Eğitim ve Bilim
2014, Cilt 39, Sayı 172
Geniş Ölçekli Test Uygulamaları Özel Sayısı

Education and Science
2014, Vol. 39, No 172
Large-Scale Assessment Special Issue

Comparability between the American and Turkish Versions of the TIMSS Mathematics Test Results

TIMSS Matematik Test Sonuçlarının Amerika ve Türkiye Arasında Karşılaştırılabilirliği

Rubab G. ARIM[1]
Ottawa Hospital Research Institute

Kadriye ERCİKAN[2]
University of British Columbia

Abstract

This study examined the degree of comparability between two versions of the Trends in International Mathematics and Science Study's 1999 mathematics test results from the United States of America and Turkey. Measurement invariance was assessed between the two language versions of tests using differential item functioning analyses and exploratory factor analyses. The impact of the differences on the score scale comparability was also examined by comparing the test characteristic curves. Approximately 23% of the items were identified as differentially functioning for the two countries. The factor analyses indicated differences in the structure of the two tests. However, the effect of these differences on score scale comparability was minimal as was demonstrated by very similar test characteristic curves for the two language versions.

Keywords: TIMSS, measurement comparability, adaptation effects, differential item functioning, test characteristic curves

Öz

Bu çalışmada, Amerika ve Türkiye'de elde edilen 1999 Uluslararası Matematik ve Fen Eğilimleri Araştırması matematik test sonuçlarının ne ölçüde karşılaştırılabilir olduğu ele alınmıştır. Ölçme değişmezliği farklı işleyen madde analizleri ve açıklayıcı faktör analizleriyle incelenmiştir. Bu düzeyde görülen farklılıkların puanlama ölçeğine etkisi ise test karakteristik eğrileri karşılaştırılarak incelenmiştir. Matematik testindeki maddelerin yaklaşık %23'ünün bu iki ülke arasında farklı işlediği belirlenmiştir. Diğer yandan faktör analiz sonuçları testlerin yapıları arasında da farklılık olduğunu göstermiştir. Bununla birlikte, iki farklı dildeki testlere ait test karakteristik eğrileri incelendiğinde bu farklılıkların puanlama ölçeğine etkisinin oldukça düşük olduğu görülmüştür.

Anahtar Sözcükler: TIMSS, ölçme karşılaştırılabilirliği, uyarlama etkisi, farklı işleyen madde analizi, test karakteristik eğrileri

[1] Rubab G. Arim, PhD; Ottawa Hospital Research Institute, 501 Smyth Rd. Ottawa, O.N. Canada K1H 8L6; rarim@ohri.ca

[2] Kadriye Ercikan, PhD, Professor; Dept. of ECPS, 2125 Main Mall, University of British Columbia, Vancouver, B.C. Canada V6T 1Z4; kadriye.ercikan@ubc.ca

Introduction

Results from international assessments such as those from the Trends in International Mathematics and Science Study (TIMSS) are used by the countries participating in these assessments for making important policy decisions. One of the key interpretations of international assessment results is comparison of a particular country's performance to those of other countries. Valid performance comparisons across countries require comparability of scores across countries (AERA, APA, & NCME, 1999; Ercikan, 1998; 2002; 2003; Ercikan & Koh, 2005; Hambleton, Merenda, & Spielberger, 2005). There are many factors, such as curricular, cultural, and language differences between countries, which may affect this comparability. In international assessments, one obvious source of differences between scores from different country administrations is the different language versions of the test forms. Previous research has provided overwhelming evidence that different language versions of tests cannot be assumed to be comparable (Angoff & Cook, 1988; Berberoglu & Sireci, 1996; Ercikan, 1998; Ercikan, Gierl, McCreith, Puhan, & Koh, 2004; Ercikan & Koh, 2005; Ercikan & McCreith, 2002; Gierl, Rogers, & Klinger, 1999; Hambleton et al., 2005; Sireci, Fitzgerald, & Xing, 1998; van de Vijver & Tanzer, 1998). These studies emphasized the importance of examining comparability of test forms in different languages at the item level using *differential item functioning* (DIF) analysis methods and at the test level using statistical analyses that compare test data structure, such as exploratory or confirmatory factor analysis (CFA), and multidimensional scaling (Ercikan, Simon, & Oliveri, 2013; Sireci et al., 1998). An item is identified as displaying DIF when the item has varying psychometric properties for different groups, after controlling for differences in the abilities of these groups (Angoff, 1993). The DIF methods address whether measurement invariance at the item level holds for the comparison groups. Statistical methods used to examine comparability at the test level, such as factor analysis and multidimensional scaling, help determine if the test items have similar relationships with each other and the overall construct being assessed by the test.

