A blind item-review process as a method to investigate invalid moderators of item difficulty in translated assessment items

Enis Dogan
American Institutes for Research, Washington DC, United States

Ruhan Circi
Bogazici University, Istanbul, Turkey

The purpose of this study was to explore the utility of a blind item-review process as a method for investigating whether test items designed for cross-cultural use include invalid moderators of difficulty. An invalid moderator of difficulty is an item characteristic that affects students’ ability to demonstrate their true competence. The item review process suggested here was applied to the Third International Mathematics and Science Study-Repeat (TIMSS-R) eighth-grade science items translated from English into Turkish. First, an item review tool featuring 13 statements was developed. Each statement targeted a specific invalid moderator of difficulty. A sample of 100 Turkish teachers rated an intermixed pool of TIMSS-R and “local” science items (items developed originally in Turkish) on each statement. The teachers did not know the source of the items. Mean teacher ratings of the TIMSS-R and the local items were computed and compared. TIMSS-R items had significantly lower ratings on all 13 statements. Mean teacher ratings on five of the 13 statements correlated significantly with the differences between *p*-values for the Turkish sample and the average *p*-values for the international cohort.

INTRODUCTION

We conducted this study in order to determine if a blind item-review process provided a method for investigating whether test items designed for cross-cultural use include invalid moderators of difficulty. An invalid moderator of difficulty is an item characteristic that affects students' ability to demonstrate their true competence. Invalid moderators of difficulty can potentially lead to construct-irrelevant variance in test scores. Invalid moderators emerge when an item has unnecessarily complex language (for a given grade or age level) and unfamiliar graphs, charts, and tables (i.e., material that is not commonly used in classrooms at a given grade level). Other problematic features include unfamiliar technical terms (e.g., scientific and mathematical terms that have not been introduced in the classrooms), words, and phrases. The context in which the item is posed can also be an invalid moderator, depending on the learning experiences of the target student population. In this study, we used Third International Mathematics and Science Study-Repeat (TIMSS-R) eighth-grade science items translated from English into Turkish in order to explore the utility of our item review process.

TIMSS and the Validity of International Assessments

The Trends in International Mathematics and Science Study (TIMSS) is one of the world's most comprehensive international comparative studies of educational achievement. Designed to assess and compare the mathematics and science achievement of students from different countries, the study also allows for cross-national comparisons of educational background variables. TIMSS has been administered every four years thus far—in 1995, 1999, 2003, and 2007.

TIMSS-R was conducted under the auspices of the International Association for the Evaluation of Educational Achievement and included 38 countries. It assessed the mathematics and science achievement of Grade 8 students (ages 13 and 14) and was based on the mathematics and science curricula of the participating countries. Content areas included in the science assessment were earth science, life science, physics, chemistry, environmental and resource issues, scientific inquiry, and nature of science (Gonzalez & Miles, 2001). The science assessment included 146 items: 42 constructed-response (CR) and 104 multiple-choice (MC). All items used in TIMSS-R were first developed in English and then translated into 32 languages, including Turkish. TIMSS-R translation guidelines called for two independent translations of each test instrument from English to the target language. A translation review team compared the two translations to create the final version. O'Connor and Malak (2000) document and discuss the details of these processes.

International studies such as TIMSS are designed to provide data useful to educational policymakers. Multiple factors have the potential, however, to undermine the validity of the results obtained from international assessments. For example, Hambleton, Yu, and Slater (1999) argue that the alignment between the topics covered in international assessments and the national curriculum of each country can affect countries' performances. Different degrees of alignment can thus undermine the validity of comparisons made across countries. Ramseier (1999) also argues that close alignment between a national curriculum and an international assessment indicates the relevance of comparisons of achievement for that country.

As Pollitt and Ahmed (2001) argue, if the items used in international assessments, such as TIMSS, do not measure the intended constructs (e.g., science or mathematics), the results regarding the relative performance of the participating countries cannot be valid. Pollitt and Ahmed also point out that unless the cognitive processes invoked in students' minds match the ones intended by the item writers, the items lose their validity. The two authors explain that the level of familiarity students have with the context of the items (the story around which the problem is constructed) is a key factor in students' understanding of the tasks that these items require. For example, a mathematics item that discusses subway stations might not be familiar to students living in areas without a subway system.

