Makara Hubs-Asia, 2014, 18(2): 127-139
DOI: 10.7454/mssh.v18i2.3467

Differential Item Functioning: Item Level Analysis of TIMSS Mathematics Test Items Using Australian and Indonesian Database

Fitriati

Mathematics Department, STKIP Bina Bangsa Getsempena, Banda Aceh 23112, Indonesia

E-mail: fitri_kindy@yahoo.com

Abstract

The Trends in International Mathematics and Science Study (TIMSS) aims to provide a broad perspective for evaluating and improving education. This assessment also ranks the participant countries based on their performance and makes inferences about factors affecting achievement and learning. However, the study may not function as it was expected because of differences in curricular, cultural, or language settings among countries. Consequently, this challenges assumptions about measurement equivalency. The present study aims to assess the equivalency of mathematics items on the TIMSS (2007) study across Australian and Indonesia. Students' responses were subjected to Rasch analysis to determine DIF items. The results revealed that many items of mathematics tests are problematic because they showed significant bias. The study also found that Australian students performed better and found mathematics items on the test easier than their Indonesian counterparts did. Several factors such as curricular differences, methods used to solve mathematics problems, availability of textbooks and teachers' quality might explain the existence of DIF between the countries. These findings indicate that serious limitations of using TIMSS results in comparing the performance of students across countries. Thus, further empirical evidence is needed before TIMSS 2007 results can be meaningfully used in research.

Differential Item Functioning: Analisis Butir Soal Matematika Studi TIMSS 2007 dengan Menggunakan Database Australia dan Indonesia

Abstrak

The Trends in International Mathematics and Science Study (TIMSS) bertujuan menyediakan perspektif yang luas dalam mengevaluasi dan meningkatkan mutu pendidikan. TIMSS juga merangking negara-negara peserta studi berdasarkan kemampuan serta membuat prediksi tentang faktor-faktor yang memengaruhi capaian belajar siswa mereka. Akan tetapi, karena perbedaan kurikulum, budaya atau bahasa dari negara-negara tersebut, TIMSS ini tidak berfungsi sebagaimana yang diharapkan. Akibatnya, kondisi ini menantang asumsi-asumsi tentang pengukuran yang ekuivalen. Penelitian ini bertujuan untuk menguji keekuivalenan soal-soal matematika dari studi TIMSS 2007 dengan menggunakan jawaban siswa Australia dan Indonesia. Rasch analisis digunakan untuk menemukan soal-soal yang bias. Hasil analisis menujukkan bahwa banyak soal matematika dalam studi TIMSS 2007 bermasalah karena soal tersebut memperlihatkan bias yang signifikan. Penelitian ini juga menemukan bahwa kemampuan siswa Australia lebih baik dari siswa Indonesia. Soal matematika terlihat lebih mudah bagi siswa Australia dibandingkan bagi siswa Indonesia. Perbedaan kurikulum sekolah, metode dalam pemecahan masalah dan ketersediaan buku dan kualitas guru diduga sebagai faktor penyebab munculnya DIF item. Temuan-temuan dalam penelitian ini mengindikasikan adanya keterbatasan yang serius dalam menggunakan hasil studi TIMSS untuk membandingkan negara-negara peserta studi. Oleh karena itu, bukti-bukti empiris lainnya sangat diperlukan sebelum hasil studi TIMSS 2007 dapat digunakan dengan bermakna sebagai dasar penelitian.

Keywords: DIF items, mathematics test items, Rasch model

Citation:
Fitriati (2014). Differential item functioning: Item level analysis of TIMSS mathematics test items using Australian and Indonesian database. *Makara Hubs-Asia*, *18*(2): 127-139. DOI: 10.7454/mssh.v18i2.3467

Makara Hubs-Asia, 2014, 18(2): 127-139
DOI: 10.7454/mssh.v18i2.3467

1. Introduction

One of the major developments in mathematics education is the growing interest in international comparisons of student achievement. International comparative studies, such as the Trend in International Mathematics and Science Study (TIMSS) and the Programme for International Students Assessment (PISA) were implemented decades ago. TIMSS is an ambitious series of international assessments conducted in nearly 60 countries to measure trends in learning mathematics and science (IEA, 2008). Since the 1960s, this cross-cultural study has been conducted, based on the idea that this assessment can provide a broad perspective for evaluating and improving education. In addition, the participant countries can assess their relative positions in mathematics achievement in relation to their competitors in the global world. Analyzing the data collected from this large-scale comparative study of mathematics achievement may enable us to understand educational processes and to identify new issues relevant to reform movements in the educational system. In addition, analysis within and across countries may determine the link among students' achievement, teachers' instructional practice, and curriculum content. This information then can be used to guide educational decision- making and practice in the area of mathematics (IEA, 2008).

However, to be able to meet the objectives stated above, it is clear that international studies need to confirm the validity and reliability of the test (Wu, 2009). This is urgent because international studies, such as TIMSS, originally used test instruments in English, which then were translated into the language of instruction of the students. Many researchers have argued that adapted tests should possess adequate validity and reliability within each language in order to make valid comparisons across these groups of students (Sireci & Gonzales, 2003; Yildirim, 2006; Chen, Gorin, Thomson, & Tatsuoka, 2008; Wu, 2009). Therefore, the present study on test adaptation meets this need.

Related to test adaptation, the TIMSS (2007) study administered tests in 39 different languages in 59 participating countries. Although TIMSS (2007) implemented rigorous translation verification to achieve maximal linguistic equivalence and to set test items that are simple and context free (IEA, 2008), the test instruments may not function in the same way in all cultures because of differences in curricular, cultural, or language settings among the countries (Sireci & Gonzales, 2003; Ercikan & Koh, 2005; Schulz & Fraillon, 2009; Yildirim 2006; Arim & Ercikan, 2014). Consequently, this international test may not function as expected. Hence, the test may not be equivalent or fair among different cultures. According to Gierl (2000: 281), 'if the construct measured by the two forms is not equivalent, it may change the validity for one set of test scores and adversely influence their comparability, meaning, and interpretability'. Hence, the validity of the score of any translated achievement tests depends on the accuracy of test adaptation, indicating the need for the evaluation of test equivalents to achieve valid test adaptation.

The issues of validity and reliability can be defined from multidimensional perspectives. That is, in the case of international assessment, different groups of participants may have differently distributed multidimensional ability because of differences in language, culture, and curriculum (Ercikan, 1998; Byrne, 2002; Arim & Ercikan, 2014). These differences may cause a test item to function differently between two groups. It has been argued that when test items exhibit Differentiate Item Functioning (DIF), the validity and reliability of the test are not yet achieved (Wu, 2009; Arim & Ercikan, 2014). It is believed that this may affect the equivalence or non-equivalence of the test items. Therefore, the investigation of DIF is required to assure the validity and reliability of the assessment.

