to a somewhat lesser extent self-confidence were effective in explaining achievement differences between these two countries.

Keywords: DIF, mathematics achievement, measurement invariance, propensity score, TIMSS

INTRODUCTION

In this study we aimed to identify factors predicting mathematics achievement of Turkish students by comparing the TIMSS achievement of Turkish and Australian students using powerful and novel statistical techniques such as differential item functioning [DIF], measurement invariance and propensity score procedures. Comparing countries according to their mathematics achievement levels and

Correspondence: Serkan Arikan,
Department of Elementary Education, Mugla Sitki Kocman University, Kötekli, Muğla,
Turkey.
E-mail: serkanarikan@mu.edu.tr

Copyright © 2016 by the author/s; licensee iSER, Ankara, TURKEY. This is an open access article distributed under the terms of the Creative Commons Attribution License (CC BY 4.0) (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original paper is accurately cited.

providing clues that are important in predicting achievement has become an important research topic using data from international assessment programs such as Trends in International Mathematics and Science Study [TIMSS]. These large-scale studies provide data to understand cross-national differences and similarities. Turkish students, on average, get scores below the TIMSS global mean score. For instance, the Turkish mean mathematics scores of 432 in TIMSS 2007 and 452 in TIMSS 2011 were well below the global average score of 500 (Mullis, Martin, & Foy, 2008; Mullis et al., 2012). By comparing the achievement of Turkey with a higher achieving country, it is expected to identify the factors behind the fairly low Turkish scores and to provide policy-relevant information about how to remedy this situation. We chose Australia as a reference country because TIMSS items were originally prepared in English and Australia is one of the English speaking countries that got higher achievement scores than Turkey in TIMSS 2007 and TIMSS 2011; the mean scores for Australia were 496 and 505, respectively.

Cross-cultural comparability

In making cross-cultural comparisons a number of methodological issues need to be taken into consideration. It is important to establish to what extent observed cross-cultural differences in scores refer to mathematics differences and to what extent these are influenced by nuisance factors. These nuisance factors are collectively known as bias. Construct, item, and method bias are three types of bias that can challenge cross-cultural comparability (Van de Vijver & Poortinga, 1997; Van de Vijver & Tanzer, 2004). Construct bias occurs when a target construct is not identical for students of different groups or nations. For instance, when measuring practical intelligence, the main focus is on cognitive performance, such as reasoning or previously acquired knowledge in western countries (Van de Vijver & Hambleton, 1996). However, asking what you would do when your house is on fire might have different answers for a student in the USA and in Zambia (Serpell, 1993). Therefore, the context of the question is important and might not be generalized to every culture. Multi-group confirmatory factor analysis methods, one of the procedures in structural equation modeling (SEM), are often used to show measurement invariance of constructs (absence of construct bias) measured in a test (Wu, Li, & Zumbo, 2007). Item bias, also called differential item functioning, occurs when students of different groups/nations with the same level of ability on the underlying construct show a different mean score on the item (Van de Vijver & Leung, 1997). For instance, in an international assessment, if an arithmetic item involves the mental manipulation of money given in Turkish Lira, it would be more realistic and easier to do the computations for Turkish students than other students who use another currency. Differential familiarity with the stimuli can create cross-cultural differences in scores that are undesirable as such familiarity differences do

State of the literature

- In making cross-cultural comparisons a number of methodological issues need to be taken into consideration as studies that compare multiple language forms of large-scale assessments, differentially functioning items are found, which threatens the comparability.
- Turkish students show scores below the average score in large-scale mathematics tests, such as TIMSS and PISA.
- In many achievement models, three general groups of factors are defined to influence student academic achievement, namely student aptitude variables, environmental variables, and instructional variables.

Contribution of this paper to the literature

- We evaluated construct and item bias of the cognitive domain structure of TIMSS. This study is the first to evaluate measurement invariance of cognitive domains of TIMSS.
- We used a propensity score approach to identify effectiveness of significant factors that predict mathematics achievement. In this novel procedure, students are first completely matched on background variables and then countries are compared.
- Our study suggests that Turkish students experience more problems with applying their knowledge in both routine and non-routine problem situations than Australian students.

not reflect country differences in mathematics, which is the target of the instrument. Using statistical methods to detect items showing DIF and removing these items from the instrument is required for item bias free and valid score comparisons. Otherwise, bias could create differences in observed scores which are not related to differences in the underlying trait or ability (He & Van de Vijver, 2013). Finally, method bias occurs when samples are not comparable or when familiarity of students to the test is different or when there are some problems related to the administration of the tests (Van de Vijver & Tanzer, 2004).

TIMSS assessment framework

In the TIMSS 2007 mathematics assessment framework, Mullis et al. (2005) explained how mathematics is conceptualized in TIMSS. For the fourth and eighth grades, TIMSS mathematics assessment emphasized two domains: content domain and cognitive domain. The content domain of TIMSS assessment is related to numbers, geometric shapes and measures, and data display for the fourth grade, and to numbers, algebra, geometry, and data and chance for the eighth grade. The cognitive domain of TIMSS assessment is related to the 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; applying is related to students' ability to apply knowledge and conceptual understanding to solve problems; finally, reasoning is related to dealing with unfamiliar situations, complex contexts and multistep problems. The three domains, knowing, applying and reasoning, have increasing levels of cognitive complexity. Measuring mathematics in such a structure can be expected to give detailed feedback about mathematics outcomes of countries.

Turkish and Australian mathematics curriculum in 2007

The Turkish mathematics curriculum for primary and secondary schools were fundamentally revised in 2005. The Turkish curriculum was reformed to emphasize process-based outcomes (Koc, Isiksal, & Bulut, 2007). Before the revision, teacher-centered pedagogical approach and rote-learning were the main characteristics of the curriculum. The focus in the mathematics curriculum was on the result rather than on the process. In the 2005 Turkish curriculum, the focus on complex cognitive processes, measured in TIMSS, was partially reflected. In 2007, when TIMSS was administered, the 2005 curriculum that pays more attention to solution processes was not yet fully implemented. The teachers had not gone through comprehensive in-service teacher training, therefore they probably relied on their old pedagogical knowledge. As a result, the classes were still predominantly teacher-centered and learning activities were mostly focused on acquiring knowledge rather than problem-solving skills.

In 2008, it was decided that the Australian curriculum had to be renewed. In this new mathematics curriculum, started to be used in 2012, the aim was to use deep knowledge to create new ideas and translate them into practical applications. The new curriculum aimed to be effective in a way that teachers would be flexible to care for the diverse needs of their students. The mathematics curriculum focused on four proficiency strands of understanding, fluency, problem solving and reasoning which are parallel to the TIMSS cognitive domain classification. Before 2012, there was no national curriculum for mathematics. Although each state had similar strands and outcome expectations, there was a need to increase the level of standards across the country (ACARA, 2012). Therefore, in both Turkey and Australia, the students who took the exam in 2007 would not have benefited extensively from new curriculums.

Factors important in predicting mathematics achievement

As TIMSS provides not only student achievement data but also student background data via self-report questionnaires, it is possible to identify student level factors that are important in predicting mathematics achievement. Comparing countries in terms of factors that are important in predicting achievement gives a broad understanding of the nature of cross-cultural score differences and can help to establish whether achievement differences between countries are related to specific student background characteristics, such as parental education.

