Retaining Translated Verbal Reasoning Items by Revising DIF Items

Avi Allalouf

National Institute for Testing and Evaluation, Israel

Paper presented at the annual meeting of the
American Educational Research Association
New Orleans, LA, April 2000

Abstract

When a test is translated from a source language to a target language, the result is generally not two psychometrically equivalent tests. Analyzing the item, the basic element of the test, which sometimes functions differently across languages, can help in understanding the difference between the source and the target language tests. If the sources of differential item functioning (DIF) across languages could be predicted, this could have important implications for test adaptations. In addition, the likelihood of producing items that do not have DIF can be increased by revising items with DIF. The results of Allalouf, Hambleton & Sireci (1999) served as the basis for the current study. In that study, verbal reasoning items – analogies, sentence completions, logic and reading comprehension items – that were administered in Hebrew and in Russian, were analyzed for DIF using the Mantel-Haenszel method. A panel of translators suggested reasons for the DIF in each item. These reasons included differences in word difficulty, item format, content and cultural relevance. The current study examines item revision as a tool for improving test adaptations. A panel of translators and researchers revised the DIF items previously detected in Allalouf et al. (1999), based on the reasons for DIF found in that study. The revised items were then re-administered. The challenge of the study was to reduce the DIF. The revised target language items were compared to the original source language items, and to the original translation. Results showed that the revision succeeded in reducing DIF: out of the 37 items that were revised by the panel, 27 exhibited reduced DIF (12 of which exhibited significantly reduced DIF), and eight exhibited increased DIF (two, significantly increased DIF). An attempt was made to determine which sources of DIF and which item types can be revised most effectively. Empirical guidelines for using a panel to reduce DIF are presented.

I would like to thank Lev Novikov for coordinating the revision process and Ronald K. Hambleton, Stephen G. Sireci and Mark J. Gierl for their helpful comments.

When a test is translated from a source language to a target language, the result is generally not two psychometrically equivalent tests. DIF (differential item functioning) analysis reveals those items whose psychometric characteristics change following translation. DIF items should be removed from equating and scoring, which lowers the reliability and validity of the translated tests. The DIF items should also be removed from item banks, so that they will not be used in future tests. Removing them from item banks also involves a financial aspect, since new translated items are costly to produce.

Analysis of translated items can enlighten us as to the possible sources of DIF. Angoff and Cook (1988) analyzed the equivalence between the SAT and its Spanish-language counterpart, the *Prueba de Aptitud Academica* (PAA). They concluded that the amount of text in an item is significant; items with less text tend to have more translation DIF, while items with more text are more likely to retain their meaning. Gafni and Canaan-Yehoshafat (1993), and Beller (1995) studied the translation of the Israeli Inter-University *Psychometric Entrance Test* (PET) from Hebrew into Russian, and arrived at the same conclusions as Angoff and Cook. Allalouf, Hambleton and Sireci (1999) found causes of DIF in verbal reasoning items that were associated with specific item types. In a recent study, Gierl and Khaliq (2000) identified four sources of DIF in Canadian achievement tests administered in English and French. Gierl and Khaliq created an eleven-member panel that, by using these sources, had significant success in predicting the language group that would perform better on item bundles.

Translated verbal items with DIF can be retained through successful revision. The purpose of the present study is to determine whether DIF in translated verbal items could be reduced or eliminated by revising these items. Successful revision can achieve four important goals: (1) retaining translated items and maintaining the size of item banks; (2)

providing a better understanding, through an efficient revision process, of the sources of DIF, (3) determining for which sources of DIF and for which item types revision can be more effective. This can be regarded as the second big challenge of this study (the first, of course, was reducing the DIF); and (4) setting up empirically-based guidelines for reducing DIF in translated items in which DIF was already detected.

Studies of item revision for DIF between different language versions have never been conducted, although similar studies have examined revised items in a single language, where DIF was found for ethnic and gender groups. Curley and Schmitt (1993) revised 23 ethnic and gender DIF items from the SAT-Verbal with a success ratio of 67%: 12 of the 18 large DIF items displayed moderate or no DIF after revision. The authors concluded that "…revising items is feasible and likely to succeed often enough to make it practical to do so, particularly when prior research on hypothesized DIF factors and/or DIF factors based on observed occurrence of extreme DIF inform the revisions" (p. 15).