TIMSS recognizes that it is important when "comparing student achievement across countries … that the comparisons be as fair as possible" (Mullis et al., 2000, p. 379). Because of concerns about the relationship between assessment results and curriculum-to-assessment alignment, TIMSS conducted a test-curriculum matching analysis (TCMA). The national research coordinator (NRC) from each participating country was asked to choose someone who was familiar with the mathematics and science curricula of the grade tested to determine the extent to which the tests were relevant to those curricula. During this process, the rater deemed an item appropriate if it was in the intended curriculum for more than 50% of the students. The details of this process can be found in Mullis et al. (2000). A group of Turkish experts concluded that over 95% of all science items in TIMSS-R fit the intended national curriculum for Turkey.

Despite this close alignment, Turkey ranked 34th out of the 38 participating countries in terms of TIMSS-R science achievement. We can offer a number of potential explanations for the relatively poor performance of Turkish students on the TIMSS-R science assessment besides the obvious one that the achievement of Turkish students after eight years of formal education is low. For example, Turkish students might have been less motivated to complete a low-stakes assessment, such as TIMSS-R, compared to the high-stakes assessments to which they are accustomed. It is also possible that both low motivation and poor performance resulted from the differences between the enacted curriculum to which the Turkish students were exposed and the content of the TIMSS-R science assessment. This conjecture, however, does not readily accord with the Turkish experts' opinion that 95% of the TIMSS-R items were covered in the national curriculum.

These possible explanations for Turkey's poor performance on TIMSS-R also apply to the other participating countries. For instance, TIMSS-R was a low-stakes assessment for all students participating in the study. Also, differences between what TIMSS-R measured and what was taught in the classrooms existed for all countries because the curricula of each differed.

Translation/Adaptation of Achievement Tests and Invalid Moderators of Item Difficulty

A major challenge in TIMSS-R is that the assessment is developed in English and translated and administered in different languages to students with different learning experiences. A substantial body of literature illustrates how the difficulty and meaning of test items can be affected when they are administered in different languages to students with different learning experiences (Abedi, 2006; Abedi & Gandara, 2006; Ercikan, Gierl, McCreith, Puhan, & Koh, 2004; Gierl & Khaliq, 2001; Hambleton, 2005; Sireci, Patsula, & Hambleton, 2005; Solano-Flores, Contreras-Nino, & Backhoff, 2006; Solano-Flores & Trumbull, 2003; van de Vijver & Poortinga, 2005).

Several researchers (e.g., Allalouf, Hambleton, & Sireci, 1999; Gierl & Khaliq, 2001) have used differential item functioning (DIF) analyses to identify translated/adapted test items that exhibit different psychometric features for different groups of students taking the test in different languages. After identifying DIF items, the analyst's next step is to examine the items in order to uncover the sources of DIF. Our approach was different. We decided not to examine items known to display DIF because we considered that these *post hoc* comparisons could bias experts' judgment of what might be wrong with a given item. We hypothesized that certain characteristics of the translated TIMSS-R science items contributed to the poor performance of Turkish students. We called such characteristics "invalid moderators of difficulty." An invalid moderator of difficulty is an item characteristic that affects students' ability to demonstrate their true competence (Leong, 2006).

In his 2006 article, Leong discusses factors that affect the difficulty of test items. He introduces an item difficulty framework that includes content difficulty, stimulus difficulty, task difficulty, and expected response difficulty. Leong defines content difficulty as the difficulty of the subject matter assessed: "In the assessment of knowledge, the difficulty of a test item resides in the various elements of knowledge such as facts, concepts, principles and procedures" (p. 3). The content difficulty level has three categories—basic (very familiar to the learner), appropriate (central to the core curriculum), and advanced (something the learner may not have had the opportunity to learn). He argues that content difficulty increases as more knowledge elements are involved in an assessment.

Stimulus difficulty is related to understanding the words, phrases, and representations (e.g., diagrams, tables) used in an item. Test items containing words and phrases that require only simple and straightforward comprehension are usually easier than those that require careful or technical comprehension. Task difficulty refers to the complexity of the process involved in producing an answer or formulating a solution. Generally, items that include more than one step to reach the solution are classified as harder items than those that do not require this step. The level of guidance provided in the items and the complexity of the cognitive processes also affect the difficulty of a task.

Expected response difficulty is reflected in the scoring rubrics. The level of detail in an expected response to the item determines the response difficulty level. In Leong's (2006) framework, this aspect of difficulty applies to CR items only, and is most likely to occur when examinees are unclear about the demand of a response and do not produce an answer that is sufficient to earn marks that reflect their abilities. Leong discusses moderators of difficulty that prevent a valid measurement of the construct of interest. Moderators can prevent examiners from assessing students' knowledge in terms of the intended construct, because, as we noted earlier, test-takers who are faced with invalid moderators might not be able to demonstrate their true ability or competence.