Walberg's (1981, 1984, 2004) Model of Educational Productivity is one of the widely used achievement models that formed a basis for many achievement studies (Young, Reynolds, & Walberg, 1996). In the model, three general groups of factors are defined to influence student academic achievement, namely student aptitude variables, environmental variables, and instructional variables. Student aptitude variables are related to information about characteristics and background information of students. Examples of environmental variables are social-psychological climate of home, classroom, and peer group. Instructional variables refer to classroom activities that might affect learning. Student aptitude is generally measured as variables such as ability, prior achievement, motivation, self-concept, age, and developmental level; environmental variables generally involved characteristics of the home, classroom, peers, and exposure to mass media related variables; instructional variables are generally represented by quantity, time and quality of instruction. Examples of student aptitude variables of Walberg's model in TIMSS are students' self-confidence, positive affects toward mathematics, and valuing mathematics; an example of environmental variables is educational resources at home; an example of instructional variables is student-centered classroom activities.

Many scholars showed that as self-confidence level increases, so does the achievement level of the students (Liu & Meng, 2010; Shen & Pedulla, 2000; Stankov, Morony, & Lee, 2014). Similarly, Marsh et al. (2013), investigating TIMSS 2007 data, reported that mathematics achievement had a higher correlation with self-concept than positive affect towards mathematics or valuing mathematics in four Arab-speaking and four English-speaking countries. Shen (2002) investigated correlates of mathematics achievement in TIMSS 1999 for 38 countries including Australia and Turkey. Students who reported that they liked math, did well in math, and students who thought that math is easy were more successful in math in almost all countries. Educational resources at home are also one of the important factors that were found to be related to achievement by many scholars (Chevalier & Lanot, 2002; Fuchs & Wöbmann, 2007; Song, Perry, & McConney, 2014). Educational resources at home can enhance the effectiveness of learning time out of school and are positively correlated with student achievement (Kaya & Rice, 2010; Walberg, 2004). The research results about student centered instruction are less clear-cut. Some studies found that student centered instruction is positively associated with achievement (Sabah & Hammouri, 2010) whereas others found that student-centered instruction is either neutral or negatively associated with achievement (Atar & Atar, 2012; Kalender & Berberoglu, 2009; Von Secker & Lissitz, 1999).

Present study

There are some studies testing measurement invariance of TIMSS' content domain (Wu, Li, & Zumbo, 2007) and TIMSS' socioeconomic status indicators (Hansson & Gustafsson, 2013); however, to our knowledge no research has addressed measurement invariance of cognitive domains of TIMSS. This study is novel in that we evaluated construct and item bias of the cognitive domain structure of TIMSS. Testing measurement invariance of cognitive domains among groups is

required to have valid achievement comparisons in terms of these cognitive domains. All language versions of TIMSS are translated from English (Olson, Martin, & Mullis, 2008) and all students are assessed in their language of instruction. In studies that compared many language forms of TIMSS mathematics tests, differentially functioning items were found (Arim & Ercikan, 2014; Ercikan & Koh, 2005; Klieme & Baumert, 2001).

This study is also novel in a way that we used a propensity score approach (e.g., Rubin 2006) to identify the relative importance of significant factors that predict mathematics achievement. Among the significant variables that predict mathematics achievement, the best predictor was identified. In a propensity matching procedure, students are completely matched on background variables (educational resources at home and self-confidence) and countries are then compared as if they were fully matched on these background variables. So, the question is addressed what the results would be if Turkish students had the same background variables as Australian students. The propensity score approach identifies remaining differences among these students. In educational comparative research, the propensity score approach is increasingly used (Lottridge, Nicewander, & Mitzel, 2011; Ruzek, Burchinal, Farkas, & Hibel, 2010; Sullivan & Field, 2013).

In this study we aimed to identify factors predicting mathematics achievement of Turkish and Australian students. In order to achieve this goal, construct bias and item bias were evaluated to assess comparability of data. Then, by building an achievement model, measurement invariance of the model was tested, significant factors were identified and using a propensity score approach, presumably important antecedents of mathematics achievement were evaluated. This study contributes to the literature in a methodological way by showing steps of conducting a comparative study and in a substantive way by identifying correlates of mathematics achievement measured by various cognitive dimensions. This study is mainly based on TIMSS 2007 data and the same procedures were repeated with TIMSS 2011 data. The research questions of this study are:

1. To what extent can the TIMSS mathematics achievement differences between Turkish and Australian students be accounted for by bias (construct and item bias)?
2. To what extent can the TIMSS mathematics achievement differences between Turkish and Australian students be accounted for by pupil background characteristics (attitudes/motivations and home resources)?

METHOD

Participants

The data of this study were obtained from the TIMSS 2007 and TIMSS 2011 database. In TIMSS, the target population comprises all students at the fourth and eighth grade of participating countries. This study used all Turkish and Australian eighth grade students who answered released items. The 2007 data from 1588 Turkish students (732 females and 856 males) and 1463 Australian students (646 females and 817 males) and the 2011 data from 503 Turkish students (270 females and 233 males) and 524 Australian students (256 females and 268 males) were investigated.

Measures

TIMSS gathered data on student's mathematics achievement and student's background characteristics via achievement tests and a student questionnaire, respectively. The mathematics achievement items in TIMSS are either in multiple choice format or in open-ended format. The TIMSS report mentions for each

achievement item which of the three cognitive domains it represents. The questionnaire items are generally in Likert format asking endorsement. In the TIMSS questionnaire, related to student aptitude variables of Walberg's model, information was collected about self-confidence, positive affects toward mathematics, and valuing mathematics. The indicators of the latent variables of self-confidence, positive affect toward math, and valuing mathematics were selected in line with TIMSS definitions. The TIMSS report also mentions indexes for each of these latent variables (Mullis, Martin, & Foy, 2008). *Self-confidence* was represented by four questionnaire items, such as "I usually do well in mathematics", and "Math is harder for me". *Positive affect toward mathematics* was represented by three items, such as "I enjoy learning mathematics" and "Mathematics is boring". *Valuing mathematics* was represented by four items such as "I think learning mathematics will help me in my daily life" and "I need mathematics to learn other school subjects". Related to environmental variables of Walberg's model, information about home educational resources was collected in TIMSS. The observed items related to educational resources at home that had at least three response categories were used to represent educational resources at home latent variable. In the study, the construct of *educational resources at home* was represented by items such as "the number of books in student's home" and "the highest educational level of mother". Finally, *student-centered classroom learning activities* were represented by four items about practices in mathematics lessons, such as "we explain our answers" and "we decide on our own procedures for solving complex problems".

Data analysis

As a first analysis, construct equivalence and item bias were evaluated to ensure that the released achievement items represented the same cognitive constructs and they were item bias free for Turkish and Australian students. For construct equivalence analysis, the three-dimensional cognitive domain structure proposed by TIMSS in mathematics test was evaluated. In the analysis, each item was associated with its own cognitive domain and invariance of this structure between the two countries was tested. This analysis was conducted for booklet one to five, as these booklets had more than one item associated with the relevant cognitive domains. These five booklets were administered to both Turkish and Australian students. Following construct equivalence evaluation, item bias was investigated using structural equation modeling (SEM). In SEM, the metric and the scalar model are expected to have values of CFI and TLI that are .90 or above (Cheng & Rensvold, 2002). If the difference in fit between metric and scalar model is larger than .01, modification indices are investigated to identify items that affect this difference, possibly followed by the removal of these items.