The item level and test level comparability address item and construct bias. Item bias includes incomparability of test items due to translation/adaptation effects, differential familiarity with item context and content (Ercikan, 1998; Ercikan & McCreith, 2002; Hambleton et al., 2005). Construct bias on the other hand includes conceptual inequivalence of the construct in different cultures, inconsistency in theoretical definitions or inconsistency in the measurement of the construct across cultures (Ercikan & Lyons-Thomas, 2013; Geisinger, 1994; Hambleton, 1993, 1994; 2005; Hui & Triandis, 1985; Oliveri, Olson, Ercikan, & Zumbo, 2012; Reise, Widaman, & Pugh, 1993; Sireci, Bastari, & Allalouf, 1998; van de Vijver & Tanzer, 1997). Research has identified one additional source of bias in comparability of scores as method bias (Sireci, Patsula, & Hambleton, 2005). Method bias includes differences in test administration procedures, as well as differential familiarity of examinees with item and test formats. All three types of bias are important in examining score comparability. While examinee response patterns may be used for evaluating item and construct bias using statistical methods such as DIF and factor analysis, additional data are needed to examine method bias. In order to examine method bias, data on how the test was administered, whether the examinees in each country had similar levels of familiarity with test format are needed among many other factors that may affect score comparability due to test administration and format. The current study focused on item and construct bias in TIMSS 1999 mathematics tests administered in Turkey and the United States of America (USA). Method bias could not be examined because of the lack of information on the test administration and format in two countries.

A final important remark should be made with respect to the score comparisons across countries. The comparisons are usually not performed at particular examinee score levels, but rather, using score distributions across countries. Such score comparisons require a more lenient level of comparability and is referred to as scalar comparability (Cook, 2006; Sireci, 1997). The scalar comparability can be examined by comparing the scores that would be assigned to examinees at the same latent ability level who took different versions of tests. In the item response theory (IRT) framework, scores are assigned to examinees using the test characteristic curves (TCCs) that are defined for each θ level as the sum of item characteristic curves. In this study, in addition to the DIF analyses and factor analyses, the TCC comparisons were conducted to examine the extent of differences at the test score level.

Method

Overview of TIMSS 1999

The TIMSS 1999 was originally prepared in English and was translated into 33 languages. The adaptation procedures are described in the TIMSS User Guide (Gonzales & Miles, 2001). TIMSS data were first collected in 1994 -1995. To measure trends in student achievement, the eighth grade assessments were administered as a follow up in 1999. Therefore, TIMSS 1999 is also known as TIMSS-Repeat (TIMSS-R). Thirty-eight countries participated in the TIMSS-R. The resultant database is complex, containing the following questionnaires for each country: (a) students' responses to cognitive mathematics and science items, (b) students' responses to the background questionnaire, (c) teachers' responses to the background questionnaire, and (d) principals' responses to the background questionnaire (Gonzales & Miles, 2001). The survey was administered from September to November 1998 in the Southern Hemisphere countries and from February to May 1999 in the Northern Hemisphere countries.

The main objective of the TIMSS-R survey was to measure student achievement in mathematics and science in the participating countries. The mathematics test was comprised of 162 items representing 5 separate content areas: fractions and number sense (38%), measurement (15%), data representation, analysis and probability (13%), geometry (13%), and algebra (22%). The performance expectations for the mathematics portion of the TIMSS-R were the following: knowing (19%), using routine procedures (23%), using complex procedures (24%), investigating and solving problems (31%), and communicating and reasoning (2%). Approximately one quarter of the items in the mathematics tests was in the free response format.

Data

TIMSS-R mathematics assessment data for a sample from the USA (n = 8815) and a sample from Turkey (n = 7738) were used in this study. The items were divided into 22 mutually exclusive clusters, labeled A to V. These item clusters were then grouped to create eight overlapping booklets, each containing up to 7 item clusters. Table 1 presents the order of the item clusters in all the booklets.

Table 1.
Order of Item Clusters in Each of the Eight Booklets

	Booklets							
Cluster Label	1	2	3	4	5	6	7	8
A	2	2	2	2	2	2	2	2
B	1				5		3	1
C	3	1				5		
D		3	1				5	
E	5		3	1				
F		5		3	1			
G			5		3	1		
H				5		3	1	
I	6							
J		6						
K			6					
L				6				
M					6			
N						6		
O							6	
P								6
Q								3
R								5
S	4							
T	7		4					
U			7		4			
V					7		4	

For example, item cluster A appeared as the second cluster in all of the booklets. In addition, Booklet 1 consisted of item clusters A, B, C, E, I, S, and T. This table is adapted from the TIMSS User Guide (Gonzales & Miles, 2001, p. 39). There were 33 to 45 items in each booklet. Only one booklet was administered to any given student.

Analysis Procedures

Two statistical methods were used to examine the comparability of scores for the two countries: DIF and the exploratory factor analysis (EFA). Analyses of DIF identify test items that have different psychometric properties for the comparison groups, therefore provides information about the degree to which items assess similar constructs for the two countries. Two statistical procedures that are sensitive to both uniform and non-uniform DIF were used to identify DIF items. The first DIF detection procedure was based on logistic regression (Swaminathan & Rogers, 1990). The second procedure was an IRT based DIF procedure that utilizes the Linn-Harnisch method in capturing differences in item characteristic curves for the comparison groups (Ercikan, 2003; Linn & Harnisch, 1981). The two DIF methods were utilized for verifying the DIF status of items.