Note that Leong's classification in itself does not say anything about the use of more or fewer difficult items for comparisons across countries. While the measurement accuracy may decrease with increasing difference between item pool difficulty and population ability, modern psychometric methods, such as the ones used in TIMSS, help prevent bias in those situations.

To make a moderator an *invalid* moderator, there needs to be construct-irrelevant variance that affects a subgroup or all of the population. In addition to determining the level of task difficulty, stimulus difficulty, and expected response difficulty, the analyst needs to assess whether these factors are affecting one or more subgroups in ways that differ from the ways affecting other subgroups. For example, if a physics item requires familiarity with subway maps in addition to content knowledge about classical mechanics, as taught in physics, these features would constitute invalid moderators. If the knowledge required in classical mechanics is simply higher than what is typically taught in a certain country, but nevertheless aligns with the curriculum, as agreed by the experts judging the TIMSS item, this feature might make the item too difficult for certain populations, but it would not involve invalid moderators of difficulty. This consideration would remain true as long as the probability of a correct response to the item increases with nothing other than physics-related skills and knowledge; in other words, knowledge and skills about things other than physics would not be required.

METHOD

We developed an item review tool, written in Turkish and based on the literature mentioned above and Leong's (2006) framework of moderators of item difficulty. The tool featured, along with Leong's moderators, descriptors that brought the total number of statements to 13. We then asked a sample of 100 school teachers to use these statements to rate, on a Likert scale, a set of 100 science items. We did not tell the teachers the source of the items.

As we describe below, 10 teachers rated each item. Fifty of these items were translated (i.e., items from TIMSS-R) and the rest were "local" items (items developed originally in Turkish to assess Turkish students). Thirteen of the 50 TIMSS-R items were CR items. Local items were either from a pool of items from the 1999 national test for eighth-graders[1] or they were items that teachers developed for formative classroom assessments. The national test, developed by the Turkish Ministry of Education, contained 24 items, all of which were MC. Of the remaining 25 classroom assessment items, six were CR and 19 were MC.

We assembled the local and the TIMSS-R items into blocks of five items, resulting in 10 blocks of local items and 10 blocks of TIMSS-R items. We then arranged these blocks in 100 booklets, each of which included one local block and one TIMSS-R block. We randomly assigned blocks to booklets in a way that ensured that each of the 100 test items was rated by 10 teachers. (During this assignment step, we kept in place the constraint mentioned above, that is, each booklet to contain one local block and one TIMSS-R block.) Because each teacher received one booklet and rated only 10 items, the teachers were not overburdened. We asked the teachers to rate items according to the 13 statements in the item review tool. Some of the statements applied to MC items only.

The Raters

We recruited the 100 participating teachers through an email posted on a number of professional list-servers on internet. Participation was voluntary, and the teachers were not offered any incentive to participate. All teachers with at least two years of teaching experience at the sixth- , seventh- , or eighth-grade within the boundaries of Istanbul[2] at the time of the study (2008) were eligible to participate. We invited teachers to participate in a study to evaluate the quality and the appropriateness of science test items. We did not mention TIMSS during the recruitment and the data-collection phases. Nor did we mention that some of the items had been translated. Our aim, in this regard, was to keep the review process blind by ensuring that the teachers did not have the opportunity to detect hints from the items suggesting that they had been translated.

Seventy percent of the 100 teachers who agreed to participate were female. Thirty-one percent taught at public schools. The rest were teachers from various private schools. The median age of the participating teachers was 31. The median number of years of teaching experience was six.

Instrument

Table 1 sets out the wording of the item review tool in English. Table 2 provides the original version, written in Turkish. As noted above, we based the statements largely on Leong's (2006) taxonomy of invalid moderators of item difficulty. We asked the teachers to rate each of the test items on a Likert scale, where 1 indicated *strongly agree* and 5 indicated *strongly disagree*, and we instructed them to think about a typical student in their classroom while rating the items. In keeping with Leong's taxonomy, we divided the survey questions into four clusters: content (C), stimuli (S), task (T), and expected response (R) difficulty. Not all statements were applicable to all science items reviewed by the teachers. T2 and R2 applied to MC items only.

[1] The national test determines students' access to high school.

[2] This restriction was put in place to minimize the cost of conducting the study. Note, however, that roughly 15% of all students and 13% of all teachers in Turkey reside in Istanbul.

Table 1: Item review tool: A survey of difficulty factors and additional invalid moderators of item difficulty

Carefully read this science question. Think about how your students might approach this question. Think about the challenges they might face in understanding or solving this question. Now, rate this question on each of the following statements.