Many international comparative studies have been conducted to determine the existence of DIF. For example, Ercikan (1999) reported that 41% of science items from TIMSS displayed moderate or large DIF when Canadian English and French examinees were compared. She also found that 18% of mathematics test items exhibited DIF. Allalouf, Hambelton, and Sireci (1999) found that 42 of 125 verbal items (34%) displayed moderate or large DIF in the Israeli Psychometric Entrance Test when Hebrew and Russian examinees were compared. Yildirim (2006) assessed the Turkish and English versions of TIMSS 1999 and found that the rate of DIF items within the test was high and differential discriminating was an issue. Arim and Ercikan (2014) also found that approximately 23% of mathematics items in a TIMSS (1999) study were identified as functioning differentially in American and Turkish versions. However, few studies have focused on Australian and Indonesian data. Such studies are urgently needed because tests that were administered in both countries were written in different languages. Australian students were tested in the source language of English, whereas Indonesian students were tested in the Bahasa Indonesia version adapted from the source language. DIF-related problems may appear during the process of test translation and adaptation between the languages of both groups. Investigation of the equivalence of English and Bahasa Indonesia versions, in the context of cultural differences, can be minimized. In addition, the performance of students in the eighth grade students in both countries is below the international average (500). DIF analysis may provide some information about the difficulty of test items faced by students in both countries. Therefore, the aim of this

study is to conduct an item-level analysis, in which the test items are investigated through utilizing the DIF method.

Because valid and reliable assessments are not easy to develop (Wu, 2010), the main purpose of this study is to examine the equivalence of mathematics items in TIMSS (2007) across cultures and languages. This study also provides an overview of statistical methods that can be employed to assess flaws in the items caused by test translation in the context of mathematics achievement testing. Several DIF methods seek evidence of the differential performance of subgroups, in order to detect biases. These include item response theory with Rasch model analysis (Hungi, 2005); item response theory with likelihood ratio analysis (IRT-LR) (Yildirim, 2006); and the Mantel-Haenszel (M-H) technique (Yildirim, 2006; Gierl & Khaliq, 2001). However, this study employed only item response theory with Rasch model analysis. The reason that this method was selected is explained in the methods section of this paper.

The current study addresses the following research question: "Do the mathematics items of TIMSS (2007) operate differently between Australian and Indonesian students?" For this purpose, the study will assess responses to Indonesian and Australian TIMSS (2007) mathematics items with respect to the psychometric characteristics of the items. Because this study evaluates the possible presence of item bias caused by test translation, the results of such analyses should provide information that is useful in understanding how differences in items may relate to educational differences across countries. In short, the results of these analyses then might provide some insights into the reasonableness of the assumption that TIMSS (2007) mathematics items are equivalent and fair across countries. Based on previous research on test adaption and test translation within international comparative studies and the appearance of DIF during that process, it is hypothesized that mathematics test items administered for Australian and Indonesian students may function differently.

2. Methods

This study used the TIMSS (2007) mathematics achievement test. A dataset of the test is publicly available on the International Association for the Evaluation of Educational Achievement (IEA) website. The test consists of numerous items designed to collect information about the mathematical ability of students. There are 63 number items, 64 algebra items, 47 geometry items, and 41 data and chance items, which is a total of 215 items. The subjects under the four content areas were as follows: Number area includes whole numbers, fractions and decimals, integers, ratios, proportions and percentage. The algebra areas include patterns, algebraic expressions, equations and formulae, and functions. This included three subject areas of geometry: geometric shapes, geometric measurements, location, and movement. Finally, the section on data and chance included data organization and representation, as well as data interpretation and chance. All aspects of the test content represent the subject matter of school mathematics that is covered by the eighth-grade curriculum in both Australia and Indonesia.

Of 215 items, 81 were classified as measuring knowledge, 88 as measuring application, and 46 as measuring reasoning skills. More than half the items (117) were multiple-choice and the rest (98) were constructed responses (CR) that required students to generate and write their own answers. These mathematics items then were matrix sampled into fourteen booklets. The pool of items was divided into 28 sets of items or cluster. These were then arranged variously to make 14 overlapping test booklets, which were distributed systematically in each classroom. The examinees were administered one of the 14 test booklets.

This present study investigates two booklets—the Booklet 8 and Booklet 9. These booklets were selected because they contain a higher number of test items than the other booklets do, so more items would be investigated. The number of TIMSS (2007) mathematics items by type and reporting category in these booklets is given in Table 1.

Table 1. TIMSS 2007 Mathematics Test Items of Two Booklets by Type and Reporting Category

Reporting Category	Item Type					
	Booklet 8			Booklet 9		
	MC	CR	Total	MC	CR	Total
Number	4	4	8	5	2	7
Algebra	4	7	11	5	4	9
Geometry	7	2	9	6	3	9
Data and Chance	2	2	4	2	4	6
Total	17	15	32	18	13	31

MC=Multiple choice; CR=Constructed response

Makara Hubs-Asia, 2014, 18(2): 127-139
DOI: 10.7454/mssh.v18i2.3467

Thus, the number of possible score points available for the analysis exceeded the number of items, whereas the total score for Booklet 8 and Booklet 9 were 33 and 32, respectively.

For the purposes of this study, a total of 1,178 grade 8 students were included across the two booklets: 578 were Australian students, and 600 were Indonesian students. The examinees were administered one of the two test booklets (Booklet 8 or Booklet 9). The Australian students were tested in the source language of English, whereas the Indonesian students were tested in the Bahasa Indonesia version that was adapted from the source language. The selection of these countries allowed for the investigation of the equivalence of English and Bahasa Indonesia versions when cultural differences were expected to be minimal.

Because many countries, cultures, and language backgrounds were involved in the TIMSS (2007) study, test adaptations play an important role. Hence, TIMSS (2007) followed strict verification procedures to ensure translation equivalence. These procedures were also used to minimize semantic, psychometric, and linguistic differences between the source and translated language versions of the test. TIMSS (2007) instruments were developed in English and then translated into 39 other languages, by following a complex verification procedure of translation and adaptation appropriate for the cultural contexts of participating countries. Professional translators and subject matter experts were involved in ensuring that the meaning and the difficulty of items did not change between the source and target versions. Additionally, a series of statistical checks to detect differences in the performance of the items were conducted (IEA, 2008). A double translation procedure was also used in TIMSS (2007) to ensure that the materials were equivalent across language versions.