A revision procedure should be based on the experience gathered from studies on judgmental methods for DIF detection. Van de Vijver & Hambleton (1996) recommended that a group that combines linguistic and psychological expertise should be involved in the translation process. Hambleton & Jones (1994) stated the importance of training the judges, and of face to face group discussions between them. They found that a short standard bias form increased the effectiveness of the item review. Examples of several item bias review forms appear in Tittle (1980).

Method

Tests and Item Types

The study was conducted on the verbal subtest of the *Psychometric Entrance Test* (PET) – a high-stakes test used for admissions to universities in Israel – translated from Hebrew to Russian. Item types in the verbal subtest include analogies, sentence completions, logic and reading comprehension. Six test sections were used, consisting of 30 items, of which about 20 were translated in each section.

DIF items

The item revision in the study was performed on 37 (out of 42) verbal items having DIF found in Allalouf et al. (1999). The most problematic item types were analogies (14 DIF items) and sentence completions (13 DIF items). Reading comprehension items were less problematic (7 DIF items) and logic items were the least problematic item type (only 3 DIF items). The main causes of DIF discovered in Allalouf et al. (1999) were: differences in word difficulty, in item format, in cultural relevance and, sometimes, differences in content between the source language and the translation.

Revision Design and Administration of Revised Items

An eight-person panel of four Russian translators and four Hebrew-speaking researchers was set up for reviewing the DIF items. The panel included experts in language editing, linguistics and test translation. The panel proposed revisions based on the causes of DIF attributed to each of the items. The reviewers were provided with item analysis data (in the two languages) and MH-D-DIF statistics, and also with examples of no-DIF translated items. The revisions had to preserve the original content of the items from the source language while attempting to eliminate the causes of DIF. The six test

sections from which the DIF items were originally taken were re-administered in April 1999 using the reformulated DIF items. The review process involved the following steps:

(a) **First draft** - One of the translators, a trained linguist, served as the coordinator-translator of the study. He reviewed the 42 DIF items and examined the explanations suggested for them in the previous study. He approved some of the explanations, modified others and offered several new hypotheses. Based on this work, the first draft of the revised items was prepared.

(b) **Second draft** - Each of the causes of DIF was thoroughly examined and discussed in separate meetings between the coordinator-translator and each of the other three translators. The purpose of these meetings was to arrive at a consensus regarding the optimal revision for each item. After this, the second draft of the revised items was prepared.

(c) **Final draft** - The entire panel met three times in order to arrive at a final revised version of each item. During these meetings, panel members received an Item Revision Form which included, for each item: (1) the original Hebrew version; (2) the original translated Russian version (i.e. the version displaying DIF); (3) item statistics for each language version: proportion correct, proportion choosing each response alternative, and the correlation between choosing an item alternative and the raw score for the verbal domain; (4) data on the performance of the Russian and Hebrew examinee groups (in terms of magnitude of DIF) on the item; (5) the optimal revised item in Russian (sometimes, two alternative revised versions); and (6) an empty space for writing the final revised version. An example of the Item Revision Form is presented in Appendix A.

(d) **Concluding meeting** - A fourth meeting was devoted to preparing some guidelines for constructing no-DIF translated items. During this meeting, the panel was also provided with many examples of no-DIF translated items.

(e) **Sources of DIF** – Five causes of DIF found in this study were categorized. As mentioned previously, the revision process had an effect on the perception of the actual cause of DIF for each item. The four causes that were found in Allalouf et al. (1999) were expanded to five causes. The study appears to have brought us closer to the real causes of DIF. Table 1 presents these causes.

(f) **Documentation** – The main cause of DIF in each revised item was documented. Table 2 summarizes the main changes by item type for the 37 items that were revised.