The TIMSS study employed standardized procedures for administration, which reduced the likelihood of administration differences as a source of method bias. Van de Vijver, Hofer, and Chasiotis (2010) suggested several steps to minimize the method bias in cross-cultural studies. Test administrators should be given an intensive training, a detailed instructions and a manual for administration, scoring, and interpretation should be prepared, and important variables related to samples should be balanced. The technical report published after each TIMSS administration reported how operations were done and how qualities of procedures were assured (Martin, & Mullis, 2012). It is stated that the TIMSS & PIRLS International Study Center prepared a document that describes step by step all operational activities. National research coordinators of each country were provided special software to support sampling, tracking classes and students, administering questionnaires, documenting scoring reliability and creating and checking data files. The TIMSS & PIRLS International Study Center also provided intensive training for each country

in these administrative jobs. Additionally, International Quality Control Monitors (IQCM) had visited at least 15 schools to control the administration of the assessment and data collection process, and to evaluate the quality of the testing sessions. All of these strict controls are expected to lead a measure that minimizes administration differences as sources of method bias. Additionally, Hooper, Arora, Martin, and Mullis (2013) examined whether extreme response sets occurred for each country in TIMSS 2011 questionnaire and they reported that in both Turkey and Australia, only 0.1% of the students marked “agree a lot” regardless of the direction of the items. This value ranged from 0.0% to 2.2% among other countries. This report also showed that for Turkey and Australia, method bias was not a serious problem for the results obtained in this study. An important source of method bias in the cognitive instruments that could not be controlled was familiarity with the types of tests used in TIMSS. The released data set does not provide information about factors such as previous test experience. As a consequence, we could not evaluate the influence of test familiarity as a source of method bias.

After having evaluated construct bias and item bias, the nature of mathematics achievement differences among Turkish and Australian students was explored. In this analysis, firstly, a structural equation model related to factors predicting mathematics achievement was formed and tested for measurement invariance. Secondly, factors predicting mathematics achievement of Turkish and Australian students were identified. Lastly, mean differences of these latent factors were analyzed to understand whether means of significant latent factors also differed between these countries by conducting latent mean structure analysis. In the analysis, Turkey was chosen as reference country. The MPLUS 7.11 program was used for testing construct equivalence and item bias, for examining the relationships between background variables and test achievement, and for latent mean structure analysis.

Finally, we used propensity score matching to identify the remaining achievement differences between these countries (if any), when students in both groups are matched based on background variables with a presumed bearing on mathematics achievement scores. Consequently, we matched students based on the background variables and evaluate what would be the results if Turkish students had the same background variables as Australians. In order to conduct this analysis, propensity scores of each student were estimated by full optimal matching procedure (e.g., Guo & Fraser, 2009), which is a novel approach. A full optimal matching procedure was selected because in this method all students were kept in the analysis, whereas in other matching procedures unmatched parts of the sample are discarded from the analysis (e.g., exact matching, neighbor matching, and optimal matching) (Stuart, 2010). The MatchIt R package (Ho, Imai, King, & Stuart, 2007) was used to do the matching and to estimate propensity scores. Then, the propensity scores were used as a covariate in a MANCOVA analysis to test remaining differences among these students.

RESULTS

We first present the results of the TIMSS 2007 dataset, followed by the cross-validation of the results based on the TIMSS 2011 dataset.

Internal consistency analysis of the instrument

The values of Cronbach’s alpha reliability coefficients in TIMSS 2007 mathematics test booklets showed values ranging from .732 to .903 for Turkish students and

from .710 to .875 for Australian students. These values are satisfactory (Cicchetti, 1994).

Construct equivalence of TIMSS cognitive domain structure and item bias

In order to answer the first research question, construct equivalence of the cognitive structure and item bias for Turkish and Australian students were evaluated. A central question to be addressed in construct equivalence was whether the three-dimensional structure proposed by TIMSS was invariant and therefore had the same meaning for Turkish and Australian students. Configural model results showed that for all five booklets, the fit values were within acceptable ranges as CFI and TLI values were .90 or above (Cheung & Rensvold, 2002) (see Table 1). This evidence supportive of construct equivalence implied that mathematics skills as measured by the TIMSS mathematics tests were free of construct bias for both groups of students.

Table 1 showed that for all five booklets the differences between the metric and scalar model were larger than .01, which implied that there might be some biased items. Modification indices were used to evaluate the difference between the metric and scalar model smaller and to identify biased items. For booklet one, modification indices suggested that items M022049 and M042301A were problematic in terms of bias. The problematic items in the first booklet are given in Appendix A. According to the TIMSS classification, item M022049 was related to the geometry content domain and the reasoning cognitive domain whereas M042301A was related to the algebra content domain and knowing cognitive domain. These two items were evaluated in terms of language by three experts and no major problems with translation of the items were identified. However, in these items visualization and generalization skills were measured. Turkish students generally work with routine problems that require a solution process mainly consisting of conducting calculations. For instance, if only the sum of interior angles was asked in item M042301A, the items might not be identified as biased. When it is required to link and generalize the sum of interior angles with the number of triangles in it, the item

Table 1. Construct equivalence and item bias results of TIMSS 2007 mathematics test for Turkish and Australian students

Booklet	Number of Items	Model	χ^2/df	$\Delta\chi^2/\Delta df$	RMSEA	CFI	ΔCFI	TLI	ΔTLI
1	29	Configural	1.382*		.035	.962		.959	
		Metric	1.473*	4.075*	.039	.951	.011	.949	.010
		Scalar	1.786*	10.145*	.051	.916	.035	.915	.034
		Scalar-NoDif	1.551*	.807	.043	.945	.006	.944	.005
2	31	Configural	1.304*		.031	.964		.961	
		Metric	1.304*	1.294*	.031	.963	.001	.961	.000
		Scalar	1.729*	13.506*	.048	.908	.055	.907	.054
3	32	Configural	1.157*		.023	.974		.972	
		Metric	1.292*	5.587*	.031	.949	.025	.947	.025
		Scalar	1.817*	17.409*	.052	.854	.095	.853	.094
4	29	Configural	1.208*		.026	.971		.969	
		Metric	1.387*	6.551*	.036	.945	.026	.942	.027
		Scalar	1.921*	16.169*	.055	.863	.082	.862	.080
5	12	Configural	1.418*		.037	.973		.965	
		Metric	1.614*	3.830*	.045	.957	.016	.948	.017
		Scalar	3.317*	19.070*	.087	.818	.139	.805	.143

* $p < .001$

probably became relatively difficult for Turkish students. Therefore, the reason why these items functioned differentially might be that these items required visualization and generalization which were not well covered in the Turkish educational system at that time. Removing these two items produced a model ($\Delta CFI = .006$) that was invariant for the two countries.

For the other four booklets, removing items that were suggested by modification indices did not produce an invariant model (See Table 1) unless many items were removed, thereby challenging the coverage of the underlying concept by the remaining items. Therefore, it was concluded that only the first booklet, with some revision, was invariant and free of item bias across groups, a prerequisite for our propensity scoring analyses. These two biased items in booklet one were removed from following analyses.

In Table 2, achievement differences between Turkish and Australian students before and after removing biased items were reported. One biased item from knowing and one biased item from reasoning were removed from the first booklet. Then achievement differences before and after removing were evaluated. This was done by conducting a MANOVA with country as the independent variable and the achievement scores as dependent variables. In the first step all item scores were used to compute the achievement score, in the second step scores on the presumably biased items were not used to compute the total scores. We followed the same procedure for the second booklet where we removed four items in two steps of two items. As can be seen in Table 2, effect sizes did not change after presumably biased items were removed. So, our results suggested that mathematics achievement differences between Turkish and Australian students could not be accounted for by item bias. Therefore, the answer for the first research question is that mathematics achievement differences between Turkish and Australian students could not be accounted for by item bias. It is unlikely that achievement differences between Turkish and Australian students are due to the fact that the items were not originally developed in Turkish but in English.