Because descriptions of data procedures and rationales for selecting sub-groups of item were given, some statistical and judgmental procedures used in the analyses were also defined. Item response theory with the Rasch Model approach was used in the DIF analyses of the items selected in this study. The Rasch model (Rasch, 1960) was used to determine the equivalence of the test items, particularly in the item-level analysis. The justification for using this model is that Rasch modeling is widely used to measure invariance and determine equivalence across groups of items (Schulz & Fraillon, 2009). Additionally, the Rasch model proposes that responses to a set of items can be explained by a person's ability along a continuum of the unidimensional construct underlying the items and by the characteristics of the items, or item parameters. Several advantages of Rasch measurement have been described (Andrich, 1988; Wright, 1997). A key characteristic of the model is that Rasch measurement can be considered sample independent, as well as instrument independent. That is, if a Rasch model fits a set of data, item characteristics are not dependent upon a specific sample; therefore, item parameters estimated across different groups and contexts will be equivalent (Andrich, 1988). Consequently, the Rasch model can be used to assess the extent to which a set of test items is sample-or context-free (Raczek *et al*., 1998). Rasch procedures also enable the test developer to examine the equivalence of item calibrations across different samples and contexts, including various cultural-linguistic settings and translations. In this case, the Rasch analysis enables a more detailed (item level) examination of the structure and operation of the scales on the tests.

Within Rasch model, DIF analysis will be employed to investigate the items that operate differently across Australian and Indonesian groups. To perform this analysis, the data of mathematics achievement tests from Australian and Indonesia student data set were subjected to Rasch analysis using Conquest 2.0 software (Wu, Adam, Wilson, & Handale, 2007). Inspecting the infit mean squares (IMS) provides evidence of the fit of the data to the model. The infit mean squares are used to determine the fit of the item within the construct. In this study, critical values chosen for the IMS fit statistic were 0.72-1.30 (Linacre, Wright, Gustafsson, & Martin-Lof, 1994). Items where IMS values fall above 1.30 are generally considered misfitting and do not discriminate well, while those below 0.72 are overfitting and provide redundant information (Tilahun, 2004). Additionally, various statistics and probability curves were also used to judge the results. For instance, parameters were estimated separately for each group to determine whether the underlying model fit the data. If the given indicators are equivalent across groups, item bias is not supported (Little, 1997). In detecting biased items, the item threshold approach was also used. As suggested by Hungi (2005), two criteria in this approach are as follows:

a) Items whose differences in threshold (estimate mean) values between two groups are outside a predetermined range. The range is

d_1-d_2>±0.50

where:

d_1=the item's threshold value in group 1, and

d_2=the item's threshold value in group 2.

b) Items whose difference in the standardized item threshold between any of the group fall outside a predefined range. Adam and Khoo (1993) employed the range -2.00 to 2.00:

st (d_1-d_2)>±2.00

3. Results and Discussion

Descriptive summary. Because this study used secondary data, it is important to show the descriptive statistics of the data to describe their condition. Table 2

shows the scale statistics for selected booklets of TIMSS (2007). The results indicate that Australian students performed significantly better than the Indonesian students did in Booklet 8. Although the Australian students also performed better in Booklet 9 than the Indonesian students did, two independent t-tests were conducted to compare the mean analysis, showing that the differences were not significant. In addition, the score distribution for the Indonesian students was found more slightly skewed (1.404) and (1.054) than that of the Australian students (0.472) and (0.232) in both booklets, respectively. These results are in line with the TIMSS (2007) international mathematics report, which showed similar statistical data (Mullis, Martin, & Foy, 2008).

Country differences. DIF analysis was used to investigate the presence of item bias and the significant differences between Australian and Indonesian groups. The number of items in the two selected booklets was subjected to analysis. Two criteria were applied to determine the biased items, which were based on IMS values and significant differences in threshold. Two separate analyses were conducted, and the results of each analysis are presented in the following sub-section.

*Country differences in Booklet 8***.** The 32 items in Booklet 8 were analyzed using the DIF model. This was carried out to test whether the items operate differently between Australian and Indonesian students. The Australian and Indonesian student mean estimates in Booklet 8 were examined. The results are shown in Tables 3 and 4.

The results of the analysis of the IMS of the items (Table 4) showed that most items in Booklet 8 had IMS within an acceptable range (0.72-1.30), with only a few falling outside the range. The results showed the IMS value of five items fell outside the predetermined range in the Australian group, which indicated that the items did not fit the Australian group. However, these items fit the model of the Indonesian group quite well. Similarly, item (m032477=1.33) recoded the IMS value outside the range in the Indonesian group, but the IMS value in the Australian group (m032477=1.29) was within the acceptable range, indicating that the item fit the model of this group.

Table 4 also shows that five items recoded the IMS value outside the desired range in both the Australian and the Indonesian groups. Four of these items (m042248=0.60, 0.70; m042229a=0.59, 0.59; m042229b=0.70, 0.42; m032064=0.67, 0.45) had IMS values below 0.72, which indicates that the items did not fit the model. The IMS of another item (m032662=1.42, 2.30) was above 1.30, which indicated that the items did not fit or discriminate well.

Because these items did not fit the models of either the Australian or the Indonesian group, they were identified as bad items, indicating that the inclusion of these items on the test should be reconsidered. Thus, based on the criterion of item IMS, the results showed that country bias was a problem in the TIMSS (2007) mathematics tests.

Examining the items based on significant differences is also important in determining the existence of DIF within the group. The results in Table 3 show that the Australian students generally performed better and found the items in Booklet 8 relatively easier than the Indonesian students did.