Table 1 – Causes of DIF Detected and Revised in the Study

1.	**Word Difficulty (WD)**
	Definition - Accurate translation, but a word or an expression became easier or more difficult. The level of difficulty of the word(s) was not preserved. **Assessing the cause** – A judgmental decision by both source and target language speakers, plus an objective measurement , when available, regarding the relative frequency of the two corresponding lexical units (words or expressions) in the source and the target language forms of the test. The relative frequency can be assessed separately by a native speaker of each of the two languages and then compared.
2.	**Content (CO)**
	Definition – A significant difference in the meanings and connotations of the two corresponding words (or expressions). This includes cases in which the translated word has a different set of meanings (for example, translating a word that has one meaning into a word that has more than one meaning). **Assessing the cause** – A judgmental decision that involves both source and target language speakers. Dictionaries are essential for making a well-founded decision. **Marginal cases** – lack of lexical uniformity, i. e. the use of several close synonyms in the target language instead of the same word and its derivatives used in the source language (In some languages, it is preferred to avoid repetition of words in writing.)
3.	**Format (FO)**
	Definition - A different order of clauses in a complex sentence or a change in the subject of a sentence when translating some specific grammatical constructions that have no parallel in the target language. The translation could be problematic if it results in longer sentences. **Assessing the cause** – A judgmental decision that involves both source and target language speakers. Format differences are relatively easy to assess, since they are usually obvious. **Example** - In a translated sentence completion item, words that originally appeared only in the stem appeared instead in all four alternative responses, thus making the item awkward. **Marginal example** – Different numbers of blanks in a sentence completion item.
4.	**Grammatical Form (GF)**
	Definition - Use of a different grammatical form in the target language translation, usually because of differences in the structure of the two languages, or incorrect use of the same form. **Assessing the cause** – Linguistic experts are necessary for finding this cause of DIF. **Example** - Using the infinitive in place of the past indefinite, or a literal, and obviously incorrect, translation of always-plural words (*Pluralia Tantum*) - words which end in a plural suffix, have a plural meaning and do not have a singular counterpart to a singular form.)
	5. Idiomatic Relationship (ID) **Definition** – Individual words used in the item form an expression in one language, but not in the other language. **Assessing the cause** – Test translators' awareness of this cause is low, since it is a rare problem.

Table 2 – Main DIF Causes by Item Type

Item Type [a]	DIF Items	Word Difficulty	Item Content	Item Format	Grammatical Form	Idiomatic Relationship
AN	**14**	3	7	--	3	1
SC	**13**	2	4	7	--	--
LO	**3**	--	3	--	--	--
RC	**7**	1	4	1	1	--
ALL	**37**	**6**	**18**	**8**	**4**	**1**

[a] AN – Analogies, SC – Sentence Completions, LO – Logic, RC – Reading Comprehension

It can be seen that the main DIF cause, across almost all item types, was content difference (18 out of 37 items). This is followed by item format (8), word difficulty (4), grammatical form (4), and idiomatic relationship (1).

Most of the original DIF items (37 out of 42) were revised. The six test sections that originally contained the DIF items were reconstructed using the reformulated DIF items, and were re-administered in April 1999.

DIF Method

The Mantel-Haenszel (MH) DIF detection method was applied in this study. The MH method (Holland & Thayer, 1988) is used to determine whether reference and focal group item performance is equal at various ability levels along the ability continuum. The DICHODIF computer program (Rogers, Swaminathan & Hambleton, 1994) was used. The DIF classification rules used in this study were based on the DIF classification rules of the Educational Testing Service (Dorans & Holland, 1993). Two categories were defined: (1) Large (the absolute value of MH D- DIF was at least 1.5) and (2) Moderate (the absolute value of MH D- DIF was at least 1.0). In this study, a statistically significant difference in performance (at the 0.05 level) between the reference and focal

groups was found for both categories. A DIF detection design was used for each of the six test sections separately.

Results

The number of examinees in each of the six test sections appears in Table 3. The abbreviations OH, OR and RR stand for Original Hebrew, Original translated Russian and Revised Russian, respectively. These abbreviations are used in the table below and throughout the paper. It should be mentioned, that in the RR test sections, some of the items were revised, and the other items, which did not exhibit DIF, were re-administered exactly as they were administered in the OR sections.

Table 3 – Number of Examinees in Each of the Test Sections [a]

Section	OH	OR	RR (April 1999)
1	6298	1501	395
2	6298	1501	389
3	5837	2033	435
4	5837	2033	453
5	7150	1485	380
6	7150	1485	840

[a] OH = Original Hebrew, OR = Original translated Russian, RR = Revised Russian.