Nature of mathematics achievement differences

The second research question dealt with factors that can predict mathematics achievement, whether these factors differ in magnitude, and whether these factors were equally predictive in Turkey and Australia. A SEM model was formed by using related TIMSS student factors as achievement predictor and by using students' TIMSS mathematics cognitive domain scores as achievement indicator. In the model, the latent factors that were expected to predict mathematics achievement were self-confidence, positive affect toward mathematics, valuing mathematics, educational resources at home, and classroom learning activities. In the model, mathematics

Table 2. Effect sizes before and after removing biased items

Variable	Country	Booklet 1				Booklet 2					
		Before removing biased items		After removing biased items		Before removing biased items		After removing two biased items		After removing four biased items	
		<i>M</i>	η^2	<i>M</i>	η^2	<i>M</i>	η^2	<i>M</i>	η^2	<i>M</i>	η^2
Knowing score	Turkey	.095	.00	.095	.00	-.260	.07**	-.272	.08**	-.285	.09**
	Australia	.121		.116		.277		.289		.303	
Applying score	Turkey	-.151	.08**	-.151	.08**	-.200	.04**	-.185	.04**	-.209	.05**
	Australia	.443		.443		.212		.197		.223	
Reasoning score	Turkey	-.111	.03**	-.091	.02*	-.125	.02*	-.125	.02*	-.125	.02*
	Australia	.252		.237		.133		.133		.133	

* $p < .01$. ** $p < .001$.

achievement (based on the first booklet only to avoid bias problems) was represented by knowing, applying, and reasoning cognitive domains and the students' domain scores were estimated by using an IRT two-parameter logistic model.

The first step in evaluating the model was to test measurement invariance of the achievement model for Turkish and Australian students (See Figure 1). The demonstration of measurement invariance was a necessary condition to make evaluations about effects of these factors in both countries. Configural, weak (factor loadings are invariant), and strong (factor loadings and regression coefficients are invariant) invariance were supported as the ΔCFI value was less than .010 (See Table 3). This result implied that the same factors were effective in predicting mathematics achievement of both countries. Therefore, the same model given in Figure 1 could be used to identify significant predictors of mathematics achievement. The model explained 52% of the variance of mathematics achievement.

The second step was to identify which factors of the model could predict mathematics achievement. Standardized regression coefficients showed that having

Table 3. Measurement invariance of the model for factors predicting achievement

Model	χ^2/df	$\Delta\chi^2/\Delta df$	RMSEA	CFI	ΔCFI	TLI	ΔTLI
Unconstrained	1.627*		.045	.940		.928	
Loadings invariant	1.638*	1.942*	.046	.936	.004	.926	.002
Loadings and intercepts invariant	1.646*	2.218*	.046	.935	.001	.926	.000

* $p < .001$

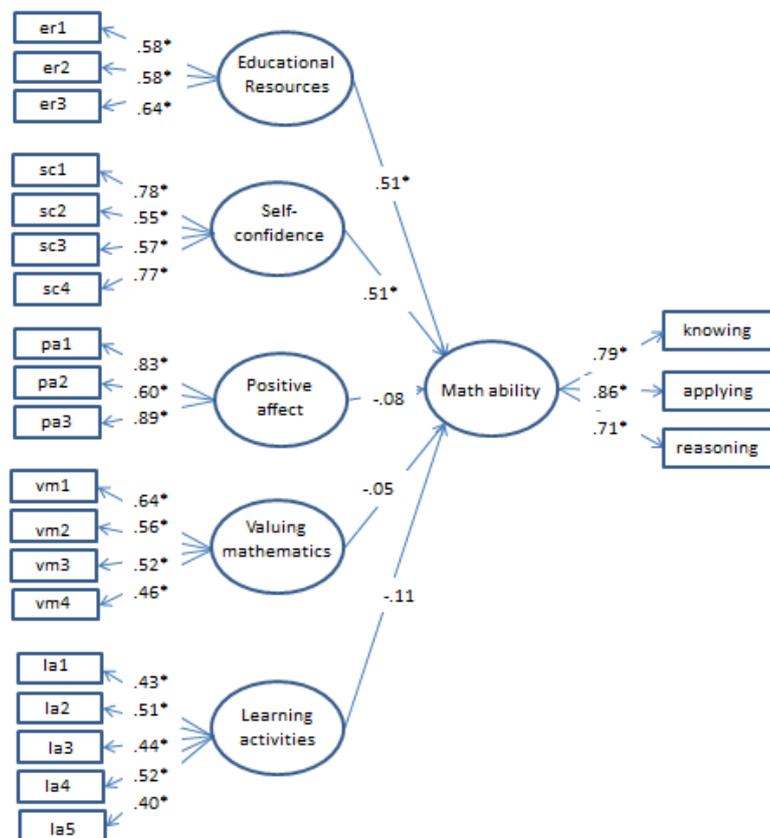


Figure 1. Mathematics Achievement Model for Turkish and Australian students

high self-confidence and having more educational resources at home were significantly and positively related with mathematics achievement of students in both countries (See Figure 1). However, positive affect toward mathematics, valuing mathematics and classroom learning activities were not significantly related to mathematics achievement in any country.

The third step was to evaluate latent factor means on all latent variables. Especially, latent mean differences on significant factors that were important in predicting mathematics achievement could be expected to explain the nature of mathematics achievement difference between these two countries. Table 4 showed that Australian students had a significantly higher level of educational resources at home and had higher self-confidence in the TIMSS 2007 data set whereas Turkish students had a significantly higher level of positive affect towards mathematics, valuing mathematics and classroom learning activities. Having observed that there are significant country differences in antecedent variables, we now turn to the question of how effective these variables are to explain country differences in achievement, using propensity score matching.

Effects of propensity scores on cross-national achievement differences

We were interested in the relative importance of educational resources at home and self-confidence in predicting cross-cultural achievement differences, given that these variables are very different in nature. To achieve this, propensity scores for both groups were estimated by matching Turkish and Australian students firstly on educational resources at home, then on self-confidence, and finally on both variables using full optimal matching procedure. Propensity score matching provides an estimate of what the achievement difference would be between Turkish and Australian students if they had an equal level of these background variables. Table 5 shows MANOVA results before correcting for propensity scores and MANCOVA results after correction. Before correcting, Australian and Turkish students had similar scores on knowledge items but Australian students had higher achievement scores on applying and reasoning cognitive domain scores, with effect sizes of .08 and .02, respectively (medium and small according to Cohen's, 1988, guideline). After using only the self-confidence propensity score as a covariate, the effect size of the difference decreased; yet, Australian students were still more successful than

Table 4. Latent Mean Comparison TIMSS 2007

Latent Factors	<i>M</i>	<i>p</i>
Educational Resources at Home	1.357	< .001
Self-confidence	.202	.049
Positive Affect toward Mathematics	-1.080	< .001
Valuing Mathematics	-.254	.018
Classroom Learning Activities	-1.102	< .001

Turkey is the reference country; a score above/below zero indicates that Australia has a higher/lower score

Table 5. MANOVA and MANCOVA results before and after correcting for propensity scores

Variable	Country	Before correcting for propensity scores		After correcting for self-confidence		After correcting for educational resources		After correcting for both	
		<i>M</i>	η^2	<i>M</i>	η^2	<i>M</i>	η^2	<i>M</i>	η^2
Knowing score	Turkey	.095	.00	.162	.01	.310	.05***	.294	.04***
	Australia	.116		-.002		-.262		-.235	
Applying score	Turkey	-.151	.08***	-.099	.05***	.067	.00	.047	.00
	Australia	.443		.352		.058		.094	
Reasoning score	Turkey	-.091	.02**	-.049	.01*	.129	.01*	.115	.01*
	Australia	.237		.163		-.151		-.126	

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

Turkish students in applying and reasoning cognitive domains. After using only the educational resources at home propensity score as a covariate, the direction of achievement difference was reversed and Turkish students were more successful than Australian students in the knowing and reasoning domains. Finally, when correcting for both sets of variables, the results were similar to those found when only educational resources at home were used for matching. Therefore, it was concluded that educational resources and self-confidence were indeed effective in explaining achievement differences (with the former being more effective than the latter).