A common method of evaluating the construct equivalence of assessments across different language groups is exploratory factor analysis (EFA; Butcher & Garcia, 1978; Poortinga, 1991; Sireci, Bastari, Xing, Allalouf, & Fitzgerald, 2003; van de Vijver & Poortinga, 1991), which provides information about the interrelationships among test items that can be used for examining the comparability of constructs assessed at the test level. Due to the explanatory nature of the research focus, EFA was selected over CFA. Moreover, this study focuses on the degree of comparability of the test data structure rather than whether there is statistically significant evidence of differences in test data structure. Furthermore, there is not a theoretical basis for testing the fit of the data to a pre-

determined model. Consequently, EFA (Butcher & Garcia, 1978; Poortinga, 1991; van de Vijver & Poortinga, 1991) and the DIF methods (Cook, 1996; Ercikan, 1998, Ercikan & Koh, 2005; Hambleton, 2003; Sireci & Berberoglu, 2000) have been used in examining score comparability for different language groups. The analyses were conducted separately for each booklet.

DIF detection procedures. The logistic regression (Swaminathan & Rogers, 1990) procedure states that the standard logistic regression model for predicting the probability of a correct response to an item is

$$P(u=1\mid\theta)=\frac{e^{(\beta_0+\beta_1\theta)}}{[1+e^{(\beta_0+\beta_1\theta)}]},$$

where u is the response to the item, θ is the observed ability of an individual, β_0 is the intercept parameter, and β_1 is the slope parameter. Separate equations for comparison groups can be specified to identify DIF. To test uniform DIF (i.e., group membership) and non-uniform DIF (i.e., the interaction between group membership and the ability level), this logistic regression model is reformulated as follows:

$$P(u=1)=\frac{e^{z}}{[1+e^{z}]},$$

where

$$z=\tau_0+\tau_1\theta+\tau_2 g+\tau_3(\theta g).$$

In this model, g represents group membership, τ_2 corresponds to the group difference, and τ_3 refers to the interaction between group and ability. If τ_2 is nonzero while τ_3 is zero, uniform DIF is concluded. If τ_3 is nonzero, whether or not τ_2 is zero, we can conclude non-uniform DIF. The null hypotheses are $\tau_2 = 0$ and $\tau_3 = 0$ against the alternative hypotheses that are $\tau_2 \neq 0$ and $\tau_3 \neq 0$. This procedure was implemented by using the EZDIF program developed by Niels Waller for the analysis of DIF items (see Waller, 1998).

In the second procedure, an application of Linn-Harnisch statistic (Linn & Harnisch, 1981) to IRT based parameters was used (Oliveri & Ercikan, 2011). The observed and expected mean responses, and the difference between them are computed for each item. The expected values are calculated using the IRT parameter estimates obtained from the entire sample and the ability estimates for the members of the specified subgroup. The differences between observed and expected mean responses are used to compute a chi-square statistic. For large sample sizes (greater than 30), the chi-square statistics with k degrees of freedom can be approximated by the standard normal distribution using $Z_p = (\chi_p^2 - k) / \text{sqrt}(2k)$, where Z_p is the pth percentile of the standard normal distribution. Thus, items with Z-statistic ≥ 2.58 were identified as functioning significantly differently for the two comparison groups at $\alpha = 0.005$ level. In this procedure, both the three-parameter logistic model (3PL; Lord, 1980), and the two-parameter partial credit model (2PPC; Yen, 1993) were used to calibrate multiple-choice items and open-ended items, respectively. This procedure was implemented using IRT calibration software PARDUX (Burket, 1991).

In the IRT-based Linn-Harnisch DIF detection method, it is possible to combine all test booklets in a single analysis to detect DIF; however, since the logistic regression approach requires the use of total test score for each examinee and each examinee completes only one test booklet, DIF analyses could only be conducted separately for each booklet. In order to use DIF identification from both procedures to verify the DIF status of items, the Linn-Harnisch analyses were also conducted separately for each booklet. In this study, an item was considered differentially functioning, only if it was identified as such by both procedures, except for polytomous items that can only be analyzed by the Linn-Harnisch method.

Exploratory factor analysis. A separate factor analysis was conducted for each booklet, using Promax rotation to compare the factor structure in the two samples. For oblique rotations, the Promax rotation is appropriate because it provides a simple structure by allowing the factors to be correlated (Kim & Mueller, 1978). Given that the items assessing different subscales of mathematics achievement, such as fractions and problem solving, they are expected to be highly correlated. Promax rotation was used to accommodate these correlations. The factor analyses were conducted in two steps. In the first step, the number of factors was determined in each test. If different numbers of factors are obtained, the factor loadings cannot be directly compared for items in the two language versions of test booklets. Therefore, a second set of factor analyses was conducted after constraining the number of factors to the smallest number obtained between the two groups.

Test characteristic curve comparisons. The TCCs depict the relationship between the estimated θ scores and the expected raw scores of the examinees based on the item parameters estimated for each comparison group, thus, they can be used to examine the effect of differences in psychometric properties of DIF items on score scale comparability. In fact, while DIF analyses provide comparisons of item parameters and item characteristic curves, differences in TCCs reflect an accumulation of DIF across test items. It is important to note that item parameters based on separate sample calibrations, which is the case when separate country comparisons are conducted, cannot be expected to be on the same scale. However, by using a *linking* with some common anchor items, the two sets of item parameters can be put on the same scale. In such linking analyses, anchor items are typically the ones that can be assumed to be measuring the same construct across languages. In international assessments, there are no anchor items because all the items are in different languages. Hence, the linking can only be done by using test items that can reasonably be expected to function the same way for the two language groups. In general, items that do not display DIF can be used as linking items. In this study, we used non-DIF items in each pairwise comparison to do the linking between the English and Turkish versions of tests. Specifically, a Stocking-and-Lord (S-L; Stocking & Lord, 1983) linking procedure, which uses a linear transformation to place item parameters on the same scale, was used to link the Turkey score scale to the USA score scale. Next, the TCCs were plotted and compared.