Table 2. Scale Statistics for Mathematics Test of Two TIMSS 2007 Booklets

Scale	B8 (32 items)		B9 (31 items)	
Statistics	Aus	Idn	Aus	Idn
Examinees	289	302	289	298
Mean	49.55	48.47	51.89	50.15
Std. dev	10.3	9.99	9.42	10.31
Skewness	0.472	1.404	0.232	1.054
Kurtosis	-0.669	1.976	-0.632	0.953
Alpha	0.945	0.904	0.947	0.909

These scales were derived from standardized math score (50, 10)

Table 3. General Country Differences in Booklet 8

Country	Estimate	Error	IMS	CI	T
Australia	-0.575	0.053	1.00	(0.84, 1.16)	0.0
Indonesia	0.575	0.053	0.9	(0.84, 1.16)	-1.3

Chi-square test of parameter equality =119.55, df=1, Sig Level=0.000
IMS: Infit mean square; **CI**: Confidence Interval (the estimate will vary from lower value to higher values);
T: Ratio between the estimate and its standard errors (if $|t|>\pm 2$ = estimate is significantly different from 0)

Makara Hubs-Asia, 2014, 18(2): 127-139
DOI: 10.7454/mssh.v18i2.3467

Table 4. Country Differences in Booklet 8

Items	IMS Approach		Threshold Approach				d_1-d_2	SE dif	sd (d_1-d_2)
	Aus	Idn	Aus		Idn				
	IMS	IMS	d_1	SE	d_2	SE			
m042183	0.91	1.09	-0.079	0.094	0.079	0.094	-0.158	0.133	-1.19
m042060	0.85	1.01	-0.364	0.097	0.364	0.097	-0.728[a]	0.137	-5.31[b]
m042019	1.40	0.85	-1.548	0.143	1.548	0.143	-3.096[a]	0.202	-15.31[b]
m042023	0.69	0.93	0.571	0.098	-0.571	0.098	1.142[a]	0.139	8.24[b]
m042197	0.73	0.83	-0.011	0.127	0.011	0.127	-0.022	0.180	-0.12
m042234	0.91	1.09	0.106	0.094	-0.106	0.094	0.212	0.133	1.59
m042066	1.05	1.23	-0.195	0.097	0.195	0.097	-0.390	0.137	-2.84[b]
m042243	0.71	0.79	-0.326	0.110	0.326	0.110	-0.652[a]	0.156	-4.19[b]
m042248	0.60	0.71	0.138	0.125	-0.138	0.125	0.276	0.177	1.56
m042229a	0.59	0.59	0.214	0.171	-0.214	0.171	0.428	0.242	1.77
m042229b	0.70	0.42	-0.460	0.160	0.460	0.160	-0.920[a]	0.226	-4.07[b]
m042080a	1.22	1.21	0.132	0.099	-0.132	0.099	0.264	0.140	1.89
m042080b	1.03	0.96	1.09	0.213	-1.090	0.213	2.180[a]	0.301	7.24[b]
m042120	0.98	1.30	-0.183	0.095	0.183	0.095	-0.366	0.134	-2.72[b]
m042203	0.91	0.98	0.168	0.094	-0.168	0.094	0.336	0.133	2.53[b]
m042264	1.03	0.83	0.183	0.131	-0.183	0.131	0.366	0.185	1.98
m042255	1.05	1.01	0.309	0.093	-0.309	0.093	0.618[a]	0.132	4.70[b]
m042224	1.35	0.85	-0.351	0.097	0.351	0.097	-0.702[a]	0.137	-5.12[b]
m032094	0.94	1.00	0.377	0.091	-0.377	0.091	0.754[a]	0.129	5.86[b]
m032662	1.42	2.30	0.150	0.126	-0.150	0.126	0.300	0.178	1.68
m032064	0.67	0.45	-0.277	0.125	0.277	0.125	-0.554[a]	0.177	-3.13[b]
m032419	1.29	1.33	0.357	0.097	-0.357	0.097	0.714[a]	0.137	5.20[b]
m032477	0.96	1.20	-0.114	0.102	0.114	0.102	-0.228	0.144	1.58
m032538	0.74	0.82	0.372	0.102	-0.372	0.102	0.744[a]	0.144	5.16[b]
m032324	1.29	1.01	0.466	0.101	-0.466	0.101	0.932[a]	0.143	6.52[b]
m032116	1.26	0.75	0.259	0.098	-0.259	0.098	0.518[a]	0.139	3.74[b]
m032100	0.89	1.05	-0.601	0.101	0.601	0.101	-1.202[a]	0.143	-8.42[b]
m032402	1.31	1.24	0.842	0.092	-0.842	0.092	1.684[a]	0.130	12.94[b]
m032734	0.92	1.10	-0.898	0.109	0.898	0.109	-1.796[a]	0.154	-11.65[b]
m032397	1.02	1.10	-0.090	0.094	0.090	0.094	-0.180	0.133	-1.35
m032695	1.38	0.98	0.056	0.063	-0.056	0.063	0.112	0.089	1.26
m032132	1.27	0.84	-0.295		0.295		-0.590[a]		
Separation Reliability=0.948									

IMS Infit mean square; a difference in item difficulty outside the range ±0.50; b st (d_1-d_2) outside the range ±2.00
Australian (N=289); Indonesian (N=305)

The results also showed that the Australian students scored 1.150 lower than the Indonesian students did. The fact that the parameter estimate is more than twice its standard error indicates that this difference is statistically significant (Wu *et al*., 2007). The significant variance within the items is shown in Table 3.

The negative value of difference in item estimate (d_1-d_2), as shown in Table 3, indicates that the item was relatively easier for the Australian students than for the Indonesian students, while positive values implied the opposite. Using this criterion, the analysis found that most items in Booklet 8 apparently favored one group or the other. However, it is important to remember that a mere difference between the estimate values of an item for the Australian and Indonesian groups may not be sufficient evidence to imply bias for or against a particular group. Nevertheless, a difference in item estimates outside the ±0.50 range is large enough to raise a concern. Similarly, differences in standardized difference in item threshold outside the ±2.00 range should raise a concern (Adam & Khoo, 1993; Hungi, 2005). Using this criterion, it is important to note that the standardized DIF for the last item could not be calculated. The standard error of this item was not estimated because the last item was fixed to the average difficulty equal to 0. Therefore, the last item was judged only according to the difference between the groups (d_1-

d_2). This case was also applied in each country's DIF analysis of each booklet in this study.

From the above criteria, 20 items were identified as DIF items because they fell outside the predefined ranges (d_1-d_2>±0.50; and st (d_1-d_2)>±2.00). It was found that 10 items (m042060, m042066, m042019, m042243, m042229b, m042224, m032064, m032100, m032734, and m032132) were markedly easier for the Australian students compared to the Indonesian students. On the other hand, 10 items (m042023, m042080b, m042203, m042255, m032094, m032419, m032538, m032324, m032116, and m032402, were markedly easier for the Indonesian students compared to the Australians students. These items are somewhat problematic because significant variance found in them.

Figure 1 (item m042019) and Figure 2 (item m032734) show that the item characteristic curves (ICC) for Australian students are clearly higher than those of the Indonesians, which means that the Australian students stood greater chances than Indonesian students of getting this item correct at the same ability level. On the contrary, the ICC for Indonesian students for item m042080b (Figure 3) was mostly higher than that of the Australian students. Based on this evidence, it can be concluded that country bias was an issue in Booklet 8.