C1: The item includes concepts unfamiliar to students.
S1: There is inaccuracy or inconsistency in the information given in the item.
S2: There is insufficient information in the item to reach a clear answer.
S3: There is uncommon vocabulary in the item.
S4: There are grammatical errors in the item that can lead to misunderstanding.
S5: There are unfamiliar representations (diagrams, graphs, tables, pictures) in the item.
S6: The item includes unfamiliar terminology.
T1: There are unfamiliar sentence structures used in the item.
T2: Alternatives contain concepts unfamiliar to students.
T3: The problem is presented in an unfamiliar context.
T4: The stem of the item is misleading to the students.
R1: The item has a number of plausible correct answers.
R2: Alternatives are insufficient to reach the correct answer.

Note: C = Content difficulty, S = Stimulus difficulty, T = Task difficulty, R = Expected response difficulty.

Table 2: The original version of the item review tool in Turkish

Önünüzdeki fen bilgisi sorusunu dikkatlice okuyun. Öğrencilerinizin bu soruya nasıl yaklaşacağını düşünün. Soru çözmede yada anlamada karşılaşabilecekleri zorlukları düşünün. Simdi, bu soruyu asağıdaki önermeler için değerlendirin.

1: Soru öğrenciler için tanıdık olmayan kavramlar içermekte.
2: Soruda verilen bilgilerde tutarsızlık var.
3: Soruda doğru yanıta ulaşmak icin yetersiz bilgi verilmiş.
4: Soruda öğrenciler için tanıdık olmayan kelimeler kullanılmış.
5: Soruda yanlış anlamaya sebep olabilecek dilbilgisi hatası var.
6: Soruda verilen kaynaklar (diyagram, grafik, resim) ögrencilerin sık karşılaşmadığı türden.
7: Soru öğrenciler için tanıdık olmayan terminoloji içermekte.
8: Sorudaki kullanılan kelime dizilimi öğrencilerin aşina olmadığı türden.
9: Cevap şıkları öğrencilerin alışık olmadıkları kavramlar içermekte.
10: Sorunun içeriği (bağlam) öğrencinin ilgi kurabileceği türden değil.
11: Soru cümlesi öğrencilerileri yanlış yönlendirecek tarzda.
12: Sorunun birçok alternatif doğru cevabı var.
13: Cevap şikları öğrencilerin doğru yanıtı bulmaları icin yetersiz.

Analysis

Our first step was to compute, for each science item rated in the study, the mean ratings and the associated variance (across raters) for all 13 statements in the item review tool. We recorded items with mean ratings lower than 3 for any of the 13 statements and labeled these as items with ”poor” mean ratings. We then used a hierarchical linear modeling (HLM) approach to compare the items from different sources (i.e., TIMSS-R or local) according to their mean ratings on all 13 statements. Here, we treated the items as nested under teachers, an acknowledgment that ratings from different teachers on the same item constitute dependent observations.

RESULTS

Mean Ratings and Variation Across Ratings

We computed the mean rating and the associated variance (across 10 raters) for each item on the item review tool. Table 3 displays the range and the average value of these means and variances for each statement across all 100 items rated. The average mean ratings ranged from 3.67 (C1: unfamiliar concepts) to 4.03 (R2: insufficient alternatives). The average variance associated with these means ranged from 0.10 (T4: misleading stem) to 0.28 (T1: unfamiliar sentence structures). Note that lower variances indicate higher agreement among raters.

Table 3 also displays the range of the means and variances. For all 13 statements, the minimum value of the variance of ratings (across the 10 raters) was 0.00, indicating that there was at least one item where all raters gave the same rating. The actual number of items where there was a perfect agreement across all 10 raters ranged from 37 items for C1 (unfamiliar concepts) to 53 items for S4 (grammatical errors).

Table 3: The range and the average of means and variances of ratings on each of the 13 statements in the item review tool across the 100 items rated

	Mean			Variance		
	Min	*Max*	*Average*	*Min*	*Max*	*Average*
C1: unfamiliar concepts	1.00	4.50	3.67	0.00	1.21	0.27
S1: inaccurate/inconsistent information	1.40	4.90	3.93	0.00	1.11	0.15
S2: insufficient information	1.50	5.00	3.93	0.00	1.07	0.16
S3: uncommon vocabulary	2.00	5.00	3.88	0.00	1.07	0.15
S4: grammatical errors	1.30	5.00	3.94	0.00	0.99	0.11
S5: unfamiliar representations	2.00	4.70	3.86	0.00	1.11	0.23
S6: unfamiliar terminology	2.00	4.70	3.83	0.00	1.11	0.19
T1: unfamiliar sentence structures	2.10	4.80	3.80	0.00	1.11	0.28
T2: alternatives with unfamiliar concepts	1.40	5.00	3.89	0.00	1.07	0.16
T3: unfamiliar context	2.10	4.80	3.91	0.00	1.11	0.19
T4: misleading stem	2.10	5.00	3.99	0.00	0.77	0.10
R1: multiple plausible correct answers	2.00	5.00	3.89	0.00	1.11	0.22
R2: insufficient alternatives	2.30	5.00	4.03	0.00	1.11	0.11