Item Difficulty Correlation

If DIF is reduced or eliminated, the correlation of item difficulty across language groups is expected to increase. Table 4 presents the item difficulty (Pi Index) correlations for items that exhibited DIF and those that did not exhibit DIF for OH, OR and RR.

Table 4 - Item Difficulty Correlations for DIF and No-DIF Items [a b]

Previous Study							
81 No-DIF Items				37 DIF Items			
		OH	OR			OH	OR
Previous Study:	OR	.855	----	Previous Study:	OR	.387	----
Current Study:	OR	.843	.934	Current Study:	RR	.572	.822

[a] OH = Original Hebrew, OR = Original translated Russian, RR = Revised Russian.

[b] Source: Allalouf et al. (1999).

The cross lingual correlations for items with no-DIF (.855, .843) were very similar. This was expected, since these items were not revised. For the DIF items, there was a great increase in the cross lingual correlations, from a low correlation of .387 to .572. This increase indicates that the revisions were successful in increasing the similarity in the level of difficulty of items, i.e. they exhibited less DIF. The correlations between Russian forms were very high for the no-DIF items (.934), which was expected, as these were identical items. However, the relatively high correlation for the DIF items (.822) was not expected, since the items were revised. This result shows that even though the revision increased the similarity of the item to the Hebrew version, the revised Russian item bore much greater similarity to the original Russian item.

Table 5 presents the correlations between the Hebrew and Russian item difficulties (Pi Index) by item type (for the DIF and no-DIF items, together). It was expected that the base correlations would increase following the revisions. However, this occurred only twice: in the analogies (a great increase from .253 to .496) and in the sentence completions (a moderate increase from.710 to .785). This was due to the fact that both item types contained many revised items (56% and 42% in each item type). In logic, with only 3 revised items (only 8%), the high correlation remained unchanged; in reading comprehension (where 26% of the items were revised), the correlation dropped unexpectedly.

Table 5 - Item Difficulty Correlations for all Items by Item Type [a]

	Previous Study (OH)			
	AN	**SC**	**LO**	**RC**
All Items	25 (14)[b]	31 (13)	36 (3)	26 (7)
Previous Study[c]: **OR**	.253	.710	.889	.854
Current Study: **OR+RR**	.496	.785	.889	.746

[a] OH = Original Hebrew, OR = Original translated Russian, RR = Revised Russian.
[b] The number of DIF items appears in parentheses.
[c] Source: Allalouf et al. (1999).

DIF Analysis

DIF analyses were performed. Each revised Russian version was compared to the original Hebrew version and to the original translated Russian version. Table 6 presents the MH D-DIF statistics for the 37 DIF items that were revised for the three groups: OH, OR and RR. The MH D-DIF values for the OH-OR analysis were computed in Allalouf et al. (1999). All of the values indicated DIF, i.e. values of over 1.0. The MH D-DIF values for the OH-RR analysis were computed in the current study, with the expectation that they would be smaller, i.e. less than 1.0. The MH D-DIF values for the OR-RR analysis were computed in the current study, with the expectation that they would be greater than 1.0, i.e. that the revision changed the item difficulty. The table also shows where DIF was reduced. Four categories of the change in DIF state were defined:

1. Significantly Reduced (S-RED) The MH D-DIF absolute value decreased (OH-RR value is less than OH-OR value), while the OR-RR MH D-DIF absolute value was above 1.
2. Reduced (RED): The MH D-DIF absolute value decreased (OH-RR value is less than OH-OR value), but the OR-RR MH D-DIF value was less than 1 (not significant).
3. Significantly Increased (S-INC): The MH D-DIF absolute value increased (OH-RR value is higher than OH-OR value), while the OR-RR MH D-DIF absolute value was above 1.
4. Increased (INC): The MH D-DIF absolute value increased (OH-RR value is higher than OH-OR value), but the OR-RR MH D-DIF value was less than 1 (not significant).

In addition, all items that changed from DIF to No-DIF were marked by (n).

