Appraising item equivalence across multiple languages and cultures

Stephen G. Sireci  University of Massachusetts, Amherst and
Avi Allalouf  National Institute for Testing and Evaluation, Jerusalem

Activity in the area of language testing is expanding beyond second language acquisition. In many contexts, tests that measure language skills are being translated into several different languages so that parallel versions exist for use in multilingual contexts. To ensure that translated items are equivalent to their original versions, both statistical and qualitative analyses are necessary. In this article, we describe a statistical method for evaluating the translation equivalence of test items that are scored dichotomously. We provide an illustration of the method to a portion of the verbal subtest of the Psychometric Entrance Test, which is a large-scale postsecondary admissions test used in Israel. By evaluating translated items statistically, language test-developers can ensure the comparability of tests across languages and they can identify the types of problems that should be avoided in future translation efforts.

I Introduction

Language tests typically measure a person’s proficiency in a single language. When more than one language is involved, it is usually in the context of second language testing, for example, when the test directions or item stems are presented in the test-taker’s dominant language, but the test questions are presented in the second language of interest. However, in the context of measuring more general language constructs such as verbal proficiency or reading comprehension, test-takers from many different language groups may be involved. In such cases, different language versions of a test are often used to accommodate the multilingual assessment context.

There are several contemporary examples of multilingual assessment of verbal skills. For example, in many school districts in the USA, the reading comprehension of elementary school children is tested in their native language (see also Stansfield, this issue). In Newark, NJ, for example, reading tests are administered in English,

Address for correspondence: Stephen G. Sireci, School of Education, 156 Hills South, University of Massachusetts, Amherst, MA 01003–4140, USA; email: Sireci@acad.umass.edu
French, Portuguese, and Spanish (Sireci, 1989). The students’ performance on these tests is a critical factor in determining their promotion to the next grade level, and this test-based promotion standard should be equivalent across all language groups. International studies of educational achievement, such as the International Education Association’s multinational reading study (Elley, 1994), provide other examples of cross-lingual assessment. A third important area of cross-lingual verbal assessment is in postsecondary admissions testing, where multiple language versions of verbal proficiency tests are administered. For example, a Spanish-language version of the SAT (the Prueba de Aptitude Academica) is administered in Puerto Rico, and in Israel the entrance examination for college and university admission is administered in six different languages (we elaborate on this test later).

In cross-lingual assessment contexts where examinees are compared across languages, tests are often translated\(^1\) to facilitate the comparability of scores from the different language versions of the tests. In such cases, it is imperative that testing agencies closely examine the tests for any translation/adaptation problems. In particular, test-developers must establish that:

1) The construct measured exists in all groups of interest.
2) The construct is measured in the same manner in all groups.
3) Items that are thought to be equivalent across languages are linguistically and statistically equivalent; and
4) Similar scores across different language versions of the test reflect similar degrees of proficiency (van de Vijver and Tanzer, 1997; Sireci et al., in press).

Although it may be impossible to unequivocally accomplish (4) by following appropriate test adaptation guidelines (e.g., Hambleton, 1994; in press) and by performing qualitative and statistical analyses to address requirements (1)–(3), valid comparisons of proficiency can be made across examinees who take different language versions of a test (Sireci, 1997).

It should be noted that standards for fairness in testing support the use of qualitative and statistical methods for evaluating test translations (e.g., American Educational Research Association, American Psychological Association, & National Council on Measurement in

---

\(^1\) In cross-lingual assessment, the term ‘adaptation’ is considered preferable to ‘translation’ because the former term does not imply a literal word-to-word translation. The test adaptation process is typically more flexible, allowing for more complex word substitutions so that the intended meaning is retained across languages, even though the translation is not completely literal (Geisinger, 1994). In this article, we use the two terms interchangeably, for those readers who may be unfamiliar with what we mean by ‘adaptation’.
Appraising item equivalence

Education, 1999; Hambleton, in press). We focus on statistical methods in this article because they are often overlooked by many language test-developers, and because they often identify problematic items that escape a comprehensive judgmental review. The purposes of this article are to provide an overview of some statistical methods for assessing flawed items due to test translation and to provide an example of research in this area within the context of verbal proficiency testing.