Cross-validation of the results with TIMSS 2011 dataset

We investigated whether results obtained with the TIMSS 2007 dataset could be replicated in the TIMSS 2011 dataset (using booklet 1 of that data set). Booklet 1 of TIMSS 2011 had 26 items and all of these items were released. Invariance results are reported in Table 6. As can be seen there, the CFI differences between metric and scalar model were larger than .01, which implied that there might be some biased items for Turkish and Australian students. Modification indices were used to make the difference between metric and scalar model smaller and to identify biased items. Modification indices suggested that items M052061, M052214, M052408 and M052429 revealed bias. Removing these items produced an unbiased form of the test for Turkish and Australian students.

In Table 7, achievement differences between Turkish and Australian students before and after removing biased items were reported. Effect sizes did not change dramatically after presumably biased items were removed. So, our results suggested that mathematics achievement differences between Turkish and Australian students could not be accounted for by bias. Therefore, the answer to the first research question is that mathematics achievement differences between Turkish and Australian students could not be accounted for by bias. It is unlikely that achievement differences between Turkish and Australian students are due to the fact that the items were not originally developed in Turkish but in English, which replicates findings for the 2007 dataset.

Table 6. Construct equivalence and item bias results of TIMSS 2011 mathematics test for Turkish and Australian students

Booklet	Number of Items	Model	χ^2/df	$\Delta\chi^2/\Delta df$	RMSEA	CFI	ΔCFI	TLI	ΔTLI
1	26	Configural	1.664*		.036	.988		.987	
		Metric	1.662*	1.595	.036	.988	.000	.987	.000
		Scalar	2.474*	21.690*	.054	.971	.017	.971	.016

* $p < .001$.

Table 7. Effect sizes before and after removing biased items

Variable	Country	Booklet 1			
		Before removing biased items		After removing biased items	
		<i>M</i>	η^2	<i>M</i>	η^2
Knowing score	Turkey	-.220	.05***	-.240	.06***
	Australia	.211		.230	
Applying score	Turkey	-.033	.00	.026	.00
	Australia	.032		-.025	
Reasoning score	Turkey	-.072	.01*	-.046	.00
	Australia	.069		.044	

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

Table 8. Measurement invariance of the model for factors predicting achievement

Model	χ^2/df	$\Delta\chi^2/\Delta df$	RMSEA	CFI	ΔCFI	TLI	ΔTLI
Unconstrained	2.686**		.058	.949		.936	
Loadings invariant	2.646**	1.862	.057	.947	.002	.937	-.001
Loadings and intercepts invariant	2.683**	4.794*	.058	.945	.002	.936	.001

* $p < .01$. ** $p < .001$.

Table 9. Latent Mean comparison TIMSS 2011

Latent Factors	<i>M</i>	<i>p</i>
Educational Resources at Home	1.121	< .001
Self-confidence	.131	.086
Positive Affect toward Mathematics	-.769	< .001
Valuing Mathematics	.379	< .001

Turkey is the reference country; a score above/below zero indicates that Australia has a higher/lower score

In order to identify which factors predict mathematics achievement, whether these factors differ in magnitude, and whether these factors were equally predictive in Turkey and Australia, a SEM model was made by using related TIMSS student factors as achievement predictor and by using students' TIMSS mathematics cognitive domain scores as achievement indicators. In the model, the latent factors to predict mathematics achievement were self-confidence, positive affect toward mathematics, valuing mathematics, and educational resources at home. Classroom learning activities items were not included in the 2011 questionnaire; therefore, this latent variable could not be used. Configural, weak (factor loadings are invariant), and strong (factor loadings and regression coefficients are invariant) invariance were supported (See Table 8). This result implies that the same factors were effective in predicting mathematics achievement of both countries as in the TIMSS 2007 dataset.

The second step was to identify which factors of the model could predict mathematics achievement. Standardized regression coefficients showed that having high self-confidence ($\beta = .67$) and having more educational resources at home ($\beta = .34$) were significantly and positively related with mathematics achievement of students in both countries. Positive affect toward mathematics was found to be negatively related to mathematics, yet with a small effect size ($\beta = -.12$). Valuing mathematics was not significantly related to mathematics achievement in any country.

The third step was to evaluate latent factor means on all latent variables. Especially, latent mean differences on significant factors that were important in predicting mathematics achievement would expect to explain the nature of mathematics achievement difference between these two countries. Table 9 showed that Australian students had a significantly higher level of educational resources at home and they value mathematics whereas Turkish students had a significantly higher level of positive affect towards mathematics.

We were interested in the relative importance of educational resources at home and self-confidence in predicting cross-cultural achievement differences, given that these variables are very different in nature. In order to achieve this goal, propensity scores for both groups were estimated by matching Turkish and Australian students firstly on educational resources at home, then on self-confidence, and finally on both groups of variables using full optimal matching procedure. Before correcting, Australian and Turkish students had similar scores on applying and reasoning items but Australian students had higher achievement scores on knowledge items (See Table 10). After using only self-confidence propensity score as a covariate, the effect size of the difference decreased; however, Australian students were still more successful than Turkish students in knowledge cognitive domains. After using only

Table 10. MANOVA and MANCOVA results before and after correcting for propensity scores

Variable	Country	Before correcting for propensity scores		After correcting for self-confidence		After correcting for educational resources		After correcting for both	
		<i>M</i>	η^2	<i>M</i>	η^2	<i>M</i>	η^2	<i>M</i>	η^2
Knowing score	Turkey	-.240	.06*	-.142	.03*	.014	.00	.032	.00
	Australia	.230		.217		-.127		-.165	
Applying score	Turkey	.026	.00	.091	.00	.252	.04*	.251	.03*
	Australia	-.025		-.018		-.371		-.369	
Reasoning score	Turkey	-.046	.00	.059	.00	.204	.03*	.213	.03*
	Australia	.044		.020		-.300		-.320	

* $p < .001$.

the educational resources at home propensity score as a covariate, Turkish students were more successful than Australian students in applying and reasoning. Finally, when correcting for both sets of variables, the results were similar to those found when only educational resources at home were used for matching. Therefore, it was concluded that educational resources and self-confidence were indeed effective in explaining achievement differences (with the former being more effective than the latter), which we also found in the 2007 dataset.

DISCUSSION

We were interested in understanding the relatively low mathematics achievement level of Turkey in international studies. We used data from the international TIMSS 2007 and 2011 study to compare Turkey with Australia in terms of mathematics achievement. Australia was chosen as a reference as it is a country with a typical Western type of education and a higher achievement level in TIMSS than Turkey. In the study, we were interested in identifying predictors of mathematics achievement in Turkey and Australia and understanding how many achievement differences would still be remaining if we statistically correct for bias and if we match the Australian and Turkish samples on relevant background variables. Identifying significant predictors and evaluating their effects on achievement are expected to increase our insights and to contribute and help stakeholders of education to appreciate how to increase scores in these large scale tests, especially if the predictor variables could be influenced by educational policy measures.