Table 6 - MH D-DIF Statistics for the 37 Revised DIF Items

Item Type	Item No.	MH D-DIF [a] OH - OR	OH-RR	OR-RR	Success [b]
Analogies 14 DIF items Were revised	1	1.84	1.67	-0.41	RED
	2	1.86	0.12 n	-1.90	S-RED
	3	1.52	1.29	0.21	RED
	4	2.18	3.08	0.86	INC
	5	1.55	2.81	1.71	S-INC
	6	1.97	0.25 n	-2.08	S-RED
	7	-2.12	-0.93 n	1.15	S-RED
	8	5.69	2.00	-3.67	S-RED
	9	2.33	1.03	-1.42	S-RED
	10	3.25	2.77	-0.52	RED
	11	1.64	1.47	-0.46	RED
	12	-2.71	-2.48	0.27	RED
	13	3.26	2.66	-0.66	RED
	14	2.22	3.46	1.31	S-INC
Sentence Completions 13 DIF items Were revised	1	1.57	0.55 n	-1.17	S-RED
	2	-1.81	-0.03 n	1.52	S-RED
	3	1.43	1.07	-0.31	RED
	4	2.10	1.67	-0.39	RED
	5	-1.24	-1.92	-0.83	INC
	6	-1.75	-0.88 n	0.72	RED
	7	-1.67	0.51 n	1.92.	S-RED
	8	1.52	0.83 n	-0.65	RED
	9	-1.16	-2.10	-0.85	INC
	10	1.77	1.34	-0.43	RED
	11	-2.04	0.32 n	1.90	S-RED
	12	1.99	2.26	0.18	INC
	13	1.11	0.37 n	-0.45	RED
Logic 3 DIF items Were revised	1	-1.55	0.29 n	1.65	S-RED
	2	-1.22	-1.54	-0.29	INC
	3	-1.15	-0.29 n	1.32	S-RED
Reading Comprehension 7 DIF items Were revised	1	-1.99	-1.73	0.08	RED
	2	-1.33	-1.35	-0.02	NO CHANGE
	3	-1.14	-1.12	0.02	NO CHANGE
	4	-1.75	-2.18	-0.39	INC
	5	1.04	0.72 n	-0.25	RED
	6	1.19	-0.29 n	-1.37	S-RED
	7	1.05	0.71 n	-0.30	RED

Summary

FOLLOWING THE REVISIONS, THE DIF:

SIGNIFICANTLY - **RED**UCED:	12 items →	**S-RED + RED: 27 items**
REDECED:	15 items →	
NO CHANGE:	2 items	
INCREASED:	6 items →	**S-INC + INC: 8 items**
SIGNIFICANTLY - **INC**REASED:	2 items →	
		Became **No DIF (n): 15 items**

[a] OH = Original Hebrew, OR = Original translated Russian, RR = Revised Russian.

[b] For the detailed four categories, see above, page 12; In short, the success is based on the difference between the OH-OR and OH-RR values.

[n] The item did not display DIF after revision

Success Ratio

The success ratio can be calculated in two ways: (1) a success ratio of **73 percent** (Significantly Reduced + Reduced: 27 items; 27/37=73%), (2) a significant success ratio of **33 percent** (Significantly Reduced: 12 items; 12/37=33%). Overall, the challenge of reducing DIF was met. However, in 22 percent of the items, the revision increased the DIF, and in 8 percent, this increase was significant. Why? This will be referred to later in the paper. Table 7 presents the success in reducing DIF by item type.

Table 7 – Success in Reducing DIF by Item Type

Type	Revised	DIF was Reduced			DIF was Increased		
		Total	S-RED	RED	Total	S-INC	INC
AN	14	11	5	6	3	2	1
SC	13	10	4	6	3	--	3
LO	3	2	2	--	1	--	1
RC [b]	7	4	1	3	1	--	1
All	**37**	**27**	**12**	**15**	**8**	**2**	**6**

[a] AN – Analogies, SC – Sentence Completions, LO – Logic, RC – Reading Comprehension
[b] In two items, DIF did not change

Table 7 shows that DIF was reduced in most of the revised items for each item type. It is interesting that the only item type for which the revision enlarged the DIF significantly (in 2 items) was in the analogies.

Analysis of Causes

Table 8 presents an analysis of the DIF causes by item type. The analysis is based on the main DIF cause as was detected in the study.

