Detecting Sources of DIF in Translated Verbal Items

Avi Allalouf
National Institute for Testing and Evaluation (NITE), Israel

Stephen G. Sireci
University of Massachusetts, Amherst

Paper presented at the annual meeting of the
American Educational Research Association
San-Diego, CA, April 1998

Detecting Sources of DIF in Translated Verbal Items

Abstract

Translated tests are being used increasingly in educational testing for assessing the knowledge and skills of individuals who speak different languages. A difficult and under-researched problem related to test translation is the identification of translated items that do not function equivalently across languages. Furthermore, there is little research exploring why translated items sometimes function differently across languages. If the sources of differential item functioning (DIF) across languages could be predicted, these sources could be eliminated at an early stage in the test development process, and this could have important implications for decisions regarding test development process, scoring and equating.

This study focuses on two questions: Is DIF related to item type? What are the sources of DIF? The data that were analyzed consisted of three forms of the Israeli Psychometric Entrance Test (PET) in Hebrew (source) and Russian (translated). The sources of DIF were assessed by analyzing the DIF direction with the aid of translators who judged the items. The results indicated that 42 out of 125 items (34%) functioned differentially across languages. The conclusions are presented by item types: Analogies - have the highest DIF rate (65%), with Russian-speaking examinees performing better on 82% of the analogy items with DIF; Sentence Completions - have relatively high DIF (45%), with no group performing better than the other; Reading Comprehension - has moderate DIF (23%), probably related to the content of the specific passage ; Logic - almost no DIF was found (8%). The main sources of DIF were changes in word difficulty, changes in item format, translation errors and differences in cultural relevance.

Translated tests are being used increasingly in educational testing for assessing the knowledge and skills of individuals who speak different languages. An example of this is the recent (1997) Third International Mathematics and Science Study, with 45 participating countries and tests prepared in 31 languages. However, translating a test from a source language to a target language does not necessarily produce two psychometrically equivalent tests. Angoff and Cook (1988, p.2) wrote: *It can hardly be expected, without careful and detailed checks, that the translated items will have the same meaning and relative difficulty for the second group as they had for the original group before translation*. Translation is one of the first stages in the long process of test adaptation across different languages (DL). The final goal of test adaptation is to maintain construct equivalence and content representation across the two language versions (Procedures for translation adaptations can be found in Hambleton, 1994). Differential item functioning (DIF) detection plays an important role in the test adaptation process. An item functions differently across groups if examinees of equal ability but from different groups (in this study, source and target languages) do not have an equal probability of correctly responding to that item.

Empirical and judgmental methods for detecting differentially functioning test items were initially developed to detect bias for different groups tested in the same language. Empirical bias studies have sometimes found consistent differences between demographic groups (i.e. blacks versus whites and males versus females), and conducting an item bias study has become an essential part of test development and test evaluation. Some of the studies (see for example, Hambleton & Jones, 1995) used judgmental and empirical procedures separately, and checked the level of agreement between them. These methods were applied later in test translations in order to detect items that function differently in the source language and the target language. Theoretically and practically, the DIF methods can be used to *detect any difference between any groups in their responses to any kind of item* (Thissen, Steinberg & Wainer, 1993).

The equivalence of the test structure in each language version should be assessed before choosing a suitable DIF method. Sireci and Swaminathan (1996) addressed this matter and presented procedures for dimensionality assessment in dealing with test adaptations.

A difficult and under-researched problem relating to test translation is the identification of translated items that do not function equivalently across languages. Furthermore, there is little research exploring why translated items sometimes function differently across languages. Few studies have dealt with detecting sources of DIF, and of these, only a few involved translated tests. If the factors affecting the DIF of translated items could be predicted, they could be taken into account at an early stage in the test development process, resulting in improved decisions regarding test constructing, scoring and equating (e.g., excluding from the equating items that function differentially across languages). DIF in test adaptations is generally not anticipated before test administration, and as a result, researchers must rely on post-hoc explanations concerning the presumed sources of DIF.

Only a few studies to date have dealt with the relationship between item content and DIF of translated items. There has clearly been some success in explaining the sources of DIF in test items. Hulin (1987) suggested a method for assessing the source of DIF in translated items using IRT parameters. According to Hulin's theory, *b*-parameter (item difficulty) difference is due to translation error, and *a*-parameter (item discrimination) difference is due to difference in cultural relevance. Ellis (1995) performed a partial test of Hulin's theory regarding the *a*-parameter difference. The study population spoke the same language (German), but came from different cultures (East Germany and West Germany); the test was the *Trier Personality Inventory* (TPI). Each culture group consisted of some 300 people. Ellis concluded that the results did not support Hulin's theory: almost all of the differences found were in the *b*-parameter and not in the *a*-parameter. However, it is likely that Ellis' sample sizes were not large enough for a stable estimation of the *a*-parameter, and consequently she could not adequately address Hulin's theory. Another conclusion from this study was the need for conducting a complementary study using different languages but involving examinees with the same cultural background. However, such a study would be very difficult to conduct because, in order to have the same cultural background, all subjects would have to be completely bilingual (people who speak different languages cannot have the same cultural background). According to Ellis, it would be very difficult to measure the degree to which a person is bilingual.