II Statistical methods for evaluating differential item functioning

Before discussing techniques for assessing problematic items, we first differentiate three important, but distinct terms: item impact, differential item functioning, and item bias. Item impact refers to a significant group difference on an item, e.g., when one group has a higher proportion of examinees answering an item correctly than another group. Item impact may be due to true group differences in proficiency or due to item bias. Analyses of differential item functioning (DIF) attempt to sort out whether item impact is due to overall group differences in proficiency or due to item bias. To do this, examinees in different groups are matched on the proficiency being measured. Examinees of equal proficiency who belong to different groups should respond similarly to a given test item. If they do not, the item is said to function differently across groups. Analyses of item impact and DIF are statistical. Analyses of item bias, on the other hand, are essentially qualitative. An item is said to be biased against a certain group when examinees from that group perform more poorly on the item relative to examinees in the reference group who are of similar proficiency, and the reason for the lower performance is irrelevant to the construct the test is intended to measure. Therefore, for item bias to exist, a characteristic of the item that is unfair to one or more groups must be identified. Thus, statistical techniques for identifying item bias seek to identify items that function differentially across examinees who belong to different groups, but are of equal proficiency. Once these items are identified, they are subjected to qualitative review for bias. In the context of test adaptations, a potential cause of bias is improper translation.

There are several techniques that can be used to assess DIF. A detailed list of DIF methods is presented in Table 1. This table provides citations for each method and indicates the types of data for which each method is appropriate. Comprehensive discussions of DIF detection methods can be found in Holland and Wainer (1993),

There are several applications of DIF methodology to the problem of evaluating translated/adapted items (e.g., Angoff & Cook, 1988; Budgell et al., 1995; Allalouf et al., 1999; Sireci & Berberoglu, 2000). The most popular methods are the delta plot, standardization, Mantel–Haenszel methods, and methods based on item response theory (Sireci et al., in press). In this article, we describe two common methods that are computationally simpler than other methods, but have been particularly effective in identifying problems in item translations (Muñiz et al., 2001). The methods described here are the delta plot method (Angoff, 1982) and the Mantel–Haenszel method (Holland and Thayer, 1988). Both are designed for test items that are scored dichotomously (right/wrong), which is the most common item format used in large-scale language testing.
To identify items that may be biased against certain groups of test-takers, Angoff (1982) suggested producing a scatter plot of item delta values for each group. The delta value is an index of the difficulty of an item, where a higher value indicates a more difficult item. Deltas are transformations of the proportion of test-takers who answered an item correctly in a particular group (p-values). These proportions, which are on an ordinal scale, are transformed to the delta scale, which is an equal-interval scale. The delta index is a conversion from the p-value to a normal deviate, expressed on a scale with a mean of 13 and a standard deviation of 4 (i.e., $\Delta_i = 13 - (4 \times Z_{(i)})$). For example, an item with a p-value of .50 has a delta value of 13, and an item with a p-value of .84, has a delta value of 9 (i.e., 13−4).

The delta plot method proposed by Angoff and Ford (1973) for detecting DIF is used to compare the difficulty indices of the items in the two administrations (e.g., across two languages). In comparing item difficulties across cultures, the delta transformation is conducted separately for each group. According to the Angoff and Ford (1973) delta plot method, pairs of deltas – one pair for each item – are plotted as points on a graph with the two groups represented on the axes. The plot of these points will ordinarily appear in the form of an ellipse. Then, the principal axis line of the ellipse is drawn. This line minimizes the sum of the squared distances from the points to the line. The distance of each point from the line, relative to the distances of other points, serves as the evidence for DIF.

An example of a delta plot is presented in Figure 1. This figure presents the delta values computed from examinees who took either the French (vertical axis) or English (horizontal axis) version of an international certification exam (from Muñiz et al., 2001). The fact that the line representing equal item difficulties does not emanate from the origin reflects the overall difference in proficiency between the two groups, which is .77 (in favor of the English examinees) in this case. Parallel lines are drawn around this principal axis to indicate the effect size criterion for signifying DIF (point at which the difference in difficulty is thought to be meaningful). Items that fall outside these parallel lines are flagged for DIF. Those items that were flagged for DIF using the more sophisticated Mantel–Haenszel method (see below) are highlighted in the figure.