Table 8 – For which DIF Causes did the Revision Succeed? Number of DIF Items, Reduced DIF Items and Significantly Reduced Items By Item Type and for All Items [a]

		Main DIF Cause:				
Item Type	**DIF Items**	**Word Difficulty**	**Content**	**Format**	**Grammatical Form**	**Idiomatic Relationship**
AN	**14**	3 (3) [2]	7 (5) [2]	--	3 (3) [1]	1(0) [0]
SC	**13**	2 (1) [0]	4 (3) [1]	7 (6) [3]	--	--
LO	**3**	--	3 (2) [2]	--	--	--
RC	**7**	1 (1) [1]	4 (4) [0]	1 (1) [0]	1(0) [0]	--
ALL	**37**	**6 (5) [3]**	**18 (14) [5]**	**8 (7) [3]**	**4 (3) [1]**	**1(0) [0]**

[a] In each cell, the first number is the number of DIF items, the second number is the number of reduced DIF items and the third number is the number of significantly reduced DIF items.

Table 8 shows that there was no difference by cause in reducing DIF. However, in those cases where there was a *significant* DIF reduction, there was a difference by cause. The revisions was found to be more effective in reducing DIF when the cause was Word Difficulty (3/6 success ratio), less effective when it was Item Format (3/8), and still less effective when the DIF cause was Content (5/18) or Grammatical Form (1/4).

Discussion

Not every item can be translated, nor can all item characteristics be preserved in the translation, but once an item is to be adapted from one language to another, every effort should be made to preserve the item characteristics. This effort should include a critical analysis of the reasons for DIF (see Cook, Schmitt and Brown, 1999 p. 31).

Substantiated findings of other studies can aid in revising a DIF item by providing the possible general causes of DIF. In addition, the revision process itself enhances the understanding of the causes of DIF, since during the revision process, additional causes of DIF may be discovered, and the data obtained after the administration of the revised items completes the picture. In our study, the revision reduced DIF in approximately 1/3 (significantly reduced DIF) to approximately 3/4 (reduced DIF) of the items.

In the introduction to this paper, four important goals that can be achieved by successful revisions were presented. These goals will now be discussed in light of the study findings:

1. Retaining translated items and maintaining item bank size – Before the revision, 37 items exhibited DIF. Following the revision, 15 items no longer exhibited DIF, i.e. these items could be retained for future use.

2. Providing a better understanding of the sources of DIF – The study shows that the revision process enhances the understanding of DIF previously provided by a review committee. The challenge of reducing DIF and the extra amount of time devoted to reformulating the item in the revision process add to the understanding of the sources of DIF. Five translated DIF sources were defined in this study.

3. Determining the sources of DIF and the item types for which revision is more effective. Several comments can be made in this regard:

A. It is possible to reduce DIF for all of the sources of DIF detected in this study.

B. Revision was more effective in significantly reducing DIF when the cause was Word Difficulty. This might be because the difficulty of individual words can be assessed easily, and finding another word is possible. Revision was less effective in significantly reducing translation DIF when the cause was Content. This might be because it is difficult to completely solve a content problem.

C. Analogies remained the most problematic item type even after revision: (1) The difficulty correlation, which was very low (Table 5), increased, but remained lowest of all item types. (2) After the revisions, only 3 analogies (out of 14) became no-DIF items. (3) This is the only item type for which (in 2 items) the revision significantly enlarged the DIF. These items are so short that any content difference greatly affects item difficulty.

D. The 13 sentence completion items that were included in this study were initially less problematic than the analogies. Their MH D-DIF values were smaller. Thus, a larger proportion of these items (7) became no-DIF items following the revisions.

E. The reading comprehension items are problematic; there was no much DIF in this item type, but after revision the difficulty correlation decreased. Currently, there is no good explanation to this result.

F. The logic items have almost no DIF; the correlation before and after the revision is very high (.889).

4. Setting up empirically based guidelines for reducing DIF in translated items for which DIF was already detected - Van de Vijver & Hambleton (1996) and others

created very useful guidelines for test adaptations in general. The five guidelines presented here are specific for translation revisions and should be used for purposes of reducing DIF.

A. Set up a revisions committee - The panel should be heterogeneous, including both source and target language speakers to facilitate constructive discussions.

B. Prepare a booklet with all of the DIF items and some non-DIF items in both languages. Present the relevant statistics for each item. Every item should be presented on a separate sheet (see Appendix for an example).