Angoff and Cook (1988) followed Angoff and Modu (1973) in an attempt to establish score equivalence between the SAT and the Spanish-language equivalent, the *Prueba de Aptitud Academica* (PAA). Two IRT methods were used for assessing DIF. Greater DIF was found in the antonym and analogy items, and smaller DIF in the sentence completion and reading comprehension items. Their explanation was: *..items with more context probably tend to retain their meaning, even in the face of translation into another language* (page 8).

In a study of the translation of the Israeli *Psychometric Entrance Test* (PET) from Hebrew to Russian, Gafni, Canaan-Yehoshafat, and Beller (Gafni & Canaan-Yehoshafat ,1993; Beller, 1995) evaluated translated items in the verbal section of the test for three different forms. The translated items in the verbal section consisted of analogies, sentence completions, reading comprehension and logic items. In order to assess DIF, the researchers used a delta-plot technique proposed by Angoff (1972). They found the greatest DIF in the analogies and the smallest DIF in the logic and sentence completion items. These results are similar to those of Angoff and Cook (1988), with the exception of the logic items which do not appear on the SAT verbal section. The reading comprehension items showed relatively higher DIF than in the Angoff & Cook study. Ellis (1989) studied the measurement equivalence of translated American and German intelligence tests using IRT methods for DIF. Her findings showed that in most items the DIF resulted from translation errors. In other cases, some post-hoc explanations were given.

In summary, a study of the sources of DIF in translated verbal items could greatly enhance current knowledge, and the need for a study of this type is great.

Purpose

The purpose of this study is to assess sources of DIF in translated verbal items. The study consisted of two parts. The first part of the study was designed to identify the types of verbal items which were most likely to display DIF when translated from one language to another. The second part dealt with detecting sources of DIF using two procedures: (1) analyzing the DIF direction (which group performed better on which items and item types), and (2) after reviewing the findings from the previous procedures, translators analyzed the type and content of the DIF items.

Three features of the study were especially important: (a) sample sizes were considerably larger than those normally available in DIF studies, (b) items displayed high DIF across languages (in DIF studies, the differences are generally smaller), and (c) Due to small group differences in ability, identification of DIF was not confounded with type I errors.

Method

Instruments

The test which was analyzed was the verbal subtest of *PET - Psychometric Entrance Test* - which is a high-stakes test used for admissions to universities in Israel (see Beller, 1994). It is a multiple choice test consisting of three subtests: verbal, quantitative, and English as a foreign language. It is written in Hebrew and translated into Arabic, Russian, French, Spanish, and English. The study dealt with the translation from Hebrew to Russian.

The verbal subtest consists of 60 items. Item types include synonyms and antonyms (not translated), analogies, sentence completions, logic, and reading comprehension (almost all are translated). Since there are some verbal items that cannot be translated, new items are constructed specifically for the translated form. Three test forms that were administered in both Hebrew and Russian were used in this study. Three forms were used in order to increase the number of items investigated in the study and to provide a basis for replication. Table 1 presents the types of the 125 (70% of the 180 items included in the three forms) common items in the three forms analyzed in the study.

Table 1

Type, Number of Items & Abbreviations for the Hebrew/Russian Common Items

Type	Number of Items	Abbreviation
Analogies	26	AN
Sentence Completions	33	SC
Logic	36	LO
Reading comprehension	30	RC
# Total	125	

Examinees

Since 1990, approximately one million people have immigrated to Israel from the former Soviet Union. As a result, the number of PET examinees who are tested in Russian has greatly increased. The main criteria for choosing the three specific forms were their large sample size and the smaller differences in verbal ability between the Russian-speaking examinees and the Hebrew-speaking examinees. The verbal ability differences between the Russian-speaking examinees and the Hebrew-speaking examinees were estimated on the basis of the differences between the quantitative scores of the Russian-speaking examinees and the Hebrew-speaking examinees. The quantitative score has a pretty stable correlation around .68 with the verbal score in both languages (see Table 2).

Dimensionality Assessment

A related study (Allalouf et al. 1997) compared the dimensional structure of the common items in the Hebrew and the translated Russian version of PET using confirmatory factor analysis (CFA) and multidimensional scaling (MDS). Both methods found that the dimensional structure of the test appears to be equivalent across the Hebrew and the Russian versions.

DIF Method

The Mantel-Haenszel (MH) DIF detection method was used on this study. The MH procedure (introduced by Holland and Thayer, 1988) is used to determine whether reference and focal group item performance is equal at various ability levels along the ability continuum. The MH procedure provides an MH D-DIF index, given in the delta metric. The method does not require large samples and has a χ^2 index for testing statistical significance. The DICHODIF computer program (Rogers, Swaminathan & Hambleton, 1994) performs this analysis and was used in this study.

The DIF classification rules applied in this study were based on the DIF classification rules of the Educational Testing Service (Dorans & Holland, 1993). Two categories were defined: (1) Large - an item was defined as having large DIF if the absolute value of MH D- DIF was at least 1.5 and statistically significant greater than

zero (at the 0.05 level). This classification is similar to ETS category "C"; (2) Moderate - an item was defined as having moderate DIF if the absolute value MH D-DIF was at least 1.0 (and less than 1.5) and statistically significant greater than zero (at the 0.05 level). This classification is similar to ETS category "B".

Part 1: Identifying DIF across languages

A DIF detection design was used separately for each of the three test forms: 1, 2 and 3, which had 40, 44 and 41 common translated items respectively. In order to refine the matching criteria, the DIF detection process involved two stages. The first stage used the total raw score as the stratifying variable. The second stage used the total score of all items that were not flagged as having large DIF (during the first stage) as the stratifying variable. Based on the second stage results, items were classified into one of three categories: no DIF, moderate DIF, and large DIF.