Results

Descriptive Statistics

Descriptive analyses were conducted to examine the relative performance levels of the students from the two countries and to estimate the reliability of tests using Cronbach's coefficient-α (Cronbach, 1951). The findings suggested that the USA sample performed significantly better than the Turkish sample and that the reliability estimates were lower for the Turkish version of the test consistently for all booklets. The differences in raw scores were as large as one standard deviation for some of the booklets even though the differences in coefficient-α were small, equal or less than 0.1 (see Table 2).

Table 2.
Descriptive Statistics for the USA and Turkish Samples on the TIMSS 1999 Mathematics Test Booklets

Booklet #	Sample	*N*	# items	*M**	*SD*	α
1	USA	1104	45	21.60	8.97	0.91
	Turkish	961	45	15.35	7.22	0.87
2	USA	1115	33	18.81	6.87	0.87
	Turkish	974	33	14.49	5.41	0.77
3	USA	1110	42	21.45	8.42	0.91
	Turkish	962	42	15.58	6.74	0.86
4	USA	1119	34	19.75	7.20	0.86
	Turkish	966	34	15.02	6.33	0.83
5	USA	1117	42	21.77	8.19	0.92
	Turkish	960	42	14.59	6.79	0.88
6	USA	1100	33	20.65	6.80	0.87
	Turkish	967	33	15.36	6.78	0.86
7	USA	1080	39	22.73	8.39	0.92
	Turkish	969	39	15.56	6.42	0.84
8	USA	1070	41	21.86	8.08	0.89
	Turkish	979	41	16.42	7.17	0.85

Note. *Mean differences between the two countries were significant for all booklets.

DIF Analyses Results

As can be seen in Table 3, using the logistic regression procedure, more non-uniform DIF items were found than uniform DIF items, in all but two booklets. This finding suggests that the difference in the probability of obtaining a correct answer for the two groups differed across ability levels. In other words, these DIF items favored the Turkish sample at some ability levels, but favored the USA sample at other ability levels. The percentage of DIF items identified by the logistic regression procedure ranged from 24% to 54% across the 8 booklets, with the lowest percentage being identified in Booklet 2 and the highest percentage being identified in Booklet 7.

Table 3.
Classification of DIF Items in All Booklets Identified By Logistic Regression Procedure

Booklet #	Not DIF	Uniform DIF	Non-Uniform DIF
Booklet 1	25	8	10
Booklet 2	25	3	5
Booklet 3	23	7	8
Booklet 4	23	5	6
Booklet 5	23	6	8
Booklet 6	23	6	5
Booklet 7	17	11	9
Booklet 8	25	7	9

The Linn-Harnisch procedure allowed examining more items since the logistic regression could not be used to analyze the polytomous items. As can be seen in Table 4, the percentage of DIF items identified by the Linn-Harnisch procedure ranged from 37% to 69% across the 8 booklets, with the lowest percentage being identified in Booklet 1 and the highest percentage being identified in Booklet 4. Moreover, a comparison of which DIF items favored the reference group (i.e., the USA sample) and which ones favored the focal group (i.e., the Turkish sample) revealed that approximately equal numbers of DIF items favored each of the two groups, for all booklets except Booklet 4, where DIF items that were identified as biased against the reference group was twice more than the items that were biased against the focal group. Overall, half of the DIF items (36 out of 70) were in favor of the USA sample, while the other half was in favor of the Turkish sample.

Table 4.
Classification of DIF Items in All Booklets Identified By Linn-Harnisch Procedure

Booklet #	Not DIF	DIF in favour of USA sample	DIF in favour of Turkish sample
Booklet 1	28	9	8
Booklet 2	15	8	10
Booklet 3	20	11	11
Booklet 4	10	8	15
Booklet 5	22	9	11
Booklet 6	16	8	9
Booklet 7	19	9	11
Booklet 8	21	9	11

The results from the two DIF procedures are summarized in Table 5. Approximately 23% of the items were identified as displaying DIF by both methods. When the mathematics topic area of the DIF items was examined, more than half of the Fractions and Number Sense DIF items were in favor of the Turkish sample, and there were more Algebra DIF items that were in favor of the USA sample in some of the booklets. However, it should be noted that these patterns were not consistent across all 8 booklets.