C. Give the material you prepared to each committee members for individual revision – Each member should work independently, at his own pace.

D. Set up committee meetings - The purpose is to reach a consensus regarding the optimal revision of each DIF item. It should be pointed out that translators might not always agree as to the best way of adapting an item, and sometimes several formulations of an item might be suggested.

E. Administer the revised items and then perform item analysis and DIF analysis. These analyses will indicate the success in reducing the DIF.

The findings are important from a methodological perspective. The fact that a revision can succeed indicates that the cause of DIF could have been eliminated earlier, during the translation process of the item, prior to its first administration, or even earlier, when items are selected for translation. When DIF is nevertheless discovered, implementing a revision design similar to that of the present study can eliminate or reduce the DIF ,which improves the development and the score validity of translated tests. The DIF causes that were identified in this study can provide a starting point to

researchers and test developers who want to identify the specific DIF causes in their specific translated tests.

References

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

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.

Beller, M. (1995). Translated versions of Israel's Inter-university Psychometric Entrance Test (PET). In T. Oakland & R. K. Hambleton (Eds.), *International perspective of academic assessment*. Boston, MA: Kluwer Academic Publishers.

Cook, L., Schmitt, A., and Brown, C. (1999, May). *Adapting achievement and aptitude tests: A review of methodological issues.* Paper presented at the International Conference of Test Adaptation, Washington, DC.

Curley, W. E., & Schmitt, A. P. (1993). *Revising SAT Verbal items to eliminate differential item functioning* (College Board Report No. 93-2). New York: College Entrance Examination Board.

Dorans, N. J., & Holland, P. W. (1993). DIF detection and description: Mantel-Haenszel and standardization. In P.W. Holland & H. Wainer (Eds.), *Differential item functioning*. Hillsdale, NJ: Erlbaum.

Gafni, N., & Canaan-Yehoshafat, Z. (1993, October). *An examination of differential item functioning for Hebrew- and Russian-speaking examinees in Israel*. Paper presented at the Conference of the Israeli Psychological Association, Ramat-Gan.

Gierl, M. J., & Khaliq, S. N. (2000, April). *Identifying sources of translation item and bundle functioning on translated achievement tests: A confirmatory analysis*. Paper presented at the annual meeting of the American Educational Research Association. New Orleans, LA.

Hambleton, R. K (1994). Guidelines for adapting educational and psychological tests: A progress report. *European Journal of Psychological Assessment, 10,* 229-224.

Hambleton, R. K., & Jones, R. W. (1995). Comparison of empirical and judgmental procedures for detecting differential item functioning. *Educational Research Quarterly, 18*, 21-36

Holland, P. W., & Thayer, D. T. (1988). Differential item performance and the Mantel-Haenszel procedure. In H. Wainer & H. Braun (Eds.) *Test validity*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Rogers, H. J., Swaminathan, H., & Hambleton, R. K. (1993). *DICHODIF : A FORTRAN program for DIF analysis of dichotomously scored item response data* [A computer program]. Amherst, MA: University of Massachusetts.

Tittle, C., K. (1982). Use of judgmental methods in item bias studies. In R. A. Berk (Ed.) *Handbook of methods for detecting test bias* (pp. 31-63). Baltimore: Johns Hopkins University Press.

van de Vijver, F. J. R., & Hambleton, R. K. (1996). Translating test: Some practical guidelines. *European Psychologist, 1*, 89-99.

Author's Note

Please address all correspondence to Avi Allalouf, National Institute for Testing & Evaluation, P.O. Box 26015, Jerusalem 91260, Israel; Email: avi@nite.org.il.

Appendix

Item Revision Form

The original source language version

TEXT

The original translated target language version

TEXT

Item statistics for each language version (example):

Percent choosing an alternative and correlation with a criterion

Language	Alt. 1	Alt. 2	Alt. 3*	Alt. 4
Target	**11%**	**4%**	**78%**	**7%**
Correlation	-0.19	-0.22	0.37	-0.17
Source	**12%**	**28%**	**55%**	**5%**
Correlation	-0.16	-0.07	0.41	-0.22

* Correct answer choice

The target language speakers performed better; DIF is large.

The suggested revised item in the target language

TEXT

The final revised item in the target language

----------Blank ----------