The delta plot method for evaluating cross-cultural DIF is relatively easy to do and is easy to interpret. However, it has been shown to miss items that function differently across groups if the discriminating power of the items were low. Furthermore, it may erroneously flag items for DIF if the discriminating power of the items were high
Figure 1 Plot of English \((n = 2000)\) and French \((n = 1333)\) group delta values (with mean difference .77 adjusted) 
*Source:* Muñiz et al., 2001; *Note:* DIF items are represented by black dots.

(Dorans and Holland, 1993). For these reasons, the delta plot method is suggested as a preliminary check before conducting further DIF analyses, or for use in those situations where sample sizes prohibit more sophisticated statistical analyses. Muñiz et al. (2001) demonstrated that the delta plot method was effective for identifying items that exhibited large DIF across language groups, even when the sample sizes were as small as 50 examinees per group. Delta plots have also been used effectively with large sample sizes to identify translation problems (Angoff and Modu, 1973; Cook, 1996).

2 Mantel–Haenszel method

The Mantel–Haenszel (MH) method for identifying DIF explicitly matches test-takers from two different groups on the proficiency of interest, and then compares the likelihood of success on the item for
the two groups across the score scale. The MH procedure is an extension of the chi-square test of independence to the situation where there are three levels of stratification (Mantel and Haenszel, 1959). In the context of DIF, these levels are:

- examinee group, e.g., two different language groups;
- matching variable: scores upon which examinees in different groups are matched; and
- item response: correct or incorrect.

For each level of the matching variable $j$, (typically total test score), a two-by-two table is formed that cross-tabulates examinee group by item performance. An example of such a table is presented in Table 2. The MH chi-square statistic is calculated as:

$$
\chi^2_{MH} = \frac{\left( \sum_j A_j - \sum_j E(A_j) \right)^2}{\sum_j \text{var}(A_j)}
$$

where the variance of $A_j$ ($\text{var}(A_j)$) equals:

$$
\text{var} A_j = \frac{M_{xj}M_{yj}M_{1j}M_{0j}}{T_j^2(T_j - 1)}
$$

The expected value of $A_j$ ($E(A_j)$) is calculated from the margins, as in a typical chi-square analysis. The MH chi-square can be computed using standard statistical software, such as SPSS, or using specialized DIF software such as the DICHODIF program (Rogers et al., 1993). Although external variables can be used to match the examinees in

---

**Table 2** Example of a Mantel–Haenszel contingency table for one level of matching variable

<table>
<thead>
<tr>
<th>Group</th>
<th>Correct (1)</th>
<th>Incorrect (0)</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Language X</td>
<td>$A_j$</td>
<td>$B_j$</td>
<td>$M_{xj}$</td>
</tr>
<tr>
<td>Language Y</td>
<td>$C_j$</td>
<td>$D_j$</td>
<td>$M_{yj}$</td>
</tr>
<tr>
<td>Total</td>
<td>$M_{1j}$</td>
<td>$M_{0j}$</td>
<td>$T_j$</td>
</tr>
</tbody>
</table>

**Notes:** $j$ denotes the specific score level at which examinees in the two groups are being compared; $A_j$ and $C_j$ represent the number of examinees at score level $j$ who answered the item correctly in language groups $X$ and $Y$, respectively, and $B_j$ and $D_j$ represent the number of examinees in each group (at $j$) who answered the item incorrectly. The margins (row and column totals) represent the total number of examinees in each group at score level $j$ (represented by $M_{xj}$ and $M_{yj}$), as well as the total number of examinees who answered the item correctly ($M_{1j}$) and the number who answered the item incorrectly ($M_{0j}$). Finally, $T_j$ represents the total number of examinees in the two groups at score level $j$. 
different groups (into j matched groups), the most common matching variable is total test score.

An attractive feature of the MH approach to detecting DIF is that, in addition to providing a statistical test, an effect size can also be computed. The MH ‘constant odds ratio’ can be used to provide an estimate of the magnitude of DIF. This ratio ($\alpha_{MH}$) is computed as

$$\alpha_{MH} = \frac{\sum_j \frac{A_jD_j}{T_j}}{\frac{C_jB_j}{T_j}}$$

This DIF effect size estimate ranges from zero to infinity with an expectation of 1 under the null hypothesis of no DIF (Dorans and Holland, 1993). However, it is usually rescaled onto the delta metric to make it more interpretable. This transformed effect size (MH D-DIF) is calculated as

$$MH\_D - DIF = -2.35\ln[\alpha_{MH}]$$

A MH D-DIF value of 1.0 is equivalent to a difference in proportion correct of about 10%. Rules of thumb exist for classifying these effect sizes into small, medium, and large DIF (Dorans and Holland, 1993). In a subsequent section, we describe a cross-lingual DIF analysis that used a MH D-DIF effect size criterion for flagging items that were not equivalent across translations.