Table 5.
Comparison of Number of DIF Items in All Booklets Identified By the Two Differential Item Functioning Procedures: Logistic Regression (LR) and Linn-Harnisch (LH)

		LR		
	LH	Not DIF	DIF	Polytomous*
Booklet 1	Not DIF	16	10	1
	DIF	9	8	1
Booklet 2	Not DIF	13	2	0
	DIF	12	6	0
Booklet 3	Not DIF	2	10	2
	DIF	20	7	2
Booklet 4	Not DIF	7	3	0
	DIF	16	8	0
Booklet 5	Not DIF	12	8	4
	DIF	11	6	1
Booklet 6	Not DIF	12	4	0
	DIF	11	6	0
Booklet 7	Not DIF	10	8	2
	DIF	7	12	0
Booklet 8	Not DIF	17	4	0
	DIF	8	12	0

Note. *Polytomous items were only analyzed by the LH method.

Exploratory Factor Analysis Results

To further examine the comparability of the tests, the factor structure in the two samples was examined by performing separate factor analyses via maximum likelihood estimation for each booklet, using Promax rotation. Three criteria were used in determining the number of factors: (1) the eigenvalue criterion; (2) scree test criterion; and (3) percent of variance explained criterion (Hair, Anderson, Tatham, & Black, 1998). The results are summarized in Table 6. The findings revealed different numbers of factors in the two countries, indicating that the two versions of the tests were neither unidimensional nor equivalent in any of the booklets. In addition, there were consistently more factors in the Turkish-language test. The number of factors ranged from 6 to 10 for the USA sample, and from 8 to 13 for the Turkish sample. The cumulative percentage of explained variance tended to be slightly higher for the Turkish group ranging from 38.2% to 45.4% for the USA sample, and from 43.7% to 48.4% for the Turkish sample. When factor loadings from the second set of factor analyses, where number of factors were constrained to be equal (lowest number identified between the two countries) were examined, the findings revealed differences between loadings in all of the booklets. For example, when the items with the largest loadings in each factor were considered, for booklets 1-8, only 10, 11, 4, 11, 7, 8, 9, and 5 items out of 45, 33, 42, 34, 42, 33, 39, and 41 items in each booklet, respectively, loaded on the same factors. Overall, only 15 out of 70 items (across 8 booklets) loaded on the same factors. This examination of the factor loadings of the items provided additional evidence that the factor structure of the two tests were non-equivalent.

Table 6.
The Number of Factors and the Percentage of Explained Variance in Each Booklet

	# of Factors		% of Explained Variance	
	USA	Turkish	USA	Turkish
Booklet 1	10	13	45.4	48.2
Booklet 2	7	11	40.4	47.7
Booklet 3	7	11	39.9	45.4
Booklet 4	6	9	38.2	43.8
Booklet 5	8	12	43.9	48.4
Booklet 6	6	8	38.7	43.7
Booklet 7	6	11	40.5	46.5
Booklet 8	10	12	45.0	47.4

Comparisons Based on Test Characteristic Curves

The TCC comparison analyses focused on two booklets: Booklet 4 and Booklet 8, which indicated different results based on the DIF and exploratory factor analyses. For example, Booklet 4 had the highest percentage of common items loading on the same factor, whereas Booklet 8 had the lowest percentage of common items that loaded on the same factor in both tests. The two booklets had typical level of DIF identification as was observed across all the other booklets, 26% DIF items in Booklet 4 and 29% DIF items in Booklet 8. As can be seen in Figures 1 and 2, the differences between TCCs for the USA and Turkish samples were minimal. For Booklet 4, the differences were less than 1 raw sore point for most of the points on the theta scale and the largest was 1.4 out of a maximum of 32 score points (based on multiple-choice items only). For Booklet 8, the greater differences were observed at the high end of the scale. These differences were in the range of 1 to 2 raw scores, out of maximum of 36 (based on dichotomous items only). Overall, there were small TCC differences.

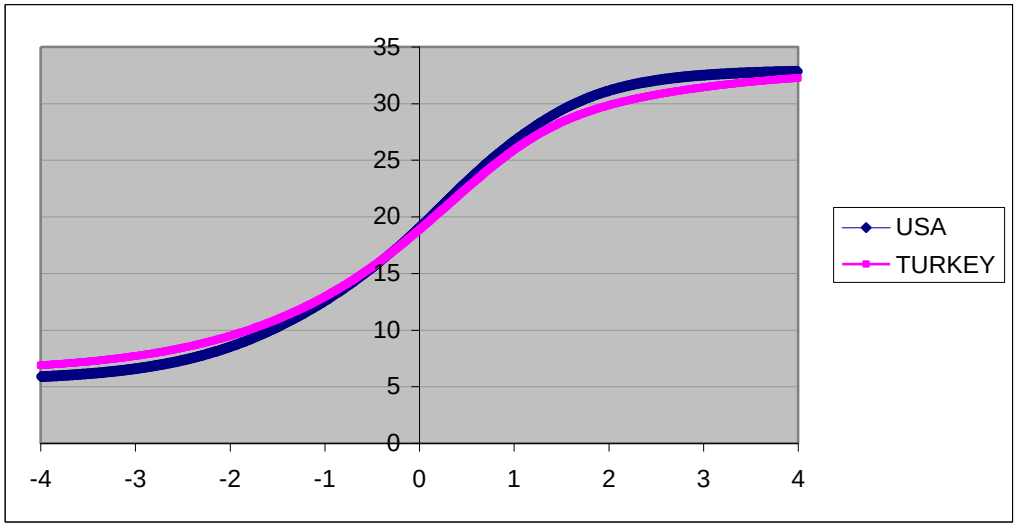

Figure 1. Comparison of the Test Characteristic Curves for the USA and Turkish Samples-Booklet 4