Country differences in Booklet 9. The DIF analysis was also carried out to examine Booklet 9. The results of the analysis of the 31 items in this booklet, for the examinees in each group, are summarized in Tables 5 and 6. As Table 6 shows, three items appear misfitting or not discriminating well in both groups because their IMS values—m032662 (1.31; 1.64); m042198c (0.67; 0.64); and m042169b (1.35; 1.50)—were outside the acceptable range.

The IMS values in the Australian group also showed that three other items—m03232 (1.65), m042198a (0.63), m042260 (0.70)—fell outside the range (0.72-1.30). However, these items behaved well when the model was fitted to the Indonesian group. Their IMS values—m03232 (1.05), m042198a (0.91), and m042260 (1.29)—fell within the range, indicating that the items fit the model of the Indonesian group. In contrast, the analysis of the IMS values in the Indonesian group found that three items did not fit the model of this group, but they fit the model of the Australian group.

Figure 1. ICC for Item m042019 (Biased in Favor of Australian, d_1-d_2=-3.096)

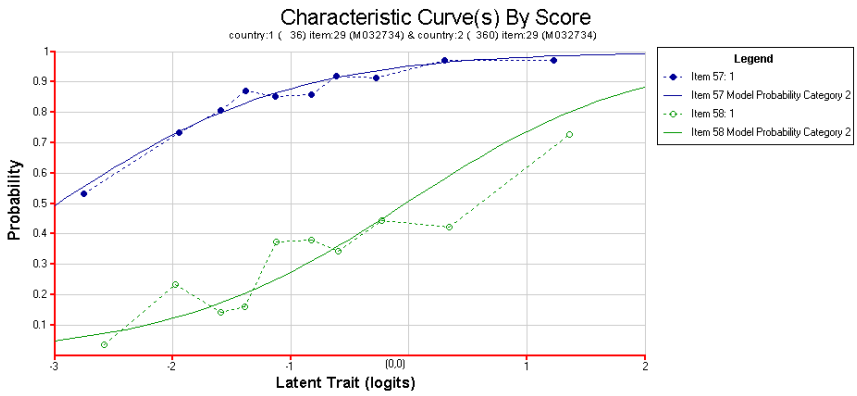

Figure 2. ICC for Item m032734 (Biased in Favor of Australian Students, d_1-d_2=-1.796)

Makara Hubs-Asia, 2014, 18(2): 127-139
DOI: 10.7454/mssh.v18i2.3467

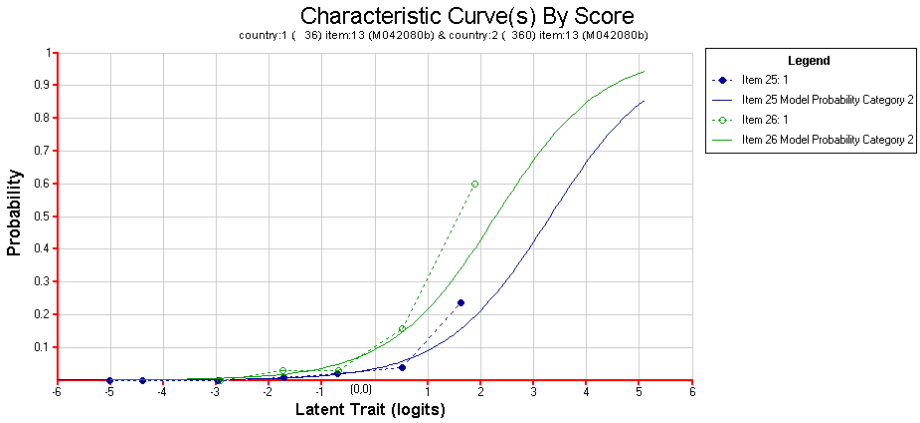

Figure 3. ICC for Item m042080b (Biased in Favor of Indonesian Students, d_1-d_2=2.180)

This is because the IMS value of the items in the Indonesian group—m032064 (0.59), m032477 (1.39), and m042300b (0.61)—fell outside the predetermined range, while the Australian group recoded the IMS values of m032064 (0.74), m032477 (0.86), and m042300b (0.89) within the range. These results indicate that these items are somewhat problematic. Thus, based on the IMS criterion, it is evident that there is a country bias in Booklet 9.

The significant DIF of the items was investigated using the threshold approach. Table 5 shows that 23 items in Booklet 9 showed significant DIF. This can be seen in the differences in the threshold values of these items, which were bigger than ±0.50, and the standardized difference values of the items were also bigger than ±2.00. In addition, 10 of these items were biased in favor of Australian students, which was indicated by the negative values of the difference in item threshold. On other hand, 13 items were biased in favor of the Indonesian students, which was indicated by the positive values of difference in item threshold. These results indicate a significant variance in this item, which is evidence of DIF. Thus, the results showed that most of the test items in Booklet 9 were biased against one group or the other.

The big gap in performance between the students in the two countries is shown in plot ICC of the items that exhibited significant DIF. The plot is illustrated in the Figures 4 and 5. Figure 4 shows that, given a particular ability level, the probability of being successful on this item is higher for Australian students than for Indonesian students, which indicates that the Australian students found this item easier than the Indonesian students did.

However, as shown in in Figure 5, the probability of being successful on this item was higher for the Indonesian students than for the Australians students because both groups were at the same ability level. The Indonesian students found this item easier than Australian students did. Many items Booklet 9 seem somewhat problematic. Therefore, it can be concluded that country bias was a concern in Booklet 9 of the TIMSS (2007) mathematics test.

In this study, the big difference in ability between the Australian and Indonesian groups in the mathematics tests of TIMSS 2007 could be explained by curriculum difference. Although this study did not investigate the degree to which DIF may be caused by curriculum difference, some evidence from the relative distribution of DIF items by content areas in each booklet indicated that some DIF items were affected by curriculum differences (Ercikan, 2002; Ercikan & Koh, 2005; Emenugo & Child, 2005; Yildirim, 2006). These differences include the sequence of mathematics courses or time spent on the topic, teacher classroom practice influenced by teacher academic training, experience, and the material available to them (Emenugo & Child, 2005).

It is assumed that this problem might also exist in the Australian and Indonesian contexts because the mathematics curricula in both countries are different. Therefore, further studies that investigate bias must be carried out, as suggested by Yildirim, Yildirim and Verheslt (2014), who said that when DIF items were detected in the test instrument, the researchers should conduct studies to determine the possible cause of DIF detected in those items.