3 Summary

A variety of DIF detection methods exist in the literature and previous research has shown that these methods can be used to identify translated items that are not statistically equivalent to their original versions. In the next section, we provide an illustration of a cross-lingual DIF study that applied the MH statistic to the problem of evaluating the translation equivalence of a subset of verbal test items. These illustrations come from Allalouf et al. (1999).

III Method

1 Instrument

The items analysed here come from the verbal subtest of the Psychometric Entrance Test (PET), which is a high-stakes test used for admissions to universities in Israel (for a complete description of the PET, see Beller, 1994). The PET contains multiple-choice items
measuring three proficiencies: verbal, quantitative, and English as a foreign language. It is developed in Hebrew and translated into Arabic, Russian, French, Spanish, and English. In this study, we report the results of DIF studies that compared the Russian translation of the verbal subtest with the original Hebrew version.

Item types in the verbal subtest of the PET include analogies, antonyms/synonyms, sentence completions, logic, and reading comprehension. The antonym and synonym items are considered non-translatable across languages and so they are not used to link the different language version of the PET to the Hebrew score scale. Instead, unique antonym and synonym items are constructed specifically for the translated form. To put the different language versions of the PET on a common scale, the translated items that are considered to be statistically equivalent are used as an anchor to equate the forms.

In this study, we used three separate test forms that were administered in both Hebrew and Russian to increase the number of items investigated and to provide a basis for comparison of results. A total of 125 verbal PET items were analysed. Table 3 presents the breakdown with respect to item type for these 125 items. Forty items were taken from Form 1, 44 items were taken from Form 2, and 41 items were taken from Form 3.

### 2 Examinees

The sample sizes for all analyses were large, ranging from 1485 examinees who took the Russian version of Form 3 to 7150 examinees who took the Hebrew version of Form 3. Some descriptive statistics for these examinees are presented in Table 4. The Hebrew examinees scored slightly higher than the Russian examinees on all three forms. The effect sizes associated with these differences were about half a standard deviation for Form 2, and about two-tenths of a standard deviation for the other two forms. The Hebrew and Russian examinees were more similar with respect to their performance on

<table>
<thead>
<tr>
<th>Item type</th>
<th>Number of items</th>
</tr>
</thead>
<tbody>
<tr>
<td>Analogies</td>
<td>26 (21%)</td>
</tr>
<tr>
<td>Sentence completion</td>
<td>33 (26%)</td>
</tr>
<tr>
<td>Logic</td>
<td>36 (29%)</td>
</tr>
<tr>
<td>Reading comprehension</td>
<td>30 (24%)</td>
</tr>
<tr>
<td>Total</td>
<td>125 (100%)</td>
</tr>
</tbody>
</table>
Table 4 Descriptive statistics for the Hebrew and Russian groups on the three test forms

<table>
<thead>
<tr>
<th>Examinees</th>
<th>n</th>
<th>Number of items</th>
<th>Verbal common items</th>
<th>Quantitativea</th>
<th>Mean</th>
<th>SD</th>
<th>Correlationb</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>Mean</td>
<td>SD</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Form 1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hebrew</td>
<td>6298</td>
<td>40</td>
<td>23.2</td>
<td>7.6</td>
<td>106.2</td>
<td>18.4</td>
<td>.72</td>
</tr>
<tr>
<td>Russian</td>
<td>1501</td>
<td>40</td>
<td>21.0</td>
<td>6.9</td>
<td>105.2</td>
<td>18.0</td>
<td>.70</td>
</tr>
<tr>
<td>Form 2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hebrew</td>
<td>5837</td>
<td>44</td>
<td>29.2</td>
<td>8.3</td>
<td>112.6</td>
<td>19.1</td>
<td>.68</td>
</tr>
<tr>
<td>Russian</td>
<td>2033</td>
<td>44</td>
<td>25.3</td>
<td>7.4</td>
<td>110.3</td>
<td>18.5</td>
<td>.65</td>
</tr>
<tr>
<td>Form 3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hebrew</td>
<td>7150</td>
<td>41</td>
<td>24.1</td>
<td>7.9</td>
<td>105.1</td>
<td>18.7</td>
<td>.70</td>
</tr>
<tr>
<td>Russian</td>
<td>1485</td>
<td>41</td>
<td>22.6</td>
<td>6.7</td>
<td>104.1</td>
<td>18.2</td>
<td>.66</td>
</tr>
</tbody>
</table>