Part 1: Results

Summary statistics for the three test forms in each language are presented in Table 2.

Table 2

Means, Standard Deviations, for the three forms, in each test language for Verbal Common Items Raw Score and Quantitative Score, and Correlations between the Scores

		Verbal Common Items [a]			Quantitative [b]		
Form	Group	N	Mean	SD	Mean	SD	Correlation
1	Hebrew	6298	23.2	7.6	106.2	18.4	.719
	Russian	1501	21.0	6.9	105.2	18.0	.698
2	Hebrew	5837	29.2	8.3	112.6	19.1	.685
	Russian	2033	25.3	7.4	110.3	18.5	.647
3	Hebrew	7150	24.1	7.9	105.1	18.7	.700
	Russian	1485	22.6	6.7	104.1	18.2	.657

a The raw score is equal to number right. There were 40 items in Form 1, 44 items in Form 2, and 41 items in Form 3.

b These scores have a Mean = 100 and SD = 20 on a basis sample.

Sample sizes were large, averaging over 6,400 for the Hebrew forms and close to 1,700 for the Russian forms. Based on the quantitative score, it can be seen that the ability differences between the Russian examinees and the Hebrew examinees was small. This is an advantage, since the more similar the groups, the more accurate is the DIF detection. The differences between the correlations with the quantitative score are small and are probably due to the differences in variance. The similarity of the verbal-quantitative correlations across the Hebrew and Russian groups is consistent with the expectations that the verbal test is measuring the same construct in both languages.

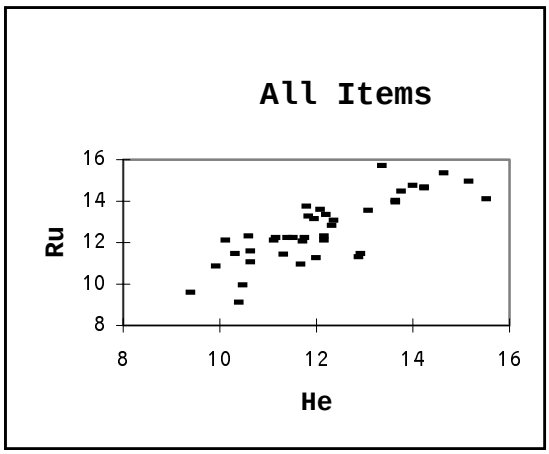
Analysis of DIF

The DIF analysis identified the items that displayed large and moderate DIF in each form. Appendix A presents the MH D-DIF values for all of the items in the comparison of the Hebrew and Russian-language test forms. In Form 1, out of the 40 items, 7 displayed large DIF and 6 displayed moderate DIF (total DIF items = 13, 33%); in Form 2, out of the 44 items, 11 displayed large DIF and 6 displayed moderate DIF (total = 17, 39%); in Form 3, out of the 41 items, 8 displayed large DIF and 4 displayed moderate DIF (total = 12, 29%).

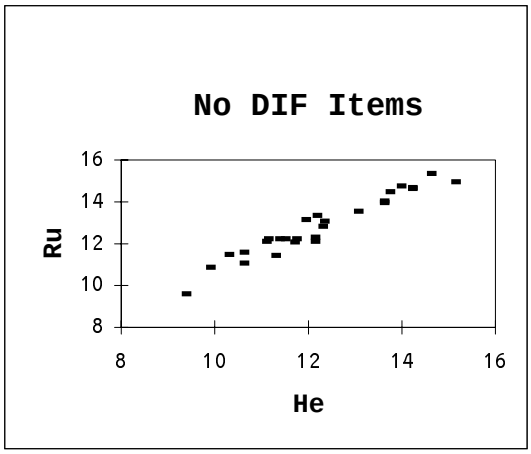
Figure 1 presents delta (item difficulties transformed to the delta metric) plots for all of the translated items in the three forms. Two plots are presented for each form. The plots on the left represent all of the common items; the plots on the right display only those items that were not flagged for moderate or large DIF. The correlation between the Hebrew and Russian deltas ranges from .64 to .82 for all common items, and from .91 to .97 for the non-DIF translated items.

Table 3 presents data regarding the number of items displaying DIF out of all the common items in each form, by degree of DIF: Large (absolute value of MH D-DIF is at least 1.5 and significantly greater than zero at the 0.05 level), and Moderate (the absolute value of MH D- DIF is at least 1.0 and less than 1.5, and significantly greater than zero at the 0.05 level).