In a first step we identified construct and item bias. Construct bias and item bias results indicated that conditions of construct and metric invariance were met in all five booklets but that only the first booklet showed scalar invariance after the removal of two items and could be used to compare Turkish and Australian students. Having a booklet free of construct and item bias was an important condition for subsequent analyses as we wanted to compare achievement levels that were not influenced by biasing factors. Our item bias analysis led to two conclusions. First, we found that even the carefully constructed educational achievement tests of TIMSS produced much item bias. Follow-up studies would be needed to address the nature of the bias. So, we can only speculate here about the nature of the bias. Van Schilt-Mol (2007) identified item bias in nation-wide administered educational tests (Cito tests) in Dutch primary schools. The immigrant sample comprised many ethnic groups, among which Turkish-Dutch. She found that there was no single reason of item bias but that specific words or concepts were less familiar to immigrant pupils compared to mainstream pupils. She found that removal of such words or concepts could eliminate ethnic group performance differences. Second, a comparison of the country differences before and after removing biased items in the first two booklets,

which contained only a few biased items, revealed that it is unlikely that achievement differences in TIMSS scores found in Turkey and Australia can be accounted for by item bias in either the 2007 or 2011 datasets.

In the achievement model, which was found to hold in both countries, the self-confidence of the students, which is related to the student aptitude variable of Walberg's model, and their educational resources at home, which is related to environmental variable of Walberg's model, were two significant predictors of achievement. Australian students reported both in 2007 and in 2011 that they had not only more educational resources at home but they also had a higher self-confidence (although the latter difference was only significant in 2007). Matching Turkish and Australian students on these two background factors using a propensity score matching approach indicated that educational resources at home were more effective than self-confidence in explaining achievement differences. This confirms earlier findings; many scholars found a similar relationship between having more educational resources at home and being more successful (Berberoglu et al. 2003; Chevalier & Lanot, 2002; Fuchs & Wößmann, 2007; Kaya & Rice, 2010) and a similar relationship between self-confidence and achievement (Abu-Hilal, 2000; Marsh, 1986; Shen & Tam, 2008).

The classroom learning activities, related to the third dimension of Walberg's achievement model, were not significant in predicting achievement in 2007. In 2011, this dimension was not represented in the questionnaire. We could not find any relationship between student-centered classroom activities and mathematics achievement. One of the reasons for this might be that student-centered activities might not be performed adequately during the lessons. In student-centered teaching, not only students but also teachers should be active. It is expected from students to construct new information by using prior knowledge under teacher guidance. The coaching role of the teacher in the student-centered learning environment is crucial. Letting students work on their own is not the target of student-centered teaching. Similar conclusions were reached by other scholars who showed that student-centered learning activities were not positively associated with achievement (Kalender & Berberoglu, 2009; Von Secker & Lissitz, 1999). Moreover, another reason for the lack of correlation could be that, as indicated earlier, in Turkey and in Australia, the new curricula were not yet effective in 2007. Therefore, teachers still adhered to teacher-centered practices. More research is needed before recommending how pupil-centered practices should be implemented in the Turkish math classroom.

In the study, mathematics achievement was modeled by using the cognitive domain structure of TIMSS, comprising knowing, applying, and reasoning. There is a hierarchical structure among these three cognitive domains. Going from knowing to reasoning, the cognitive demands to answer the items are assumed to increase. Results showed that Turkish and Australian students had similar mean scores in the knowing domain in 2007. These findings suggest that Turkish and Australian students do not show appreciable differences in knowledge of basic mathematical facts, concepts, and procedures. However, when an item is related to applying or reasoning, Turkish students tended to show a lower achievement than their Australian peers. Application items are related to students' ability to apply knowledge and conceptual understanding to solve problems whereas reasoning items require the use of mathematics in unfamiliar situations, complex contexts, and multistep problems. So, our study suggests that Turkish students experience more problems with applying their knowledge in both routine and non-routine problem situations than Australian students. Similarly, 2011 data suggested that Turkish students got a significantly lower score in knowing and reasoning and a slightly lower score in applying (although the latter was not significant). By implication, our

study points to a need to improve cognitive skills of Turkish students. In 2013, the mathematics curriculum of the fifth up to the eighth grade (middle school) was updated. The general purpose of the curriculum is stated as helping students to acquire mathematical knowledge, skills, and attitudes. It is also claimed that emphasis is given to conceptual learning, fluency in operation, mathematical communication, valuing mathematics, and developing problem solving skills (MONE, 2013). This policy addresses the problem we observed in our study. No systematic evaluation of the impact of the new policy is available yet. Data, predating the policy change, indicate that the Turkish educational system was not much focused on the development of higher-order skills. Classroom activities in Turkey were still mainly focusing on items that are related to basic skills as comprehension rather than higher order thinking skills as problem solving (Doganay & Bal, 2010). Similarly, preservice teachers in Turkey reported that they experience difficulties in finding non-routine problem situations (Temur, 2012). Based on TIMSS 2007 and 2011 data, our findings show that higher-order cognitive processes, such as problem solving and reasoning, are not adequately developed among Turkish students. In order to obtain firmer evidence, outcomes of Turkish students in coming cycles of TIMSS, for instance TIMSS 2019 and PISA 2023, need to be investigated for the achievement levels in higher-order thinking skills to see the effects of 2013 curriculum.

All findings indicate that as educational resources at home are more effective in predicting mathematics achievement and as Turkish students were less successful in many cognitive domains, it is important to emphasize that parents could be actively involved in stimulating their children by providing a cognitively enriching environment. The higher education level of parents and the possession of more books at home are proxies for a complex set of variables which are presumably related to a cognitively more stimulating environment in the home. Even if the mechanisms behind the enrichment are not yet well understood, it is clear that providing a stimulating learning at home can contribute considerably to the child's cognitive development.

Our study has some limitations. First of all, only data of one booklet was suitable to compare Turkish and Australian students. Although the number of items to estimate achievement level of students was not very small (27 in 2007 and 26 in 2011), only students who answered the first booklet could be used. As we estimated student abilities based on released items and as we did not compare results of students from different booklets, we did not include sampling weights in our estimation. This might lead to an increase in the standard error of the estimation. Another limitation is related to the choice of comparison country. Australia was chosen as TIMSS items were originally prepared in English and Australia is one of the English speaking countries that got higher achievement score than Turkey in TIMSS 2007 and 2011. In future research, the same procedures described here could be repeated by using data from other English speaking countries or from any other high achieving country. Additionally, there might be other factors, such as teacher and school level, that are significant in predicting student achievement. As this study focused on student level variables and propensity score application to student level predictors, these factors were not included in our study.

ACKNOWLEDGEMENT

This paper is a part of a research project supported by the Scientific and Technological Research Council of Turkey (TUBITAK) with program no. 2219. Any opinions expressed here are those of the authors and do not necessarily reflect the views of the TUBITAK.