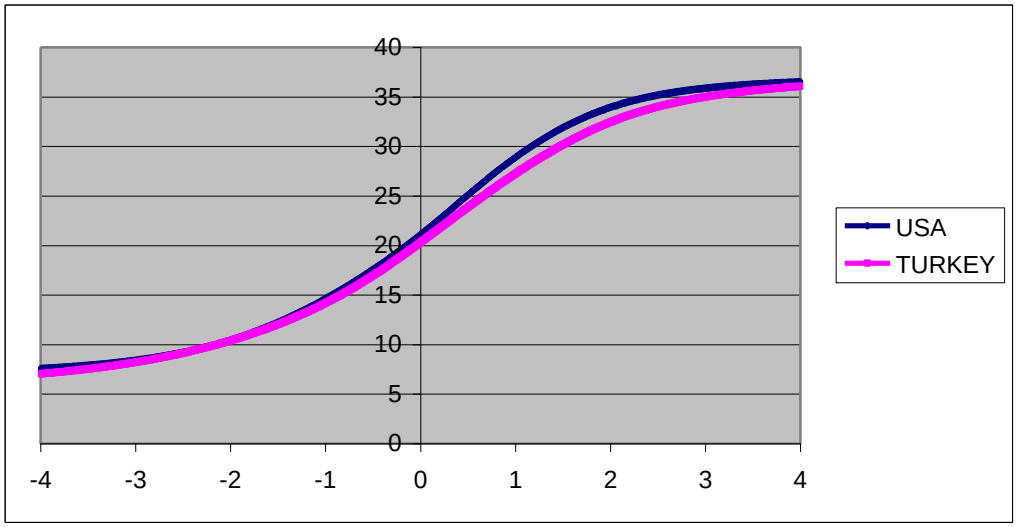

Figure 2. Comparison of the Test Characteristic Curves for the USA and Turkish Samples-Booklet 8

Discussion

This study examined item and construct comparability in the TIMSS-R USA and Turkey 1999 test administrations. Considerable differences were found between the two language versions of the mathematics test. Approximately 23% of the items were identified as differentially functioning for the two groups and the factor analyses suggested that the structures of the tests were different. The directions of the DIF items were approximately evenly distributed between the two groups. Therefore, the overall effect of DIF on score scale comparability, as was demonstrated by the TCC comparisons, was minimal. However, we need to interpret TCC comparisons with caution. Linking TCCs is necessary in order to be able to compare across countries. However, a single linear linking that is implemented through Stocking-and-Lord may overcorrect some parts of the scale and may mask potential differences created by DIF as well as exaggerate the differences. Alternative linking methods needs to be explored to verify the score scale differences obtained in this research.

Given the degree of large differences in the factor structure of the test data, the comparability of test scores for the two groups is still questionable. Furthermore, it is important to highlight some of the limitations of DIF methodology demonstrated by previous research (Ercikan, Roth, Simon, Lyons-Thomas, & Sandilands, in press; Ercikan & Oliveri, in press). These researchers demonstrated great degrees of inaccuracies, in particular underestimation of DIF, when there is great degree of diversity within the comparison groups. These researchers suggest methods such as using latent class modeling to examine possible heterogeneity within groups before conducting DIF analyses.

There are many factors that complicate the implications of these findings. First, the factor analyses clearly demonstrated multidimensionality of the test data for both groups. Both the IRT based DIF analyses and the logistic regression assume essential unidimensionality of the data. Even though the fit statistics in the IRT analyses indicated good fit of the test items to the unidimensional model, the DIF detection and the direction of DIF may have been affected by the multidimensionality in the data. Since TCC comparisons are simply comparisons of accumulated DIF across items in the test, they will be affected similarly. In essence, the similarity of TCCs indicates that neither of the countries is biased against in this pairwise comparison. However, the score scale comparability is not sufficient to accurately interpret and compare the performance of examinees (Ercikan & Gonzalez, 2008). The differences identified at the item level by DIF analyses and at the test level by factor analyses indicate that the scores imply different things in terms of examinee competence, knowledge and skills.

Conclusion

These results highlight limitations in interpretability of international assessment data for making comparisons between countries highlighted by other researchers (Ercikan, 2009; Ercikan, Roth, & Asil, in press). Differential test data structure between countries implies that learning and performance on different aspects mathematics may vary across countries. Therefore, these findings highlight limitations on the connections that can be made between performance on the assessment and learning that is generalizable across countries. Finally, the incomparability between the American and Turkish versions of TIMSS cannot be fully interpreted without identifying sources of DIF or differences in test data structure (Ercikan, 2006). However, methods such as expert reviews or think aloud protocols for identifying sources of DIF (Ercikan et al., 2010) could not be pursued in this research because the test items have not been released.

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (AERA, APA, & NCME: 1999). *Standards for Educational and Psychological Testing*. Washington, DC: American Psychological Association.

Angoff, W. H. (1993). Perspective on differential item functioning methodology. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 3-24). Hillsdale, NJ: Lawrence Erlbaum.