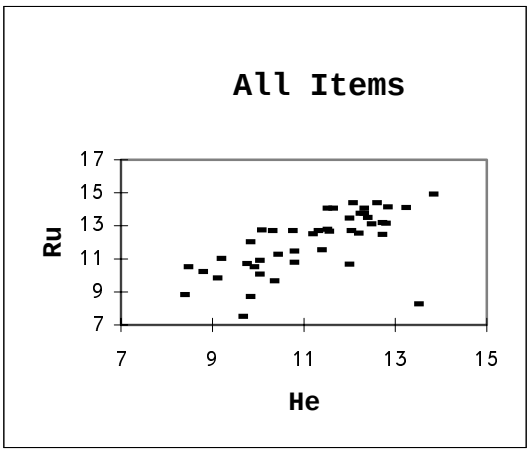
Figure 1 - Delta Plots, and correlations between Russian & Hebrew deltas



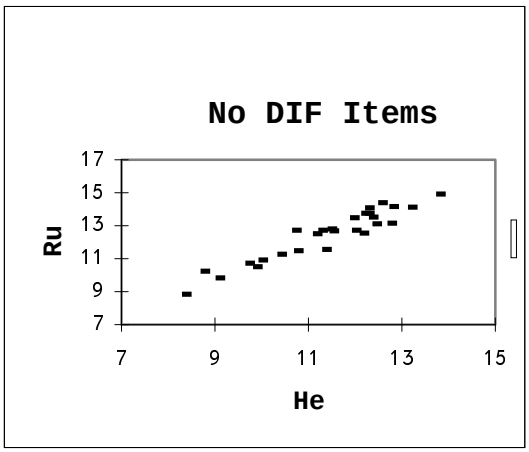
Form 1 r = 0.82



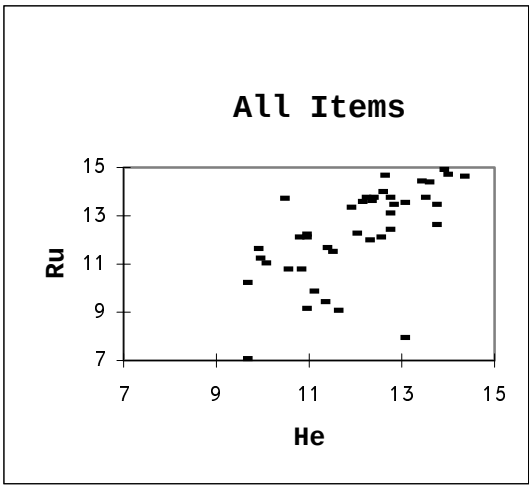
Form 1 r = 0.97



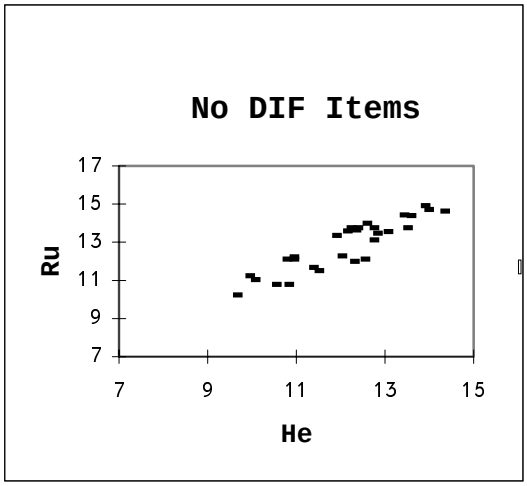
Form 2 r = 0.66



Form 2 r = 0.95



Form 3 r = 0.64



Form 3 r = 0.91

FIGURE 1 - Delta Plots for the three Forms

Table 3

Percentage [a] of Items Displaying DIF

in the Comparison of the Hebrew and Russian Test Forms by Item Type

Item Type	Form 1		Form 2		Form 3		**Total**		
	L	M	L	M	L	M	**L**	**M**	**L&M**
Analogies	29%	29%	60%	0%	78%	0%	**58%**	**7%**	**65%**
Sentence Completions	27	18	33	25	8	17	**24**	**21**	**45**
Logic	0	0	8	9	0	8	**3**	**5**	**8**
Reading Comprehension	20	20	0	20	0	10	**7**	**16**	**23**
Total	18	33	25	39	19	29	**21**	**14**	**34**

a Out of the same item types; L = large DIF, M = moderate

The findings are very consistent: The greatest DIF was found on the analogies and to a certain extent on the sentence completion items, and a lower level of DIF was found for the logic and the reading comprehension items. Prediction of the MH D-DIF values using the content area designations for each item (out of all the 125 items) revealed that the regression weight for analogies is significant (RSQ = .163 , F = 23.96 P < .0001). The results are similar to those of Angoff & Cook (1988), and Gafni & Canaan-Yehoshafat (1993) and Beller (1995), which were described earlier. However, they differ from the two previous studies in two ways: There were more sentence completion items that showed DIF in the current study than in those two studies, and there were relatively less reading comprehension items that showed DIF than in the Gafni & Canaan-Yehoshafat (1993) study. The explanation by Angoff & Cook (1988) that translated verbal items with more context are more likely to retain their meaning might also be applicable to our results.

Part 2: Detecting sources of DIF

Once items were identified as functioning differentially in Hebrew and Russian, two procedures were used for detecting possible sources of DIF. First, the direction of DIF was determined. This analysis recorded the group (Hebrew or Russian) that performed better on each DIF item. This analysis was also conducted separately for each item type (content area) to address the question of whether the translation made the item easier or more difficult when translated into Russian for any or all of the content areas. The second procedure for discovering potential sources of DIF used five Hebrew-Russian translators to analyze the type and content of the DIF items. Of the 60 items given to these translators for analysis, 42 were previously identified as DIF items; the other 18 were not flagged for DIF. The translators were unaware of the DIF classifications of the items. After exaplanation of the term "DIF," each translator independently completed an item review questionnaire (see Exhibit 1) for each item. The translators took about ten hours to complete these questionnaires.