REFERENCES

- Abu-Hilal, M. M. (2000). A structural model for predicting mathematics achievement: Its relation with anxiety and self-concept in mathematics. *Psychological Reports, 86*, 835-847. doi:10.2466/pr0.2000.86.3.835
- ACARA. (2012). *The shape of the Australian Curriculum*. Retrieved from http://www.acara.edu.au/verve/_resources/The_Shape_of_the_Australian_Curriculum_v4.pdf.
- Arim, R. G., & Ercikan, K. (2014). Comparability between the American and Turkish version of the TIMSS mathematics test results. *Education and Science, 39*(172), 33-48.
- Atar, H. Y., & Atar, B. (2012). Examining the effects of Turkish Education reform on students' TIMSS 2007 science achievements. *Educational Sciences: Theory and Practice, 12*(4), 2632-2636.
- Berberoglu, G., Çelebi, Ö., Özdemir, E., Uysal, E., & Yayan, B. (2003). Factors affecting achievement level of Turkish students in the Third International Mathematics and Science Study. *Educational Sciences and Practice, 2*(3), 3-14.
- Cheung, G. W., & Rensvold, R. B. (2002) Evaluating Goodness-of-Fit Indexes for testing measurement invariance. *Structural Equation Modeling: A Multidisciplinary Journal, 9*, 233-255. doi:10.1207/S15328007SEM0902_5
- Chevalier, A., & Lanot, G. (2002). The relative effect of family characteristics and financial situation on educational achievement. *Education Economics, 10*, 165-181. doi:10.1080/09645290210126904
- Cicchetti, D. V. (1994). Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. *Psychological Assessment, 6*, 284-290. doi:10.1037/1040-3590.6.4.284
- Cohen, J (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Mahwah, NJ: Lawrence Erlbaum Associates.
- Doganay, A., & Bal, A. P. (2010). The measurement of students' achievement in teaching primary school fifth year mathematics classes. *Educational Sciences: Theory and Practice, 10*, 199-215.
- Ercikan, K., & Koh, K. (2005). Examining the construct comparability of the English and French versions of TIMSS. *International Journal of Testing, 5*, 23-35. doi:10.1207/s15327574ijt0501_3
- Fuchs, T., & Wößmann, L. (2007). What accounts for international differences in student performance? A re-examination using PISA data. *Empirical Economics, 32*(2), 433-464. doi:10.1007/s00181-006-0087-0
- Guo, S., & Fraser, M. W. (2009). *Propensity score analysis: Statistical methods and applications*. Thousand Oaks, CA: Sage Publications.
- Hansson, A., & Gustafsson, J. E. (2013). Measurement invariance of socioeconomic status across migrational background. *Scandinavian Journal of Educational Research, 57*, 148-166. doi:10.1080/00313831.2011.625570
- He, J., & Van de Vijver, F. J. R. (2013). Methodological issues in cross-cultural studies in educational psychology. In G. A. D. Liem & A. B. I. Bernardo (Eds.), *Advancing cross-cultural perspectives on educational psychology: A festschrift for Dennis McInerney* (pp. 39-56). Charlotte, NC: Information Age Publishing.
- Ho, D., Imai, K., King, G., & Stuart, E. (2007). Matching as Nonparametric Preprocessing for Reducing Model Dependence in Parametric Causal Inference. *Political Analysis, 15*, 199-236. doi:10.1093/pan/mdl013
- Hooper, M., Arora, A., Martin, M. O., Mullis, I. V. S. (2013). Proceedings from IRC 2013: *Examining the Behavior of "Reverse Directional" Items in the TIMSS 2011 Context Questionnaire Scales*. Retrieved from http://www.iea.nl/fileadmin/user_upload/IRC/IRC_2013/Papers/IRC-2013_Hooper_etal.pdf.
- Kalender, I., & Berberoglu, G. (2009). An assessment of factors related to science achievement of Turkish students. *International Journal of Science Education, 31*, 1379-1394. doi:10.1080/09500690801992888
- Kaya, S., & Rice, D. C. (2010). Multilevel effects of student and classroom factors on elementary science achievement in five countries. *International Journal of Science Education, 32*, 1337-1363. doi:10.1080/09500690903049785

- 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*, 385-402. doi:10.1007/BF03173189
- Koc, Y., Isiksal, M., & Bulut, S. (2007). Elementary school curriculum reform in Turkey. *International Education Journal, 8*, 30-39.
- Liu, S., & Meng, L. (2010). Re-examining factor structure of the attitudinal items from TIMSS 2003 in cross-cultural study of mathematics self-concept, *Educational Psychology: An International Journal of Experimental Educational Psychology, 30*, 699-712. doi:10.1080/01443410.2010.501102
- Lottridge, S. M., Nicewander, W. A., & Mitzel, H. C. (2011). A comparison of paper and online tests using a within-subjects design and propensity score matching study. *Multivariate Behavioral Research, 46*, 544-566. doi:10.1080/00273171.2011.569408
- Marsh, H. W. (1986). Verbal and math self-concepts: An internal/external frame of reference model. *American Educational Research Journal, 23*, 129-149. doi:10.3102/00028312023001129
- Marsh, H. W., Abduljabbar, A. S., Abu-Hilal, M. M., Morin, A. J., Abdelfattah, F., Leung, K. C., Xu, M. K., Nagengast, B., & Parker, P. (2013). Factorial, convergent, and discriminant validity of TIMSS math and science motivation measures: A comparison of Arab and Anglo-Saxon countries. *Journal of Educational Psychology, 105*, 108-128. doi:10.1037/a0029907
- Martin, M. O., & Mullis, I. V. S. (Eds.). (2012). *Methods and procedures in TIMSS and PIRLS 2011*. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.
- Ministry of National Education [MONE] (2013). *Ortaokul Matematik Dersi (5, 6, 7 ve 8. Sınıflar) Öğretim Programı (Middle School Mathematics Course Educational Program)*. Ankara, Turkey.
- Mullis, I. V. S., Martin, M. O., & Foy, P. (2008). *TIMSS 2007 international mathematics report: Findings from IEA's Trends in International Mathematics and Science Study at the fourth and eighth grades*. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.
- Mullis, I. V. S., Martin, M. O., Foy, P., & Arora, A. (2012). *TIMSS 2011 International results in mathematics*. Amsterdam: International Association for the Evaluation of Educational Achievement.
- Mullis, I. V. S., Martin, M. O., Ruddock, G. J., O'Sullivan, C. Y., Arora, A., & Erberber, E. (2005). *TIMSS 2007 assessment frameworks*. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.
- Olson, J. F., Martin, M. O., & Mullis, I. V. S. (Eds.). (2008). *TIMSS 2007 technical report*. Chestnut Hill, MA: Boston College.
- Rubin, D. B. (2006). *Matched sampling for causal effects*. New York, NY: Cambridge University Press.
- Ruzek, E., Burchinal, M., Farkas, G., & Duncan, G. J. (2014). The quality of toddler childcare and cognitive skills at 24 months: Propensity score analysis results from the ECLS-B. *Early Childhood Research Quarterly, 29*, 12-21. doi:10.1016/j.ecresq.2013.09.002
- Sabah, S., & Hammouri, H. (2010). Does subject matter matter? Estimating the impact of instructional practices and resources on student achievement in science and mathematics: Findings from TIMSS 2007. *Evaluation & Research in Education, 23*, 287-299. doi:10.1080/09500790.2010.509782
- Serpell, R. (1993). *The significance of schooling: Life-journeys in an African society*. Cambridge, United Kingdom: Cambridge University Press.
- Shen, C. (2002). Revisiting the relationship between students' achievement and their self-perceptions: A cross-national analysis based on TIMSS 1999 data. *Assessment in Education: Principles, Policy & Practice, 9*, 161-184. doi:10.1080/0969594022000001913
- Shen, C., & Pedulla, J. J. (2000). The relationship between students' achievement and their self-perception of competence and rigour of mathematics and science: A cross-national analysis. *Assessment in Education: Principles, Policy & Practice, 7*, 237-253. doi:10.1080/713613335
- Shen, C., & Tam, H. P. (2008). The paradoxical relationship between student achievement and self-perception: a cross-national analysis based on three waves of TIMSS data.