The relative failure of Indonesian students in achieving most items on the TIMSS (2007), with respect to the Australian students, could be attributed to the ineffectiveness of the curriculum and instructional practices in Indonesia or the limited textbooks or other sources in most Indonesian schools to support student learning.

Table 5. General Country Differences in Booklet 9

Country	Estimate	Error	IMS	CI	T
Australia	-0.677	0.050	1.08	(0.84, 1.16)	0.9
Indonesia	0.677	0.050	0.79	(0.84, 1.16)	-2.8

Chi-square test of parameter equality = 180.60, df=1, Sig Level=0.000
IMS: Infit mean square; **CI**: Confidence Interval (the estimate will vary from lower to higher values);
T: Ratio between the estimate and its standard errors (if $|t|>\pm 2$=estimate is significantly differ from 0)

Table 6. Country Differences in Booklet 9

Items	IMS Approach		Threshold Approach						
	Aus	Idn	Aus		Idn		d_1-d_2	SE dif	sd (d_1-d_2)
	IMS	IMS	d_1	SE	d_2	SE			
m032094	1.06	1.00	0.649	0.092	-0.649	0.092	1.298[a]	0.130	9.98[b]
m032662	1.31	1.64	0.301	0.127	-0.301	0.127	0.602[a]	0.180	3.35[b]
m032064	0.74	0.59	-0.083	0.120	0.083	0.120	-0.166	0.170	-0.98
m032419	1.15	1.21	0.332	0.093	-0.332	0.093	0.664[a]	0.132	5.05[b]
m032477	0.86	1.39	0.040	0.100	-0.040	0.100	0.080	0.141	0.57
m032538	0.87	0.78	0.539	0.099	-0.539	0.099	1.078[a]	0.140	7.70[b]
m032324	1.05	1.00	0.234	0.099	-0.234	0.099	0.468	0.140	3.34[b]
m032116	1.27	1.10	0.452	0.094	-0.452	0.094	0.904[a]	0.133	6.80[b]
m032100	1.00	0.98	-0.416	0.099	0.416	0.099	-0.832[a]	0.140	-5.94[b]
m032402	1.10	1.13	1.172	0.093	-1.172	0.093	2.344[a]	0.132	17.82[b]
m032734	1.11	0.99	-0.962	0.116	0.962	0.116	-1.924[a]	0.164	11.73[b]
m032397	0.96	1.08	-0.227	0.096	0.227	0.096	-0.454	0.136	-3.34[b]
m032695	1.19	1.17	-0.027	0.065	0.027	0.065	-0.054	0.092	-0.59
m032132	1.65	1.05	-0.091	0.095	0.091	0.095	-0.182	0.134	-1.35
m042041	1.00	1.05	0.207	0.102	-0.207	0.102	0.414	0.144	2.87[b]
m042024	0.76	0.87	-0.283	0.097	0.283	0.097	-0.566[a]	0.137	-4.13[b]
m042016	0.98	1.02	0.522	0.092	-0.522	0.092	1.044[a]	0.130	8.02[b]
m042002	1.01	0.94	-0.354	0.118	0.354	0.118	-0.708[a]	0.167	-4.24[b]
m042198a	0.63	0.91	-1.162	0.133	1.162	0.133	-2.324[a]	0.188	12.36[b]
m042198b	0.99	0.76	-0.589	0.112	0.589	0.112	-1.178[a]	0.158	-7.44[b]
m042198c	0.67	0.64	-0.058	0.207	0.058	0.207	-0.116	0.293	-0.40
m042077	1.10	0.98	0.696	0.098	-0.696	0.098	1.392[a]	0.139	10.04[b]
m042235	0.78	0.99	-0.095	0.097	0.095	0.097	-0.190	0.137	-1.39
m042067	1.76	1.23	0.905	0.099	-0.905	0.099	1.810[a]	0.140	12.93[b]
m042150	1.16	1.05	0.194	0.093	-0.194	0.093	0.388	0.132	2.95[b]
m042300a	0.87	0.84	-0.187	0.099	0.187	0.099	-0.374	0.140	-2.67[b]
m042300b	0.89	0.61	-0.071	0.101	0.071	0.101	-0.142	0.143	-0.99
m042260	0.70	1.29	-0.863	0.113	0.863	0.113	-1.726[a]	0.160	-10.80[b]
m042169a	0.80	0.81	0.409	0.099	-0.409	0.099	0.818[a]	0.140	5.84[b]
m042169b	1.35	1.50	-1.026	0.192	1.026	0.192	-2.052[a]	0.272	-7.56[b]
m042169c	0.77	0.79	-0.158		0.158		-0.316		
Separation Reliability=0.961									

IMS Infit mean square; a. difference in item difficulty outside the range ±0.50; b. st (d_1-d_2) outside the range ±2.00
Australian (N=289); Indonesian (N=298)

Makara Hubs-Asia, 2014, 18(2): 127-139
DOI: 10.7454/mssh.v18i2.3467

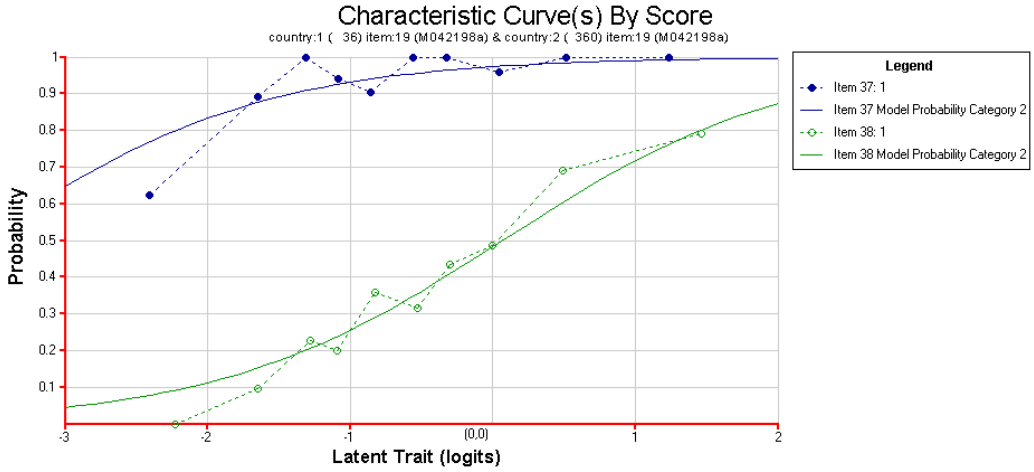

Figure 4. ICC for Item m042198a (Biased in Favor of Australian, d_1-d_2=-2.324)

Figure 5. ICC for Item m032402 (Biased in Favor of Indonesian, d_1-d_2=2.344)

This assumption is in line with the findings of some studies that documented the teaching strategy used by Indonesian mathematics teachers as a factor contributing factor to this failure (Hadi, 2004; Widjaya & Heck, 2003; Zakaria, Solfitri, Daud, & Abidin, 2013).