Exhibit 1: Questionnaire given to the translators

Translator's name _____ Date _____

Some of the 60 items provided were identified as functioning differently across groups. This means that Hebrew and Russian examiees with the same ability had a different likelihood of answering the item correctly. Make sure that you understand the concept of differential item functioning (DIF) before completing these questionnaires. For each item, please answer the following questions:

1. The correct answer for the item is _____
2. Can you rate the item distractors - the most attractive: ___ , the least attractive: __
3. Do you anticipate DIF in this item? _____

>>. If the answer to question 3 is positive, proceed to the following questions (4,5,6).

4. Is the DIF large or moderate? _____
5. For which language group is the item more difficult? _____

Explain _____

6. What was the reason for the DIF? _____
7. Additional comments. _____

After analyzing the translators' answers to the questionnaire, an eight-person panel was formed. The panel included the five translators and three Hebrew-speaking researchers. Panel members reviewed each item separately, using the statistical data and the questionnaire results. The purpose of the panel was to arrive at conclusions regarding the possible sources of DIF in each DIF item, and to draw general conclusions about sources of DIF in translated verbal items

Part 2: Results

a. DIF Direction

Looking for sources of DIF involves an analysis of the "DIF direction": Is there any relationship between item type and group performance on the DIF items? A total of 42 moderate and large DIF items were found. The Russian-speaking examinees

performed better on 25 of these items, and the Hebrew-speaking examinees performed better on 17 of them. The proportion was 60:40, which is close to the expected 50:50 ratio (the DIF detection method assumes that the test is generally not biased). Table 4 presents an item type analysis of the 25 DIF items on which the Russian-speaking examinees performed better than the Hebrew-speaking examinees.

Table 4

Number of DIF Items and Percentages [a] of the 25 DIF Items on which the Russian-Speaking Examinees Performed Better than the Hebrew-Speaking Examinees

Item Type	Form			**Total**	
	1	2	3	Percentages	DIF Items
Analogies	75%	83%	86%	**82 %**	**14/17**
Sentence Completions	80	29	67	**53**	**8/15**
Logic	0	0	0	**0**	**0/3**
Reading Comprehension	0	100	100	**43**	**3/7**
Total	54	53	75	**60**	**25/42**

[a] As percentages of the same item type that showed DIF in each form and in all three forms (total).

The findings are interesting. The Russian examinees performed much better than the Hebrew examinees on the analogy items with DIF. The translation might have made the analogies much easier in Russian than in Hebrew. Another possible explanation is that the Russian-language examinees are better at analogies than the Hebrew-language examinees. In the sentence completion DIF items, no group outperformed the other group. Because of the small number of DIF items in logic and in reading comprehension, it was not possible to draw conclusions at this stage.

Mean p-value differences

Tables 5A , 5B & 5C presents the mean p-value for the items in both languages, by item type: for all the items, for the DIF items, and for the no DIF items. The p-values of the analogies demonstrate again that the Russian examinees performed much better than the Hebrew examinees on the analogy items.

Table 5A, 5B & 5C Mean P-Values, by Language and Item Type

5A **All Items**	All Items (125)	Analogies (26)	Sentence Completion (33)	Logic (36)	Reading Comprehension (30)
Hebrew	61	66	67	56	58
Russian	55	68	60	48	47
Difference	6	-2	7	8	11

5B **DIF Items**	All Items (42)	Analogies (17)	Sentence Completion (15)	Logic (3)	Reading Comprehension (7)
Hebrew	64	64	65	72	59
Russian	62	72	59	52	47
Difference	2	-8	6	20	12

5C **No DIF Items**	All Items (83)	Analogies (9)	Sentence Completion (18)	Logic (33)	Reading Comprehension (23)
Hebrew	60	69	68	55	58
Russian	51	61	60	47	47
Difference	9	8	8	8	11

b. Panel of Translators & Researchers

Based on the translators' work, the panel found four main causes for DIF in the 42 DIF items (Appendix B presents the reasons for DIF in these items):

1. **Changes in Difficulty of Words or Sentences -** The translation was accurate, but some words or sentences became easier or more difficult. For example: An analogy item contained a very difficult word in the stem that was translated into a very trivial word. The translator was not aware of the difficulty of the original word, or of the importance of preserving that difficulty.

2. Changes in Content - This can be labeled a translation problem. The meaning of the item changed in the translation, thus turning it into a different item. This could be due

to an incorrect translation that changed the meaning of the item or the translation of a word that has a single meaning into a word that has more than one meaning.

3. Changes in Format - The format of the item was changed. For example: A sentence became much longer. Another example: In a translated sentence completion item, words that originally appeared only in the stem now appeared instead in all four alternative responses, thus making the item awkward. It should be noted that, due to constraints of the Russian language, translating the item in this way could not be avoided.

4. Differences in Cultural Relevance- The item remained exactly the same, but the two groups differed because of the culture content of the specific item. This could be the content of a reading comprehension passage that was more relevant to one of the groups, or the content of a sentence completion item that was more familiar to one of the groups.

Analyses of the causes for DIF by item type revealed that in analogy items the DIF was usually due to differences in word difficulty or translation errors. In the sentence completion items the DIF was due to all four causes stated above. In the reading comprehension items, the DIF was probably due to cultural relevance. On seven of the items, which constitute 17% of the 42 items displaying DIF (including all 3 logic items), the panel could not reach any agreement regarding the DIF source. Table 6 summarizes the reasons for DIF by item type.