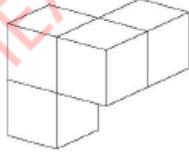
- Educational Research and Evaluation: An International Journal on Theory and Practice*, 14, 87-100. doi:10.1080/13803610801896653
- Song, S., Perry, L. B., & McConney, A. (2014). Explaining the achievement gap between Indigenous and non-Indigenous students: an analysis of PISA 2009 results for Australia and New Zealand. *Educational Research and Evaluation*, 20, 178-198. doi:10.1080/13803611.2014.892432
- Stankov, L., Morony, S., & Lee, Y. P. (2014). Confidence: The best non-cognitive predictor of academic achievement? *Educational Psychology*, 34, 9-28. doi:10.1080/01443410.2013.814194
- Stuart, E. A. (2010). Matching methods for causal inference: A review and a look forward. *Statistical Science*, 25, 1-21. doi:10.1214/09-sts313
- Sullivan, A. L., & Field, S. (2013). Do preschool special education services make a difference in kindergarten reading and mathematics skills?: A propensity score weighting analysis. *Journal of School Psychology*, 51, 243-260. doi:10.1016/j.jsp.2012.12.004
- Temur, Ö. D. (2012). Analysis of prospective classroom teachers' teaching of mathematical modeling and problem solving. *Eurasia Journal of Mathematics, Science & Technology Education*, 8(2), 83-93. doi:10.12973/eurasia.2012.822a
- Van de Vijver, F. J. R., & Hambleton, R. K. (1996). Translating tests: Some practical guidelines. *European Psychologist*, 1, 89-99.
- Van de Vijver, F. J. R., Hofer, J., & Chasiotis, A. (2010). *Methodological aspects of cross-cultural developmental studies*. In M. H. Bornstein (Ed.), *Handbook of cross-cultural developmental science* (pp. 21-37). Mahwah, NJ: Erlbaum.
- Van de Vijver, F. J. R., & Leung, K. (1997). *Methods and data analysis of comparative research*. Thousand Oaks, CA: Sage.
- Van de Vijver, F. J. R., & Poortinga, Y. H. (1997). Towards an integrated analysis of bias in cross-cultural assessment. *European Journal of Psychological Assessment*, 13, 29-37. doi:10.1027/1015-5759.13.1.29
- Van de Vijver, F. J. R., & Tanzer, N. K. (2004). Bias and equivalence in cross-cultural assessment: an overview. *Revue Européenne de Psychologie Appliquée/European Review of Applied Psychology*, 54, 119-135. doi:10.1016/j.erap.2003.12.004
- Van Schilt-Mol, T. M. M. L. (2007). *Differential Item Functioning en item bias in de Cito-Eindtoets Basisonderwijs [Differential Item Functioning and item bias in the Cito-Eindtoets Basisonderwijs]*. Amsterdam, the Netherlands: Aksant.
- Von Secker, C. E., & Lissitz, R. W. (1999). Estimating the impact of instructional practices on student achievement in science. *Journal of Research in Science Teaching*, 36, 1110-1126.
- Walberg, H. J. (1981). A psychological theory of educational productivity. In: F. H. Farley & N. Gordon (Eds.), *Psychology and education* (pp. 81-108). Berkeley, CA: McCutchan.
- Walberg, H. J. (1984). Improving the productivity of America's schools. *Educational Leadership*, 41(8), 19-27.
- Walberg, H. J. (2004). Improving educational productivity: An assessment of extant research. *The LSS Review*, 3(2), 11-14.
- Wu, A. D., Li, Z., & 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(3), 1-26.
- Young, D., Reynolds, A., & Walberg, H. J. (1996). Science achievement and educational productivity: A hierarchical linear model. *Journal of Educational Research*, 89, 272-287. doi:10.1080/00220671.1996.9941328



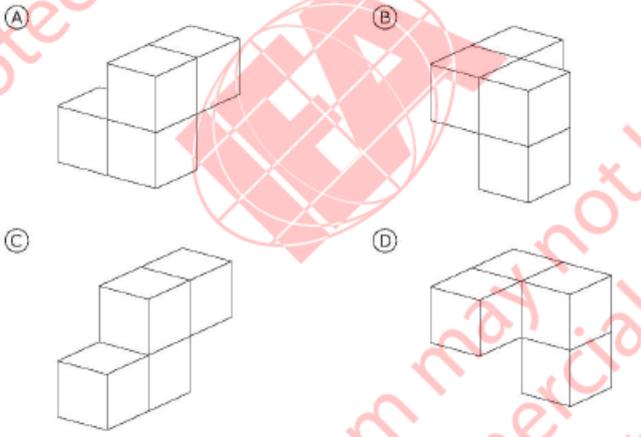
APPENDIX A: Items Showing DIF

ITEM M022049 English version

This object will be turned to a different position

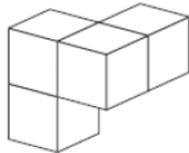


Which of these could be the object after being turned?

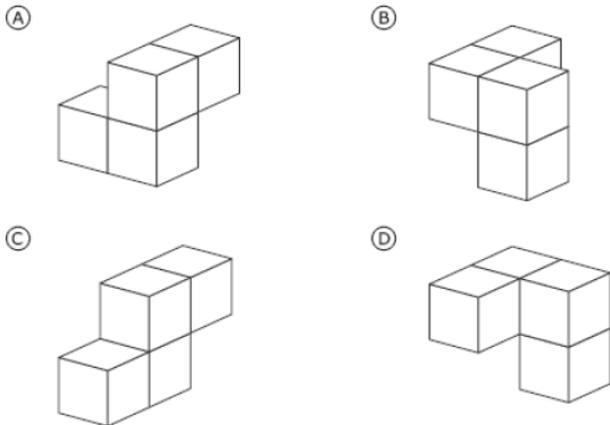


ITEM M022049 Turkish version

Bu şekil farklı bir konuma döndürülecektir.



Aşağıdakilerden hangisi bu şeklin döndürüldükten sonraki konumu olabilir?

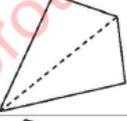
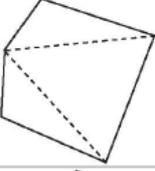
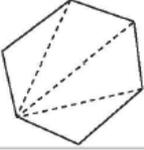


ITEM M042301A English version

Interior Angles

Jackson was investigating the properties of polygons. Jackson made up the table below to see if he could find a connection between sides and angles.

A. Complete the table by filling in the blank spaces.

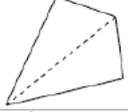
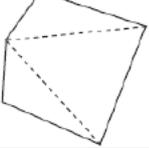
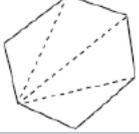
Polygon	Number of Sides	Number of Triangles	Sum of Interior Angles
	3	1	$1 \times 180^\circ$
	—	—	— $\times 180^\circ$
	—	—	— $\times 180^\circ$
	—	—	— $\times 180^\circ$

ITEM M042301A Turkish version

İç Açılar

Kağan, çokgenlerin özelliklerini inceliyordu. Çokgenlerde kenarlar ve açılar arasında bir bağlantı olup olmadığını ortaya çıkarmak için aşağıdaki tabloyu hazırladı.

A. Tablodaki boş yerleri tamamlayınız.

Çokgen	Kenar Sayısı	Üçgenlerin Sayısı	İç Açılar Toplamı
	3	1	$1 \times 180^\circ$
	—	—	— $\times 180^\circ$
	—	—	— $\times 180^\circ$
	—	—	— $\times 180^\circ$