Consequently, this may affect Indonesian students' performance on the constructed response (CR) items of TIMSS (2007), which requires students to communicate mathematically (providing explanations and reasoning), to compare various results, and to understand the real-world context.

Another reason for the big difference in ability between the Australian and Indonesian groups in the mathematics tests of TIMSS (2007) is low teacher qualification. A survey of teacher quality conducted by the World Bank (2005) showed that the preparation and attendance of teachers are inadequate. Unlike many other countries, Indonesia allows graduates of all teacher-training institutes to become teachers without checking their preparedness to impart knowledge and skills under various school conditions. The survey also found that 20% of Indonesian teachers were absent at the time of random spot check in a representative number of schools. This finding is unfortunate because absenteeism could result in the low quality of education, particularly the low achievement in mathematics among students. Another study on teacher quality, which was conducted by Saito, Harun, Kuboki and Tachibana (2006), also revealed that mathematics teachers seldom pay attention to the learning processes of students. Teachers still seem to conceive a lesson only from the perspective of teaching models, such as the "chalk and talk," demonstration, and group discussion approaches. This is evident in the dominant interest in teaching

models, the lack of attention to detail in the learning processes of students, and the lack of questioning the reasons for mistakes and the misconceptions of students. In addition, teachers used most contact time to explain and solve mathematics problems, while students remain passive and simply copy what their teacher writes on the blackboard.

However, it is possible that other factors, such as experience with similar tests or a lesser propensity to guess, contributed to a different test-taking approach. These possibilities merit further investigation to determine the reasons that DIF items exist.

The results of this study suggest that future research should investigate other areas. For example, it important to determine the ways in which the results of items and item analyses differ. The current study only predicted that curriculum differences, instructional practices, and teacher quality were some factors contributing to DIF items. Future research should attempt to investigate the sources of DIF using the same data so that appropriate intervention can be made to improve the quality of test design. This study was an initial step in assessing DIF items. Problematic items identified by the statistical procedure could be examined more thoroughly to determine any other potential sources that were not found in this study.

Future research could also use more than one DIF technique to assess TIMSS test items so that the pattern of agreement of the procedures may produce reliable, generalizable results of DIF items. Yildirim (2006) suggested that using more than one method would lead to better understanding because multiple methodologies would compensate the defects of others.

In addition, this study found that many items in the TIMSS (2007) mathematics test recoded bad IMS and exhibited item bias. However, it was difficult to establish the reasons that they showed bad fitting or bias. Therefore, it is suggested to carry out replication studies or in-depth investigations before decisions are made to eliminate items identified as bad fitting and biased items in future TIMSS mathematics tests.

4. Conclusions

The investigation of item bias using the DIF technique of the Rasch model indicated that country DIF was a problem in the mathematics test items. Using Australian and Indonesian data, the analyses of country DIF identified that about 75% of the total number of items in each booklet being tested exhibited significant bias. The findings showed that 20 items in Booklet 8 and 23 items in Booklet 9 were identified as biased items. In addition, these items had differences in threshold values, and the standardized differences in item threshold were outside the predefined ranges. Furthermore, many items were apparently biased in favor of one group or the other. Based on the results of the analyses conducted in this study, it was concluded that TIMSS (2007) has many DIF items, and there was a big difference in ability between the two groups.

In addition, the country DIF analyses revealed that the Australian students generally performed better, and they found that the items in each booklet were relatively easier than the Indonesian students did. This DIF was consistently significant in both booklets used in the country DIF analyses. The differences in item performance observed in this study indicate serious limitations in using TIMSS results to make comparisons between students in Australia and Indonesia. Thus, further empirical evidence is needed before the results of TIMSS (2007) can be meaningfully used in research.

References

Allalouf, A., Hambleton, R.K., & Sireci, S.G. (1999). Identifying the causes of differences in translated verbal items. *Journal of Educational Measurement, 36*(3), 185-198.

Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika, 43*(4), 561-573.

Arim, R.G., & Ercikan, R. (2014). Comparability between the American and Turkish version of the TIMSS mathematics test results. *Educational and Science*, *39*(172), 33-48.

Byrne, B.M. (2003). Testing for equivalent self-concept measurement across culture. In Marsh, H.W., Craven, R.G., & McInerney, D.M. (eds.), *International advances in self-research: Speaking to the future*. Greenwich: Information Age Publishing. p. 291-314.

Camilli, G., & Shepard, L.A. (1994). *Methods for identifying biased test items.* California: Sage Publications.

Chen, Y.H., Gorin, G.S., Thomson, M.S., & Tatsuoka, K.K. (2008). Cross-cultural validity of the TIMSS 1999 mathematics test: verification of a cognitive model. *International Journal of Testing*, *8*, 251-271.

Dorans, N.J., & Holland, P.W. (1993). DIF detection and description: Mantel-Haenzsel and standardization. In Holland, P.W., & Wainer, H. (eds.), *Differential item functioning: theory and practice*. Hillsdale, NJ: Erlbaum. p. 137-166.

Emenogu, B.C., & Childs, R.A. (2005). Curriculum, translation, and differential functioning of measurement and geometry items. *Canadian Journal of Education*, *28*(1&2), 128-146.

Makara Hubs-Asia, 2014, 18(2): 127-139
DOI: 10.7454/mssh.v18i2.3467

Ercikan, K. (1998). Translation effects in international assessments. *International Journal of Educational Research*, *29*(6), 543-553.

Ercikan, K. (2002). Disentangling sources of differential item functioning in multilanguage assessment. *International Journal of Testing*, *2*(3&4), 199-215.

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

Gierl, M.J. (2000). Construct equivalence on translated achievement test. *Canadian Journal of Education*, *25*(4), 280-296.

Gierl, M.J., & Khaliq, S.N. (2001). Identifying sources of differential item and bundle functioning on translated achievement test: a confirmatory analysis. *Journal of Educational Measurement*, *38*(2), 164-187.

Grisay, A., & Monseur, C. (2007). Measuring the equivalence of item difficulty in the various versions of an international test. *Studies in Educational Evaluation*, *33*, 69-86.