Angoff, W. H., & Cook, L. L. (1988). *Equating the scores of the Prueba de Aptitud Academica and the Scholastic Aptitude Test* (College Board Report No. 88-2). New York: College Entrance Examination Board.

Berberoglu, G. & Sireci, S. G. (1996). Evaluating translation fidelity using bilingual examinees. *Laboratory of Psychometric and Evaluative Research Report No. 285*. Amherst, MA: University of Massachusetts, School of Education.

Burket, G. (1991). PARDUX [Computer program]. Unpublished.

Butcher, J. N., & Garcia, R. E. (1978). Cross-national application of psychological tests. *The Personnel and Guidance Journal*, *56*, 472-486.

Cook, L. (August, 1996). *Establishing score comparability for tests given in different languages.* Paper presented at the meeting of the American Psychological Association, Toronto, Canada.

Cook, L. (July, 2006). *Practical considerations in linking scores on adapted tests.* Keynote address at the 5th international meeting of the International Test Commission, Brussels, Belgium.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, *16*, 297-334.

CTB/McGraw-Hill (1991). PARDUX. [Computer software]. CTB/McGraw-Hill. Monterey, CA.

Ercikan, K. (1998). Translation Effects in International Assessments. *International Journal of Educational Research*, *29*, 543-553.

Ercikan, K. (2002). Disentangling Sources of Differential Item Functioning in Multilanguage Assessments. *International Journal of Testing, 4*, 199-215.

Ercikan, K. (2003). Are the English and French Versions of the Third International Mathematics and Science Study Administered in Canada Comparable? Effects of Adaptations. *International Journal of Educational Policy, Research and Practice*.

Ercikan, K. (2006). Developments in assessment of student learning and achievement. In P.A. Alexander and P. H. Winne (Eds.), American Psychological Association, Division 15, Handbook of educational psychology, 2nd edition (pp. 929-953). Lawrence Erlbaum.

Ercikan, K. (2009). Limitations in sample to population generalizing. In K. Ercikan & M-W. Roth (Eds.), *Generalizing in Educational Research: Beyond Qualitative and Quantitative Polarization* (pp. 211-235), New York: Routledge.

Ercikan, K., Arim, R.,G., Law, D. M., Lacroix, S., Gagnon, F., & Domene, J. F. (2010). Application of think-aloud protocols in examining sources of differential item functioning. *Educational Measurement: Issues and Practice*, *29*, 24-35. Ercikan, K., Gierl, M. J., McCreith, T., Puhan, G., & Koh, K. (2004). Comparability of Bilingual Versions of Assessments: Sources of Incomparability of English and French Versions of Canada's National Achievement Tests. *Applied Measurement in Education*, *17*, 301-321.

Ercikan, K. & Gonzalez, E. (March, 2008). Score scale comparability in PIRLS. Paper presented at the National Council on Measurement in Education, New York, NY, USA.

Ercikan, K., & Koh, K. (2005). Construct Comparability of the English and French versions of TIMSS. *International Journal of Testing, 5,* 23-35.

Ercikan, K., & Lyons-Thomas, J. (2013). Adapting tests for use in other languages and cultures. In K. Geisinger (Ed.), *APA handbook of testing and assessment in psychology, Volume 3,* (pp. 545-569). American Psychological Association: Washington, DC.

Ercikan, K., & McCreith, T. (2002). Effects of Adaptations on Comparability of Test Items and Test Scores. In D. Robitaille & A. Beaton (Eds.) *Secondary analysis of the TIMSS results: A synthesis of current research* (pp. 391-407). Dordrecht, the Netherlands, Kluwer Academic Publishers.

Ercikan, K. & Oliveri, M. E. (in press). *Are our current methods of investigating test fairness doing justice? Population heterogeneity and DIF analysis* to be published in validity book from proceedings of ETS and NY Teacher's College joint conference.

Ercikan, K., Roth, W-M., Asil, M. (in press). Cautions about uses of international assessments. *Teachers College Record*.

Ercikan, K., Roth, M., Simon, M., Lyons-Thomas, J., & Sandilands, D. (in press). Assessment of linguistic minority students. *Applied Measurement in Education*.

Ercikan, K., Simon, M., & Oliveri, M. E. (2013). *Score comparability of multiple language versions of assessments within jurisdictions.* In M. Simon, K. Ercikan, & M. Rousseau. (Eds.), Improving large-scale assessment in education: Theory, issues and practice. (pp. 110-124). New York: Routledge/Taylor & Francis.

Geisinger, K. F. (1994). Cross-cultural normative assessment: Translation and adaptation issues influencing the normative interpretation of assessment instruments. *Psychological Assessment, 6,* 304-312.

Gierl, M. J., Rogers, W. T., & Klinger, D. (1999). *Consistency between statistical procedures and content reviews for identifying translation DIF*. Paper presented at the annual meeting of the National Council on Measurement in Education, Montréal, Quebec, Canada.

Gonzales, E. J., & Miles, J. A. (Eds.). (2001). *TIMSS 1999 User guide for the international database: IEA's repeat of the Third International Mathematics and Science Study at the Eight Grade.* Retrieved from http://isc.bc.edu/timss1999i/data /bm2_userguide.pdf

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

Hambleton, R. K. (1993). Translating achievement tests for use in cross-cultural studies. *European Journal of Psychological Assessment, 9,* 57-68.