Notes: a these are scaled scores (mean = 100, SD = 20); b between the verbal and the quantitative scores.

the quantitative subtest of the PET. Although the Hebrew group scored higher, the effect sizes associated with these differences were only about a tenth of a standard deviation. The correlations between the verbal and quantitative subtests were similar between the Hebrew and Russian samples. The correlations associated with the Russian samples were slightly smaller, which is probably due to the smaller test score standard deviations for these groups.

3 Dimensionality assessment

As pointed out by Zumbo (this issue), and others (e.g., Sireci, 1997), before performing DIF analyses in cross-lingual assessment, the viability of the matching criterion (in this case, total test score) must be defended by ruling out construct bias. We investigated the structural equivalence of the verbal subtest scores for each group by performing a weighted multidimensional scaling analysis on the common item data. The results of this analysis indicated that the same dimensions were needed to account for the relationships among the items in each group. Therefore, it was concluded that the total score derived for these common items represented the same construct in each language group (for the complete results of this dimensionality assessment, see Sireci et al., 1998).
4 DIF analyses

The DICHODIF computer program (Rogers et al., 1993) was used to perform all of the MH analyses reported here. This program provides an index of effect size, given in the delta metric, called MH D-DIF. Although chi-square tests were computed to test for statistically significant DIF, we focused on the MH D-DIF effect sizes. We use the DIF effect size classification rules developed at Educational Testing Service (Dorans and Holland, 1993) to classify items as ‘large’, ‘moderate’, or ‘no’ DIF. Items were labeled ‘large DIF’ if the absolute value of MH D-DIF was greater than or equal to 1.5 and the chi-square test was statistically significant at \( p < 0.05 \). Items were labeled as ‘moderate DIF’ if the absolute value MH D-DIF was at least 1.0 (and less than 1.5) and the chi-square test was statistically significant at \( p < 0.05 \). All other items were labeled ‘no DIF’.

All DIF analyses were conducted separately for each of the three test forms. We used a two-stage matching strategy, because if many items are initially flagged for DIF, the total test score may be inappropriate as a matching criterion. To refine our matching, in the first stage of analysis, we used the total raw score as the stratifying variable. In the second stage, we omitted those items flagged for large DIF (in the first stage) from computation of the total score. The items were classified into one of the three DIF categories based on the second stage results.

5 Detecting causes of DIF

In addition to identifying items that exhibit DIF across languages, we were interested in discovering reasons why such differences occurred. Two procedures were used for detecting possible causes of DIF. First, we investigated the direction of the DIF. This analysis identified whether the DIF was in favor of the Hebrew or Russian group of examinees. This analysis was conducted separately for each item type to address the question of whether the translation made the item easier or more difficult for any or all of the content areas. The second procedure for discovering potential causes of DIF used five Hebrew–Russian translators to analyse the type and content of the DIF items. The translators were unaware of the DIF classifications of the items. After an explanation of the term ‘DIF’, each translator independently completed an item review questionnaire for 60 items: the 42 DIF items, and 18 non-DIF items.

After the bilingual translators completed a review of the items, an 8-person committee was formed. The committee included the 5 translators and 3 Hebrew-speaking researchers. At the time of the committee meeting, the statistical information regarding DIF was given to
the translators. At the meeting, for every DIF item, each translator who surmised a cause for DIF defended his or her cause; the 18 non-DIF items served mainly for comparison purposes. The whole process was designed for the purpose of reaching agreements regarding the causes of DIF in the 42 DIF items, based on the ideas that the translators produced before they knew which items had DIF.