Table 6

Reasons for DIF by Item Type

Item Type	Reasons for DIF					**Total**
	Changes In			Cultural Relevance	Unknown Reason	
	Word Difficulty	Item Content	Item Format			
Analogies	12	5	-	-	-	**17**
Sentence Completions	4	3	5	1	2	**15**
Logic	-	-	-	-	3	**3**
Reading Comprehension	-	-	-	5	2	**7**
Total	16	8	5	6	7	**42**

The translators pointed out that the two languages differ in the number of difficult words which they contain. This fact makes the task of translating, while preserving equal difficulty of the words, almost impossible. The verbal subtest also consists of antonyms and synonyms which are constructed especially for the Russian version of the test. The translators, who construct these items, say that the Russian language contain less difficult words to choose from. In Hebrew there are two kinds of difficult words: 1. Old words, usually from the Bible, 2. Modern words which were invented in the last century and are rarely used (e.g.; Hebrew word for answering machine, banana etc.) In Russian these two sources for difficult words do not exist. Languages also differ in the length of their sentences. For example, a 30 line passage in Hebrew is translated into a 45-50 line passage in Russian. It was anticipated that the different lengths would be the reason for all, or at least most, of the DIF in the reading comprehension items, but such an effect was not found.

Discussion

Understanding DIF sources in translated verbal items is an area which warrants further empirical research. The current study focused on two questions: Is DIF related to item content or item format? What are the sources of DIF ?The data was taken from three forms in Hebrew (source) and in Russian (translated) of the verbal subtest of the Israeli PET. DIF was detected using the Mantel Haenszel method. The sources of DIF were assessed by analyzing the DIF direction and by using translators to analyze the type and content of the DIF items. The conclusions of the study can be arranged by item type and by DIF source.

Conclusions by item type

Analogies: This was the most problematic of the four item types. Many of the analogy items displayed DIF (65% of the items: 59% displayed large DIF, 7% displayed moderate DIF). An interesting finding is that in 82% of the DIF items, the Russian examinees performed better than the Hebrew examinees. One of the main factors affecting the difficulty of analogies is word difficulty (see Bejar et al., 1991, Roccas & Moshinsky, 1997). In translating an analogy, translators often choose an easier word, which makes the analogy easier. Of course, they could theoretically choose words that are more difficult, but this was not the case in this study. The few

items on which the Hebrew speakers performed better contained translation errors. Another possibility that should be mentioned but which seems to us to be highly unlikely is that the Russian-speaking examinees perform better on analogies.

Sentence Completions: This item type was also problematic, but less so than analogies. Many sentence completion items displayed DIF (45% : 24% large DIF, 21% moderate DIF). Neither group outperformed the other (in 53% of the DIF items the Russian-speaking examinees performed better). Due to language differences, it is difficult to retain the exact format of the sentences. This was the main reason for DIF in the sentence completion items. The second main reason for DIF was word difficulty.

Logic: The logic items exhibited almost no DIF. Only one item out of 36 displayed large DIF, and only two additional items showed moderate DIF. These results are consistent with the conclusion of Angoff & Cook (1988) that the longer the question, the less DIF that exists. They are also consistent with a new idea: It is easier to translate and to preserve the various characteristics of verbal logic items because the logic component of the question is less dependent upon a translation that provides an exact meaning. In this study, this item type was the best type for test adaptations.

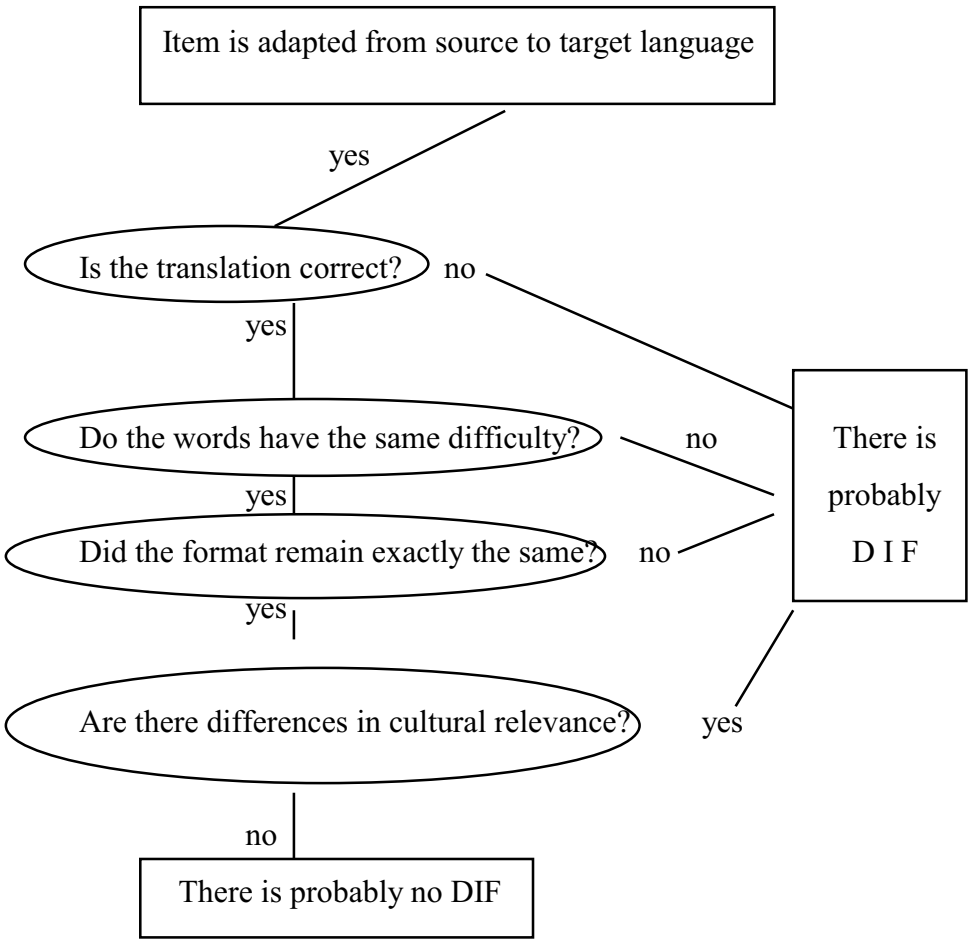
Reading Comprehension: The reading comprehension items displayed a relatively small amount of DIF. There were differences in the DIF of the six passages which were analyzed (see Appendix A). The items that displayed large DIF involved two (out of six) passages. From these results, it appears that DIF is related to the specific content of the passage. For example, three DIF items were found in a passage whose content dealt with economics, and it could be argued that the Russian speakers were less familiar with this area.