Hambleton, R. K. (1994). Guidelines for adapting educational and psychological tests: A progress report. *European Journal of Psychological Assessment*, 10, 229-244.

Hambleton, R. K. (2003). Advances in translating and adapting educational and psychological tests. *Language Testing, 20,* 127-134.

Hambleton, R. K. (2005). Issues, Designs, and Technical Guıdelines for Adapting Tests into Multıple Languages and Cultures. In R.K. Hambleton, P. Merenda, & C. Spielberger (Eds.). *Adapting educational and psychological tests for cross-cultural assessment* (pp. 93-115). Hillsdale, NJ: Lawrence Erlbaum.

Hambleton, R. K., Merenda, P. F., & Spielberger, C. D. (2005). *Adapting educational and psychological test for cross-cultural assessment.* Mahwah, NJ: Lawrence Erlbaum.

Hui, C. H., & Triandis, H. C. (1985). Individualism-collectivism: A study of cross-cultural researchers. *Journal of Cross-Cultural Psychology*, 17, 225-248.

Kim, J-O., & Mueller, C. W. (1978). Factor analysis: Statistical methods and practical issues: Newbury Park: Sage.

Linn, R. L., & Harnisch, D. L. (1981) Interactions between item content and group measurement on achievement test items. *Journal of Educational Measurement, 18*, 109-118.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, NJ: Erlbaum.

Oliveri, M. & Ercikan, K. (2011). Do different approaches to examining construct comparability lead to similar conclusions? *Applied Measurement in Education, 24,* 349-366.

Oliveri, M., Olson, B., Ercikan, K., & Zumbo, B. (2012). Methodologies for investigating item- and test-level construct comparability in international large-scale assessments. *International Journal of Testing, 12,* 203-223.Poortinga, Y. H. (1991). Conceptual implications of item bias. In P. L. Dann, S. H. Irvine, & J. M. Collis (Eds.). *Advances in computer- based human assessment* (pp. 279-290). Dordecht, Netherlands: Kluwer Academic.

Reise, S. P. Widaman, K. F., & Pugh, R. H. (1993). Confirmatory factor analysis and item response theory: Two approaches for exploring measurement invariance. *Psychological Bulletin, 114,* 552-566.

Sireci, S. G. (1997). Problems and issues in linking assessment across languages. *Educational Measurement: Issues and Practice, 16*, 2-19.

Sireci, S. G., Bastari, B., & Alallouf, A. (1998, August). *Evaluating construct equivalence across adapted tests*. Invited paper presented at the meeting of the American Psychological Association, San Francisco.

Sireci, S. G., Bastari, B., Xing, D., Allalouf, A., & Fitzgerald C. (2003). *Evaluating construct equivalence across tests adapted for use across multiple languages.* Unpublished manuscript. University of Massachusetts at Amherst.

Sireci, S.G. & Berberoglu, G. (2000). Using bilingual respondents to evaluate translated-adapted items. *Applied Measurement in Education, 35,* 229-259.

Sireci, S. G., Fitzgerald, C., & Xing, D. (1998). Adapting credentialing examinations for international uses. *Laboratory of Psychometric and Evaluative research report no. 329*. Amherst, MA: University of Massachusetts, School of Education.

Sireci, S.G., Patsula, L., & Hambleton, R. K. (2005). Statistical methods for identifying flawed items in the test adaptations process. In R. K. Hambleton, P. Merenda, & C. Spielberger (Eds.). *Adapting educational and psychological tests for cross-cultural assessment* (pp. 93-115). Hillsdale, NJ: Lawrence Erlbaum.

Stocking, M. L., & Lord, F. M. (1983). Developing a common metric in item response theory. *Applied Psychological Measurement*, *7*, 201-210.

Swaminathan, H., & Rogers, H. J. (1990). Detecting differential item functioning using logistic regression procedures. *Journal of Educational Measurement, 27,* 361-370.

van de Vijver, F., & Poortinga, Y. H. (1991). Testing across cultures. In R. K. Hambelton, J. Zaal (Eds.), *Advances in educational and psychological testing* (pp. 277-308). Boston: Kluwer Academic.

van de Vijver, F., & Tanzer, N. K. (1997). Bias and equivalence in cross-cultural assessment. *European Review of Applied Psychology, 47*, 263-279.

van de Vijver, F., & Tanzer, N. K. (1998). Bias and equivalence in crosscultural assessment. *European Review of Applied Psychology, 47,* 263-279.

Waller, N. G. (nd). EZDIF software for differential item functioning [Computer software and manual]. Retrieved from http://peabody.vanderbilt.edu/depts/psych_and_hd/ faculty/wallern/

Waller, N. G. (1998). EZDIF: Detection of uniform and nonuniform differential item functioning with the Mantel-Haenszel and Logistic Regression procedures. *Applied Psychological Measurement, 22*, 391.

Yen, W. M. (1993). Scaling performance assessments: Strategies for managing local item dependence. *Journal of Educational Measurement, 30,* 187-214.