Items With Poor Ratings

As we noted earlier, we regarded items with mean ratings (across raters) lower than 3 on the Likert scale as having poor mean rating. According to this criterion, 32 of the 100 items had poor mean ratings on one or more of the 13 dimensions in the item review tool. These items, 32 of which were TIMSS-R items, are listed with the corresponding mean ratings in Table 4.

Table 4 also displays the number of dimensions on which each of these items had poor mean ratings. The three local items, one a national test (NT) item and two classroom assessment (CA) items, had poor mean ratings on the same statement: C1 (unfamiliar concepts). Seven TIMSS items had poor mean ratings on four or more statements. One of these TIMSS items (Item 41 in Table 4) had poor mean ratings on seven of the 13 dimensions: C1 (unfamiliar concepts), S3 (uncommon vocabulary), S4 (grammatical errors), S6 (unfamiliar terminology), T2 (alternatives with unfamiliar concepts), T3 (unfamiliar context), and T4 (insufficient alternatives).

Table 4 furthermore shows the number of items with poor mean ratings for each of the 13 dimensions. S3 (uncommon vocabulary), S6 (unfamiliar terminology), and T1 (unfamiliar sentence structures) had the highest number of items with poor mean ratings. S3 (uncommon vocabulary) had 10, S6 (unfamiliar terminology) had 11, and T1 (unfamiliar sentence structures) had eight such items, all of which were from the TIMSS-R item pool.

Comparison of TIMSS and Local Items

After computing the mean ratings across raters on all 100 items for each of the 13 statements in the item review tool, we plotted the ratings according to item source (see Figure 1). As is evident from Figure 1, the mean ratings of the TIMSS-R items were lower than the mean ratings of the NT and CA items. However, the mean ratings of the NT and the CA items did not differ on any of the 13 statements. We accordingly combined these two types of items as "local items" in the next phase of analysis.

In order to test the significance of these differences, we conducted an HLM analysis in which we treated items as nested under teachers. We formulized the HLM model as follows:

Level 1: $Y_{ij} = \beta_{0j} + \beta_{1j}\,(\text{Local}) + \varepsilon_{ij}$

Level 2: $\beta_{0j} = \gamma_{00} + u_{0j}$

$\beta_{1j} = \gamma_{10}$

where Y_{ij} is the j^{th} teacher's rating for item i and Local is a dummy variable that is 0 for TIMSS-R items and 1 for local items. We ran the model separately for each statement in the item review tool.

This approach accommodated the dependency among the teacher ratings for the same science item being rated. The analysis took into account the variance between teachers (raters). We then computed, within the same framework, the generalizability (G) coefficients, and partitioned the variance in ratings across the two levels (i.e., items and raters).

Table 4: Items with poor mean ratings by source, type, and total number of items with poor mean ratings on each of the 13 statements in the item review tool

Item	Source	Type	Number of poor ratings*	Mean teacher ratings												
				C1	*S1*	*S2*	*S3*	*S4*	*S5*	*S6*	*T1*	*T2*	*T3*	*T4*	*R1*	*R2*
Item 41	TIMSS	MC	7	1.4			2.3	1.3		2.8		1.4	2.6	2.9		
Item 6	TIMSS	MC	6		2.3	2.6	2.2	2.2	2.6		2.5					
Item 17	TIMSS	MC	5		2.5		2.8		2.0	2.0			2.1			
Item 38	TIMSS	MC	5				2.6			2.4		2.3			2.5	2.3
Item 8	TIMSS	MC	4		2.8			2.4	2.3	2.4						
Item 11	TIMSS	MC	4					2.9	2.6	2.7					2.0	
Item 56	TIMSS	CR	4			2.2	2.3			2.8			2.4			
Item 15	TIMSS	MC	3			2.7	2.4			2.6						
Item 31	TIMSS	MC	3					2.5			2.8				2.9	
Item 46	TIMSS	MC	3				2.2			2.2		1.9				
Item 4	TIMSS	MC	2				2.4				2.1					
Item 36	TIMSS	MC	2		1.4	1.5										
Item 45	TIMSS	MC	2							2.8		1.7				
Item 54	TIMSS	MC	2								2.8	2.1				
Item 57	TIMSS	MC	2							2.6		2.4				
Item 59	TIMSS	CR	2					2.5			2.7					
Item 76	TIMSS	CR	2				2.7			2.8						
Item 10	TIMSS	MC	1				2.0									
Item 19	TIMSS	MC	1		2.9											
Item 21	TIMSS	MC	1								2.2					
Item 25	TIMSS	MC	1	1.6												
Item 26	TIMSS	MC	1										2.8			

Table 4: Items with poor mean ratings by source, type, and total number of items with poor mean ratings on each of the 13 statements in the item review tool (contd.)