IV Results

1 Detection of DIF

The DIF analyzes identified items that displayed large and moderate DIF in each form. In Form 1, out of the 40 items, 7 displayed large DIF and 6 displayed moderate DIF (13 total DIF items or 33%). In Form 2, 11 of the 44 items displayed large DIF and 6 displayed moderate DIF (17 total DIF items or 39%). In Form 3, 8 of the 41 items displayed large DIF and 4 displayed moderate DIF (12 total DIF items or 29%). In total, 42 of the 125 items (33.6%) exhibited large or moderate DIF. Table 5 summarizes these results with respect to item type and magnitude of DIF.

In interpreting the findings, there was a consistent relationship between amount of DIF and item type. The greatest DIF was found on the analogy items and to a lesser extent on the sentence completion items. A low level of DIF was found for the logic and the reading comprehension items.

2 Direction of DIF

Analysis of the direction of the 42 DIF items revealed that examinees who took the Russian version performed better on 25 (59.5%) of these items; the Hebrew-speaking examinees performed better on 17

<table>
<thead>
<tr>
<th>Item type</th>
<th>Form 1</th>
<th>Form 2</th>
<th>Form 3</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Analogies</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Large</td>
<td>29%</td>
<td>29%</td>
<td>60%</td>
<td>78%</td>
</tr>
<tr>
<td>Moderate</td>
<td>27</td>
<td>0</td>
<td>15</td>
<td>81</td>
</tr>
<tr>
<td>Sentence completions</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Logic</td>
<td>0</td>
<td>0</td>
<td>8</td>
<td>9</td>
</tr>
<tr>
<td>Reading comprehension</td>
<td>20</td>
<td>18</td>
<td>20</td>
<td>14</td>
</tr>
<tr>
<td>Total</td>
<td>18</td>
<td>15</td>
<td>25</td>
<td>14</td>
</tr>
</tbody>
</table>

Table 5 Percentage of items of each item type showing large and moderate levels of DIF
of these items (40.5%). Table 6 presents an item type direction analysis of the 42 DIF items. As indicated in Table 6, 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 from the empirical findings.

3 Results of qualitative analyses

Based on the translators’ work, the committee found four main causes for DIF in the 42 DIF items. These causes are summarized in Table 7, which provides a tabulation of reasons why the items were hypothesized to function differently across the Hebrew and Russian versions, broken down by item type. The four hypotheses generated by the committee were:

### Table 6 Summary of DIF direction analysis (number of items)

<table>
<thead>
<tr>
<th>Item type</th>
<th>Better performance by Russian examinees</th>
<th>Better performance by Hebrew examinees</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(Number and %)</td>
<td>(Number and %)</td>
</tr>
<tr>
<td>Analogies</td>
<td>14 (82%)</td>
<td>3 (18%)</td>
</tr>
<tr>
<td>Sentence completions</td>
<td>8 (53%)</td>
<td>7 (47%)</td>
</tr>
<tr>
<td>Logic</td>
<td>0 (0%)</td>
<td>3 (100%)</td>
</tr>
<tr>
<td>Reading comprehension</td>
<td>3 (43%)</td>
<td>4 (57%)</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td>25 (60%)</td>
<td>17 (40%)</td>
</tr>
</tbody>
</table>

### Table 7 Causes of DIF

<table>
<thead>
<tr>
<th>Item type</th>
<th>Changes in</th>
<th>Cultural relevance</th>
<th>Unknown reason</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Word difficulty</td>
<td>Item content</td>
<td>Item format</td>
<td></td>
</tr>
<tr>
<td>Analogies</td>
<td>12</td>
<td>5</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>Sentence completions</td>
<td>4</td>
<td>3</td>
<td>5</td>
<td>1</td>
</tr>
<tr>
<td>Logic</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>3</td>
</tr>
<tr>
<td>Reading comprehension</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>2</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td>16</td>
<td>8</td>
<td>5</td>
<td>6</td>
</tr>
</tbody>
</table>
1) Changes in difficulty of words or sentences: The most common explanation for DIF was a change in word difficulty due to translation. Sixteen of the 42 DIF items (38.1%) were classified into this category. For these items, the committee felt the translation was accurate, but some words or sentences became easier or more difficult after translation. For example, an analogy item contained a very difficult word in the stem that was translated into a very trivial word. This problem can occur when a translator is unaware of the difficulty of the original word, or of the importance of preserving that difficulty.