DIF Sources

Based on the results of this study, a basic flow chart representing the process involved in identifying DIF sources in an item is presented in Figure 2. This flow chart can be used routinely by the translators.

Figure 2

Flow chart for DIF sources found in this study



One limitation of this study is that there were only two specific languages involved. However, the conclusions do not appear to be language-specific. Replicating this study using other languages, and performing simultaneous analyses of "multiple DIF" between several languages is desirable.

Further research on the sources of DIF in translated items can be designed on the basis of the following ideas:

(1) Focusing on the items that did not display DIF, for example in the analogy items. How are these items different from the analogies exhibiting DIF?

(2) Give a questionnaire to examinees in both groups and asking them to explain their answer. Analysis of the explanations of those who missed DIF items may bring better understand of the causes of changes in item difficulty between the groups

(3) Changing the items with DIF (for example, changing the words difficulty) and retesting for DIF (e.g. Curley & Schmitt, 1993).

(4) Finally, for the problematic item types, the desirable degree of difficulty may be achieved only by constructing some items directly in the target language, and than translate them to the "source" language. A DIF study on these items would be of interest.

A translated test can be equated to the source test even if some items in the target language function differently from the source items. This is because only non DIF translated items should be used for equating. However, the smaller the number of DIF items, the better the translated test: it will be more similar to the source test, and the equating will be more accurate.

In summary, and most importantly for future test translation work, identification of the sources or causes of DIF is not likely to be achieved *only* through statistical analysis of item response data. Statistical analyses are very helpful in detecting problematic items, but they do not reveal the causes of the problems.

While the results are of value to individuals interested in cross-national and cross-cultural studies, the findings are also important methodologically. The study design combined with the analysis methods used in this study can both identify DIF items and also identify the causes of DIF for most of the items detected as DIF in other studies in this area as well. Such identifications can improve the development of translated tests and enhance their score validity.

References

Allalouf, A., Bastari, Sireci, S. G., & Hambleton, R. K. (1997, October). *Comparing the dimensionality of a test administered in two languages.* Paper presented at the meeting of the Northeastern Educational Research Association, Ellenville, NY.

Angoff, W. H. (1972). *A technique for the investigation of cultural differences*. Paper presented at the meeting of the American Psychological Association, Honolulu. (ERIC Document Reproduction Service No. ED 069686).

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.

Angoff, W. H., & Modu, C. C. (1973). *Equating the scores of the Prueba de Aptitud Academica and the Scholastic Aptitude Test* (Research Report No. 3). New York: College Entrance Examination Board.

Bejar I. I., Chaffin, R., & Embertson, S. (1991) *Cognitive and psychometric analyses of analogical problem solving*. New York: Springer-Verlag.

Beller, M. (1994). Psychometric and social issues in admissions to Israeli universities. *Educational Measurement: Issues and Practice, 13*, 12-21.

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.

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 and H. Wainer (Eds.) *Differential item functioning*. Hillsdale, NJ: Erlbaum.

Ellis, B.B. (1989). Differential item functioning: Implication for test translation. *Journal of Applied Psychology*, 74, 912-921.

Ellis, B.B. (1995). A partial test of Hulin's psychometric theory of measurement equivalence in translated tests. *European Journal of Psychological Assessment, 11,* 184-193.

Gafni, N & Canaan-Yehoshafat, Z. (1993). *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.

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., and Jones, (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.

Hulin, C. L. (1987). A psychometric theory of evaluations of item and test translations: Fidelity across languages. *Journal of Cross-Cultural Psychology*, 67, 115-142.

Roccas, S. & Moshinsky, A. (1997). *Factors affecting the difficulty of verbal analogies*. NITE report no. 239, National Institute for Testing and Evaluation, Jerusalem.

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.

Scheuneman, J. D. (1987) An experimental, exploratory study of the causes of bias in test items. *Journal of Educational Measurement*, 24, 97-118

Schmitt, A. P., & Bleistein, C. A. (1987). *Factors affecting differential item functioning of black examinees on Scholastic Aptitude Test analogy items* (Research Report 87-23). Princeton, NJ: Educational Testing Service.