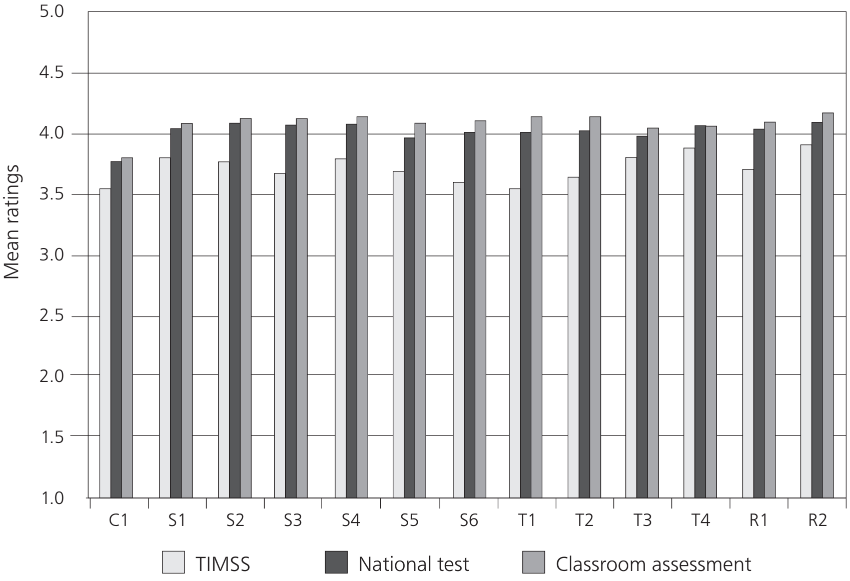
Item	Source	Type	Number of poor ratings*	Mean teacher ratings												
				C1	*S1*	*S2*	*S3*	*S4*	*S5*	*S6*	*T1*	*T2*	*T3*	*T4*	*R1*	*R2*
Item 39	CA	MC	1	1.6												
Item 61	TIMSS	MC	1								2.8					
Item 70	TIMSS	MC	1												2.3	
Item 73	CA	MC	1	1.0												
Item 82	NT	MC	1	1.2												
Item 84	TIMSS	MC	1			2.8										
Item 20	TIMSS	CR	1												2.4	
Item 40	TIMSS	CR	1								2.5					
Item 50	TIMSS	CR	1						2.9							
Item 68	TIMSS	CR	1											2.2		
Number of items with poor mean ratings				5	5	5	10	6	5	11	8	6	4	2	5	1

Notes:

* A poor item is defined as one with a mean rating lower than 3.

Only items with a mean rating lower than 3 are listed in the table.

Figure 1: Mean ratings on each of the 13 statements in the item review tool according to item source



Note: Lower ratings indicate unfavorable ratings.

C1: unfamiliar concepts
S1: inaccurate or inconsistent information
S2: insufficient information
S3: uncommon vocabulary
S4: grammatical errors
S5: unfamiliar representations
S6: unfamiliar terminology

T1: unfamiliar sentence structures
T2: alternatives with unfamiliar concepts
T3: unfamiliar context
T4: misleading stem
R1: multiple plausible correct answers
R2: insufficient alternatives

The G coefficients ranged from .88 to .96 for the 13 statements in the item review tool, indicating that only a small portion of the variance in ratings was due to teachers. Table 5 presents a summary of the results of the HLM analyses. The table also depicts the differences between mean ratings (across raters and items). All comparisons were statistically significant. TIMSS-R items had significantly lower ratings on all 13 statements, indicating that the raters judged these items as more problematic than the local items on all dimensions.

Teacher Ratings and Item Difficulty

After observing that teachers rated the TIMSS-R items less favorably than the other items, we investigated whether these ratings were associated with the relative difficulty of TIMSS-R items for the Turkish sample compared to the international sample. We explored the relationship between the mean ratings on 13 survey questions and the difference between the percentage of correct values for the international and Turkish samples ($p_{INT - TR}$) for the TIMSS-R items. This correlation was across items, not raters; that is, it was based on the ratings across items on 13 statements and the percentage correct for each of these items.