Hadi, S. (2004). *Effective teacher professional development for implementation of realistic mathematics education in Indonesia*, *Thesis*, University of Twente, Enchede, Den Haag. Accessed on April 19, 2011 from http://doc.utwente.nl/58708/1/thesis_Hadi.pdf.

Hungi, N. (2005). Employing the Rasch model to detect biased items. In Alagumalai, S., Curtis, D.D., & Hungi, N. (eds.), *Applied Rasch measurement: a book of exemplars–Papers in honour of John P. Keeves.* The Netherlands: Springer. p. 139-157.

International Association for the Evaluation of Educational Achievement. (2008). TIMSS 2007 Technical Report. Accessed on March 15, 2011 from http://timss.bc.edu/timss2007/techreport.html.

International Association for the Evaluation of Educational Achievement. (2008). TIMSS 2007 Database. Accessed on March 15, 2011 from www.timss.bc.id/index.html.

International Association for the Evaluation of Educational Achievement. (2008). TIMSS 2007 User Guide for the International Database. Accessed on March 15, 2011 from http://timss.bc.edu/timss2007/idb_ug.html.

International Association for the Evaluation of Educational Achievement. (2008). TIMSS 2007 International mathematics report; finding from the IEA's trend in international mathematics and science study at the fourth and eighth grades. Accessed on March 15, 2011 from http://timss.bc.edu/timss2007/intl_reports.html.

International Association for the Evaluation of Educational Achievement. (2008). TIMSS 2007 Mathematics Assessment Framework. Accessed on April 2, 2011 from http://timss.bc.edu/timss2007/frameworks.html.

Linacre, J.M., Wright, B.D., Gustafsson, J-E., & Martin-Lof, P. (1994). Reasonable mean-square fit values. *Rasch Measurement Transactions*, *8*(3), 370-380.

Little, T.D. (1997). Mean and covariance structures (MACS) analyses of cross-cultural data: practical and theoretical issues. *Multivariate Behavioral Research*, *32*, 53-76.

Ng, D., Mosvold, R., & Fauskanger, J. (2012). Translating and adapting the mathematical knowledge for teaching (MKT) measures: the case of Indonesia and Norway. *The Mathematics Enthusiast*, *9*(1&2), 149-178.

Raczec, A.E., Ware, J.E., Bjorner, J.B., Gandek, B., Haley, S.M., et.al., (1998). Comparison of Rasch and summated rating scale s constructed from SF-36 physical function items in seven countries: Result from the IQOLA project. *J Clin Epidemiol*, *51*(11), 1203-1214.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests.* Copenhagen, Chicago: Danmarks Pegagogiske Institut, University of Chicago Press.

Saito, E., Harun, I., Kuboki, I., & Tachibana., H. (2006). Indonesia lesson study in practice: case study of Indonesian mathematics and science teachers education project. *Journal of In-service Education*, *32*(2), 171-184.

Sireci, S.G. (1997). Problems and issues in linking assessment across languages. *Educational Measurement: Issues and Practice*, *16*(1), 12-19.

Sireci, S.G., & Gonzales, E.G. (2003). *Evaluating the structural equivalence of test used in international comparison of educational assessment*. Paper presented at the Annual Meeting of the National Council of Measurement and Education in Chicago, New York, April 22-24 2003. Retrivied on http://eric.ed.gov/?id=ED481107.

Schulz, W., & Fraillon, J. (2009). *The analysis of measurement equivalence in international studies using the Rasch model*. Paper presented at Rasch measurement: present, past and future at the European Conference on Educational Research (ECER), in

Vienna, Austria, September 28-30 2009 Retrivied on http://iccs.acer.edu.ai/uploads/File/papers/ECER09_RaschModel_InternationalStudies(Schulz&Fraillon.pdf.

The Jakarta Post (March 1 2010). *Most Students Fail Mock National Examination*. Accessed on April 15, 2011 from http://www.thejakartapost.com/news/2010/03/01/most-students-fail-mock-national-examination.html.

Tilahun, M.A. (2004). Monitoring mathematics achievement over time: a secondary analysis of FIMS, SIMS and TIMS: a Rasch analysis. In Alagumalai, S., Curtis, D.D., & Hungi, N. (eds.), *Applied Rasch measurement: a book of exemplars – Papers in honour of John P. Keeves*. The Netherlands: Springer. p. 63-79.

Widjaja, Y.B., & Heck, A. (2003). How a realistic mathematics education approach and microcomputer-based laboratory worked in lessons on graphing at an Indonesian junior high school. *Journal of Science and Mathematics Education in Southeast Asia*, *26*(2), 1-51.

World Bank. (2005). *Improving education quality*. Accessed on June 6, 2014 from http://siteresources.worldbank.org/INTIINDONESIA/Resources/Publication/280016110613030543 9/617331-1110769011447/810296-1110769045002/Education.pdf.

Wright, B.D., & Linacre, J.M. (1994). Reasonable mean-square fit values. *Rasch Measurement Transactions*, *8*(3), 370.

Wright, B.D. (1997). Solving measurement problems with the Rasch model. *Journal Educational Measurement*, *14*(2), 97-116.

Wu, M.L., Adams, R.J., Wilson, M.R., & Haldane, S.A. (2007). *Conquest version 2.0 (Generalised Item Response Modeling Software)*. Camberwell, Victoria: ACER Press.

Wu, M.L. (2010). Measurement, sampling, and equating errors in large-scale assessments. *Educational Measurement: Issues and Practice*, *29*(4), 15-27.

Yildirim, H.H. (2006). The differential item functioning (DIF) analysis of mathematic items in the international assessment programs. *Doctoral dissertation*, Secondary Science and Mathematics Education, Middle East Technical University, Turkey. Accessed on April 19, 2011 from http://etd.lib.metu.edu.tr/upload/ 12607135/index.

Yildirim, H.H., Yildirim, S., & Verhelst, N. (2014). Profile analysis as a generalized differential item functioning analysis method. *Education and Science*, *39*(172), 49-64.

Zulkardi. (2002). Developing a learning environment on realistic mathematics education for Indonesia student teachers. *Thesis*, University of Twente, Enchede, Den Haag. Accessed on April 19, 2011 from http://doc.utwente.nl/58718/1/thesis_Zulkardi.pdf.

Zakaria, E., Solfitri, T., Daud, Y., & Abidin, Z.Z. (2013). Effects of cooperative learning on secondary school students' mathematics achievement. *Creative Education*, *4*(2), 98-100.