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

Sireci, S. G., & Swaminathan H. (1996, October). Evaluating translation equivalence: So what's the big dif? Paper presented at the annual meeting of the Northeastern Educational Research Association, Ellenville, NY.

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

Thissen, D., Steinberg, L. & Wainer, H. (1993). Detection of differential item functioning using the parameters of item response models. In P.W. Holland & H. Wainer (Eds.) *Differential item functioning*. Hillsdale, NJ: Erlbaum.

Appendix A

MH D-DIF values for all of the items in the Hebrew-Russian comparison, by form. Positive values of MH D-DIF mean that the Russian examinees performed better. (AN = Analogies, SC = Sentence Completion, LO = Logic, RC = Reading Comprehension).

Form 1

Item No.	Item Type	MH D-DIF
1		1.91
2		.33
3	AN	-.65
4		-.70
5		1.92
6		.76
7		-.31
8	SC	-.55
9		1.48
10		.53
11		1.95
12		.47
13		0.02
14	LO	.12
15		-.01
16		-.55
17		-.14
18		.11
19		-.76
20	RC	.08
21		-1.95
22		-.14
23	AN	1.49
24		-1.03
25		-1.74
26		-.25
27	SC	1.49
28		2.29
29		.84
30		.21
31		.69
32	LO	31
33		-.01
34		.45
35		.18
36		-1.30
37		-.22
38	RC	-1.07
39		-.88
40		-1.68

Form 2

Item No.	Item Type	MH D-DIF
1		2.32
2		1.50
3	AN	-.39
4		2.10
5		-1.27
6		-1.73
7	SC	.18
8		.32
9		-1.65
10		1.40
11		-1.53
12		-.32
13	LO	.36
14		-1.31
15		-.94
16		-.84
17		.25
18		-.47
19	RC	.24
20		-.22
21		-.77
22		3.38
23		.19
24	AN	-.63
25		-1.96
26		.24
27		5.77
28		.59
29		-1.02
30		2.03
31		-1.87
32	SC	.16
33		-.34
34		-.01
35		.88
36	LO	-.19
37		-.32
38		.67
39		-.27
40		.50
41		1.17
42	RC	1.30
43		-.24
44		-.31

Form 3

Item No.	Item Type	MH D-DIF
1		-.05
2		2.34
3	AN	.82
4		3.30
5		1.69
6		-.80
7		-.38
8	SC	2.09
9		.16
10		1.04
11		-1.13
12		-.71
13	LO	.04
14		.29
15		.25
16		.30
17		-.71
18		-.69
19	RC	-.08
20		-.71
21		-.31
22		-2.63
23	AN	3.42
24		2.42
25		6.09
26		-.72
27		.26
28	SC	-.74
29		-1.49
30		-.83
31		.83
32		.72
33	LO	.34
34		.27
35		-.14
36		-.21
37		.58
38		1.09
39	RC	.92
40		-.87
41		-.23

Appendix B - Sources of DIF in the 42 DIF items

Item Type	No	Item	Better Performance by	Explanation for DIF
	1	1	Russian Speakers	Word Difficulty
	2	3	Russian Speakers	Word Difficulty
	3	11	Russian Speakers	Word Difficulty
	4	12	Hebrew Speakers	Translation Problem
	5	20	Russian Speakers	Word Difficulty
	6	21	Russian Speakers	Word Difficulty
	7	22	Russian Speakers	Word Difficulty
Analogies	8	32	Russian Speakers	Translation Problem
(AN)	9	34	Hebrew Speakers	Translation Problem
	10	36	Russian Speakers	Word Difficulty
	11	45	Russian Speakers	Translation Problem
	12	46	Russian Speakers	Word Difficulty
	13	47	Russian Speakers	Word Difficulty
	14	52	Hebrew Speakers	Translation Problem
	15	53	Russian Speakers	Word Difficulty
	16	54	Russian Speakers	Word Difficulty
	17	55	Russian Speakers	Word Difficulty
	1	5	Russian Speakers	Cultural Differences
	2	7	Russian Speakers	Word Difficulty
	3	13	Hebrew Speakers	Sentence Difficulty
	4	14	Russian Speakers	Changes in Format
	5	15	Russian Speakers	Translation Problem
	6	23	Hebrew Speakers	Changes in Format
Sentence	7	24	Hebrew Speakers	Changes in Format
Completions	8	27	Hebrew Speakers	Translation Problem
(SC)	9	28	Russian Speakers	Unknown
	10	37	Hebrew Speakers	Changes in Format
	11	38	Russian Speakers	Word Difficulty
	12	39	Hebrew Speakers	Translation Problem
	13	48	Russian Speakers	Word Difficulty
	14	50	Russian Speakers	Unknown
	15	57	Hebrew Speakers	Changes in Format
Logic	1	29	Hebrew Speakers	Unknown
(LO)	2	31	Hebrew Speakers	Unknown
	3	51	Hebrew Speakers	Unknown
	1	10	Hebrew Speakers	Unknown
	2	17	Hebrew Speakers	Cultural Differences
Reading	3	18	Hebrew Speakers	Cultural Differences
Comprehension	4	19	Hebrew Speakers	Cultural Differences
(RC)	5	41	Russian Speakers	Cultural Differences
	6	42	Russian Speakers	Cultural Differences
	7	59	Russian Speakers	Unknown