Table 5: Mean ratings of TIMSS-R and local items on each statement in the item review tool and the results of significance test comparing these means

	TIMSS-R items		Local items		Mean comparisons with HLM			
	Mean	SD	Mean	SD	γ_{01}	SE	*t*	*p*
C1: unfamiliar concepts	3.53	0.73	3.79	0.73	0.25	0.03	8.23	0.00
S1: inaccurate/inconsistent information	3.80	0.55	4.05	0.25	0.25	0.03	9.88	0.00
S2: insufficient information	3.76	0.61	4.10	0.20	0.35	0.04	9.25	0.00
S3: uncommon vocabulary	3.67	0.70	4.09	0.21	0.42	0.04	10.63	0.00
S4: grammatical errors	3.79	0.64	4.10	0.18	0.31	0.04	8.98	0.00
S5: unfamiliar representations	3.68	0.54	4.03	0.28	0.34	0.04	8.72	0.00
S6: unfamiliar terminology	3.60	0.64	4.06	0.25	0.45	0.04	11.43	0.00
T1: unfamiliar sentence structures	3.54	0.61	4.06	0.22	0.52	0.04	12.22	0.00
T2: alternatives with unfamiliar concepts	3.64	0.83	4.07	0.22	0.38	0.04	9.16	0.00
T3: unfamiliar context	3.80	0.50	4.01	0.27	0.21	0.04	5.73	0.00
T4: misleading stem	3.92	0.41	4.06	0.15	0.15	0.03	5.44	0.00
R1: multiple plausible correct answers	3.70	0.59	4.08	0.30	0.37	0.04	9.06	0.00
R2: insufficient alternatives	3.91	0.42	4.12	0.24	0.19	0.03	7.05	0.00

Ratings on five survey items (S3, S4, S6, T3, and R1) correlated significantly ($p < .05$) with $p_{INT - TR}$. The correlation coefficients were -.46 (S4: grammatical errors); -.39 (R1: multiple plausible correct answers); -.37 (T3: unfamiliar context); -.33 (S6: unfamiliar terminology); and -.31 (S3: uncommon vocabulary). The negative but not strong significant correlation coefficients indicate an association between unfavorable ratings on these five statements and the Turkish students finding it relatively difficult to answer these TIMSS-R items correctly. The issue here is that, of these five statements, two (T3, S6) can be viewed as difficulty factors, which are, by themselves, not necessarily invalid moderators. A third statement (S3) could be viewed as a difficulty factor as well as an invalid moderator because of contextualized items that tap into construct-irrelevant sources of variance. Difficulty in correctly answering the remaining two items (S4, R1) could be due to factors associated with the translation or adaptation of them.

DISCUSSION AND CONCLUSIONS

An increasing number of countries are interested in participating in international assessments so that they can better understand the achievement of their student populations and assess the outcomes of their educational provision in relation to inputs, such as curricular materials and teacher training (Kellaghan & Greaney, 2001). However, the validity of the results of international assessments depends on the quality of the test translation and adaptation process.

In order to address this matter (among others), the International Test Commission adopted guidelines for translating and adapting tests for cross-cultural use (Hambleton, 2005). Two of these guidelines (p. 26) are especially relevant for international assessments such as TIMSS:

> D1: Test developers/publishers should ensure that the adaptation process takes full account of linguistic, psychological, and cultural differences in the intended populations.
>
> D3: Test developers/publishers should provide evidence that item content and stimulus materials (e.g., any passages) are familiar to all intended populations of interest.

We undertook our study to illustrate the utility of a blind item-review process designed to provide information, assuming the above guidelines were followed, about whether the translation/adaptation of test items used in an international assessments was done correctly. The item review process that we used in our study allowed for the detection of invalid moderators of difficulty in translated/adapted items designed for cross-cultural use. Such moderators can stem from differences in linguistic background and from content and stimulus familiarity, along with other factors.

Our study did, however, have several limitations. First, the apparent lack of literature on invalid moderators of difficulty means that our conceptual framework and item review tool need further evaluation. Second, the released TIMSS-R items that we used in the study were not necessarily representative of the larger pool of TIMSS-R science items in terms of content, cognitive demand, and linguistic features. Moreover, the content and the item-type distribution of the TIMSS-R and local items were not identical. Third, the raters in this study were mostly teachers from private schools in Istanbul. This group of teachers was not a representative sample of science teachers in Turkey, and the method we used to recruit them was not an ideal way to obtain a representative sample. Fourth, the limited burden that could be put on the science teachers who volunteered to participate also limited the number of ratings that could be required from each of them, which limited the data base for the study. More data points would have been valuable. Fifth, we based our item review tool largely on work conducted by Leong (2006).