2) Changes in content: The committee felt that the meaning of 8 items was changed in the translation, thus turning it into a different item. This problem may occur when an incorrect translation changes the meaning of the item, or when a word that has a single meaning in one language is translated into a word that has more than one meaning in the target language.

3) Changes in format: In 5 cases, changes in the format of the item were identified as the probable cause of DIF. For example, a sentence became much longer, or in a translated sentence completion item, words that originally appeared only in the stem were repeated 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: For 6 items, differences in the relevance of the content of the item to each culture were identified as the source of DIF. In these cases, the translations were deemed appropriate, but the content of the item interacted with culture. For example, the content of a reading comprehension passage could be more relevant to one group, relative to another, or the content of an item is simply more familiar to one group.

Analyses of the causes of DIF by item type revealed that in analogy items, the DIF was due to differences in word difficulty or changes in content. In the sentence completion items, all four reasons for DIF were cited for at least one of the 15 DIF items. For the reading comprehension items, the DIF was most often linked to cultural relevance. However, it is noteworthy that the translators could not explain the cause of DIF for 7 of the 42 items (16.7%), which included the three logic items.

V Discussion

Due to the information technology revolution and other factors, the world is quickly becoming a smaller place. Language barriers remain
one of the most significant obstacles to overcome in working with and understanding individuals from other countries and cultures. Activity in foreign language testing is on the rise and more and more language tests are being administered in multilingual contexts. The adaptation of these tests across languages is likely to be a common procedure. In such cases, the comparability of the adapted tests to the original is an important validity issue.

Analysis of differential item functioning can be helpful to language testers who must use different language versions of their assessments. To maintain parallel tests across languages – or to make comparisons across individuals who take different language versions of an exam – adapting items from one language to another may be necessary. By performing DIF analyses on these adapted items, test-developers can check whether the hypothesized equivalence of the item translations holds up to empirical scrutiny.

As the results of this study show, statistical and qualitative analysis of translated items should be used together to identify successful and unsuccessful item translations (Gierl et al., 1999). Qualitative analyses facilitate the important task of interpreting DIF findings in cross-lingual assessment. Discovering reasons why certain types of translated items display DIF helps us to better understand the construct(s) measured on each language version of a test. Thus, using both statistical and judgmental analyses of cross-lingual DIF allows us to defend or qualify construct interpretations of test scores, which enhances the construct validity of the inferences derived from such test scores.

The results of the empirical study described here illustrate how DIF analyses can identify problematic items and how the information gained from these studies can be used to inform future test development and test adaptation efforts. For example, based on their cross-lingual DIF analyses, Angoff and Cook (1988), Gafni and Canaan-Yehoshafat (1993), and Beller (1995) found that analogy items exhibited higher levels of DIF compared to the other item types studied here. These results are consistent with ours. Angoff and Cook (1988) hypothesized that translated verbal items with more context are more likely to retain their meaning. That is, it may be difficult to retain the exact meaning of a word, or pair of words in one language using one or two words in another language. However, if more context were provided, such as in a reading passage, there is more opportunity to use ‘decentering’ or other techniques to maintain the original meaning across the translation.

Although there are several statistical options for investigating cross-lingual DIF, we believe the MH procedure is a sensible choice for language test-developers. The MH DIF detection procedure offers
several advantages over other procedures. It is relatively easy to compute, it has modest sample size requirements, and it provides both a statistical test for DIF as well as the ability to compute an effect size. When the numbers of translated items are large, statistical significance becomes less important due to the likelihood of committing type II errors. What remains important is identifying items that are thought to be comparable across languages, but are not. Overlooking such item bias reduces the validity of the inferences we draw from the scores derived from the different language versions of an exam. Therefore, cross-lingual DIF analyses should be conducted whenever translated items are designed to be parallel across languages.

Acknowledgements

The authors are grateful for the helpful comments provided by Ronald K. Hambleton on an earlier version of this article.

VI References


—— 1995: Translated versions of Israel’s inter-university Psychometric Entrance Test (PET). In Oakland, T. and Hambleton, R.K., editors,


—— in press: Issues, designs, and technical guidelines for adapting tests in multiple languages and cultures. In Hambleton, R.K, Merenda, P. and


Shealy, R. and Stout, W. 1993: A model-based standardization differences and detects test bias/DTF as well as item bias/DIF. *Psychometrika* 58, 159–94.