These limitations need to be addressed in future studies designed to build on the blind item-review process and to improve the use of this process. Future studies also need to examine the dimensionality of this survey tool. And the tool itself could be improved by incorporating the work of others in the field. Abedi and Gandara (2006), Ercikan et al. (2004), and Hambleton, Merenda, and Spielberger (2005) are excellent such sources.

With refined items, and possibly a revised rating scale, the item review tool and the associated blind item-review process suggested here could serve as a valuable tool for practitioners and researchers alike who are interested in better translation and adaptation procedures for items used in international assessments. Such procedures would result in more valid items, free of invalid moderators of difficulty, which in turn would generate more valid results.

References

Abedi, J. (2006). Psychometric issues in the ELL assessment and special education eligibility. *Teachers College Record*, *108*(11), 2282–2303.

Abedi, J., & Gandara, P. (2006). Performance of English language learners as a subgroup in large-scale assessment: Interaction of research and policy. *Educational Measurement: Issues and Practices*, *26*(5), 36−46.

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

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.

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

Gonzalez, E. J., & Miles, J. A. (2001). *TIMSS-R user guide for the international database.* Chestnut Hill, MA: Boston College.

Hambleton, R. K. (2005). Issues, designs, and technical guidelines for adapting tests into multiple languages and cultures. In R. Hambleton, P. Merenda, & C. Spielberger (Eds.), *Adapting educational and psychological tests for cross-cultural assessment* (pp. 3–38). Hillsdale, NJ: Lawrence Erlbaum.

Hambleton, R. K., Merenda, P., & Spielberger, C. (Eds.). (2005). *Adapting educational and psychological tests in cross-cultural assessment*. Hillsdale, NJ: Lawrence Erlbaum.

Hambleton, R. K., Yu, J., & Slater, S. C. (1999). Field-test of the ITC guidelines for adapting psychological tests. *European Journal of Psychological Assessment*, *15*(3), 270−276.

Kellaghan, T., & Greaney, V. (2001). The globalisation of assessment in the 20th century. *Assessment in Education*, *8*(1), 87–102.

Leong S. C. (2006, May). *On varying the difficulty of test items*. Paper presented at the annual meeting of the International Association for Educational Assessment, Singapore. Retrieved from http://www.iaea2006.seab.gov.sg/conference/download/papers/On%20varying%20the%20difficulty%20of%20test%20items.pdf

Mullis, I. V. S., Martin, M. O., Gonzalez E. J., Gregory K. D., Smith T. A., Chrostowski, S. J., ... O'Connor, M. K. (2000). *TIMSS-R: International science report*. Chestnut Hill, MA: Boston College.

O'Connor, K. M., & Malak, B. (2000). Translation and cultural adaptation of the TIMSS instruments. In M. O. Martin, K. D. Gregory, & S. E. Stemler (Eds.), *TIMSS 1999 technical report* (pp. 89−100). Chestnut Hill, MA: Boston College.

Pollitt, A., & Ahmed, A. (2001). *Science or reading? How students think when answering TIMSS questions*. Paper presented at the annual meeting of the International Association for Educational Assessment, Rio de Janeiro, Brazil.

Ramseier, E. (1999). Task difficulty and curricular priorities in science: Analysis of typical features of the Swiss performance in TIMSS-R. *Educational Research and Evaluation*, 5(2), 105–126.

Sireci, S., Patsula, L., & Hambleton, R. (2005). Statistical methods for identifying flaws in the test adaptation process. In R. Hambleton, P. Merenda, & C. Spielberger (Eds.), *Adapting educational and psychological tests for cross-cultural assessment* (pp. 93–116). Hillsdale, NJ: Lawrence Erlbaum.

Solano-Flores, G., Contreras-Nino, L. A., & Backhoff, E. (2006). Test translations and adaptation: Lessons learned and recommendations for countries participating in TIMSS, PISA, and other international comparisons. *REDIE: Electronic Journal of Educational Research*, *8*(2).

Solano-Flores, G., & Trumbull, E. (2003). Examining language in context: The need for new research and practice paradigms in the testing of English-language learners. *Educational Researcher*, *32*(2), 3–13.

van de Vijver, F., & Poortinga, Y. K. (2005). Conceptual and methodological issues in adapting tests. In R. Hambleton, P. Merenda, & C. Spielberger (Eds.), *Adapting educational and psychological tests for cross-cultural assessment* (pp. 39–61). Hillsdale, NJ: Lawrence Erlbaum.