# Translating and Adapting a Test, yet another Source of Variance; the Standard Error of Translation

Yoav Cohen, Naomi Gafni & Pnina Hanani

*National Institute for Testing & Evaluation (NITE), Jerusalem, Israel*

A paper submitted to the annual meeting of the IAEA

Baku, Azerbaijan, September 2007

# A Comparison of Parallel Test Transadaptations

Test transadaptation (translation and adaptation) is the process whereby a test constructed in one language and culture is prepared for use in a second language and culture. Test transadaptation involves both the translation and adaptation of items written originally in the source language and the replacement of items unsuitable for translation/adaptation with items written in the target language. In the process, the transadaptation team effects a series of changes and modifications before the test attains its final transadapted form.

One of the International Test Commission (ITC) Guidelines for Test Translation and Adaptation is (Guideline D1., ITC, 2001; Hambleton, 2005): "Test developers/publishers should ensure that the adaptation process takes full account of linguistic and cultural differences in the intended populations." The rationale provided for this guideline is that "because a single translator cannot be expected to have all of the required qualities and brings a single perspective to the task of translation, in general, it seems clear that a team of specialists is needed to accomplish an accurate adaptation."

Two principal questions must be asked with regard to the product of test transadaptation produced by a team of experts: one is whether the transadapted product is of a high quality and the other is whether another team of experts would have done a better job. The first question can be answered by investigating the equivalence of the source and the transadapted test. One way to approach the second question is to examine the variance between transadaptations of the same test produced by independent teams. Such an investigation can provide us with an "estimate" for a standard error of transadaptation. The smaller this error, the more confidence we have in the transadaptation process and the final product.

The purpose of this study was to systematically investigate the variance between tests transadapted from the same source test by two independent teams.

## Method

### Instrument

The source test consisted of two Verbal Reasoning and two Quantitative Reasoning sections taken from a Hebrew version of the Psychometric Entrance Test (PET). PET is a scholastic aptitude test constructed and administered by the National Institute for Testing and Evaluation (NITE). Israeli universities use the PET for admissions purposes (for a detailed description of the PET, see Beller, 1994). The test battery consists of eight multiple-choice sections: two pilot sections and the following operational subtests (all of which are multiple-choice):

1. Verbal Reasoning (V) – Two sections consisting of 30 items each, focusing on the verbal skills and abilities needed for academic studies: the ability to analyze and understand complex written material, the ability to think systematically and logically, and the ability to perceive fine distinctions in meaning among words and concepts.
2. Quantitative Reasoning (Q) – Two sections consisting of 25 items each, focusing on the ability to use numbers and mathematical concepts (algebraic and geometrical), the ability to solve quantitative problems, and the ability to analyze information presented in the form of graphs, tables, and charts.
3. English as a Foreign Language (E) – Two sections consisting of 29 items each, designed to test proficiency in the English language (reading and understanding texts at an academic level).

Each subtest is scored separately, using a number-right scoring-rule formula, and is standardized on a scale that, for the original norm group (Hebrew-speaking examinees in 1984), had a mean of 100 and a standard deviation of 20.

PET is transadapted into the languages spoken by the majority of non-Hebrew-speaking university applicants – Arabic, Russian, French, Spanish, and English. Special attention has been given to Arabic, which is Israel's second official language (spoken by about 20% of the population), and to Russian, which is spoken by the largest immigrant group in Israel (about 10% of the population). Results from reliability, validity, and item- and test-bias studies comparing the transadapted versions with the Hebrew source version can be found in Beller, Gafni, and Hanani (2005).

**The Study's Design**

Two Verbal Reasoning sections (V1 and V2) and two Quantitative Reasoning sections (Q1 and Q2) were transadapted into Russian by two independent teams (RU-T1 and RU-T2) and into Arabic by another two independent teams (AR-T1 and AR-T2). This process produced a total of 16 transadapted sections (eight for each target language – four Quantitative Reasoning sections and four Verbal Reasoning sections; see Table 2 below). The 16 transadapted sections were randomly assigned as pilot sections to examinees who took PET in two different test administrations.

**The Transadaptation Process**

Each of the three sub-tests comprising PET is treated differently in the transadaptation process. The English sub-test of PET is identical for all language versions. In the Quantitative Reasoning subtest all the items are transadapted into the target language;

thus, the structure of the sections is not affected by the transadaptation. However, in the Verbal Reasoning subtest, the transadaptation process results in sections that are structured somewhat differently in each language. This process is now described.

The transadaptation process meets the ITC Test Adaptation Guidelines (ITC, 2001). Each transadaptation team consists of: a coordinator, a translator, three reviewers, a back-translator and partner for the back translation, and a person who does the final review along with the answer key review. All the team members possess substantial expertise and experience in the transadaptation of tests.

The transadaptation process consists of the following six stages:

***1. Selection of a test form and items suitable for transadaptation – or – pre-translation adaptation***

The tests administered to non-Hebrew-speaking examinees are transadaptations of previously administered Hebrew test forms. This ensures that the items selected for transadaptation are all of high psychometric quality.

The forms to be transadapted are selected according to the following considerations (Beller, Gafni, & Hanani, 2005):

- **Quality of calibration** – To reduce potential calibration problems, a form that was taken by Hebrew-speaking examinees who are relatively similar in distribution of ability to the target language examinees is identified.
- **Maintenance of technical term frequency in reading comprehension texts –** Reading comprehension texts with an abundance of technical terms are avoided when selecting test forms for translation, because in many cases these terms are self-explanatory in one language but not in another. In addition, the frequency with which such terms are used may vary in different cultural and linguistic contexts.

- **Cultural context –** The cultural context of the test must be familiar to the target group examinees.
- **Sensitivity reviews** – The tests undergo item sensitivity reviews to screen for items that might be provocative or offensive in their translated version.

***2. Transadaptation into the target language by professional test translators, all of them native speakers of the target language***

A qualified and experienced translator, who is proficient and knowledgeable in both languages and cultures, especially in the target language, translates the original Hebrew version of the test into the target language. Problems arising during the transadaptation process are discussed with the coordinator.

***3. Critical independent reviews of the transadaptation by three bilingual reviewers***

The transadapted versions are critically reviewed by three bilingual reviewers. The reviewers are required to first critique the transadapted version without consulting the original Hebrew and only afterwards to compare the transadapted version with the original Hebrew version. They are then required to pay special attention to the accuracy of the transadaptation as well as to the clarity of the sentences, the difficulty of the words, and the fluency of the text.

***4. Revision of the transadaptation by the translator and the coordinator***

The coordinator and the translator discuss the reviewers' comments, and revisions are made accordingly.

***5. Back translation***

A bilingual expert, who has seen neither the original Hebrew version nor the transadapted version, orally translates the transadapted version back into Hebrew.

The back-translation is simultaneously compared with the original Hebrew version, and items are revised where necessary.

***6. Final review***

The revised version of the transadaptation is given to a native speaker of the target language who has seen neither the original Hebrew version nor the previous versions of the translation. He or she is requested to answer the questions and to ascertain that there is one, and only one, correct answer to each question. The coordinator evaluates the final reviewer's answers, especially wrong answers that may derive from transadaptation inaccuracies.

## The Final Structure of the Transadapted Verbal Reasoning Sections

A Hebrew Verbal Reasoning section consists of the following item types: four vocabulary items, four letter substitution items, six analogies, five sentence completions, five logic items, and six reading comprehension items. The process described above resulted in somewhat different versions of the Verbal Reasoning sections in Russian and Arabic.

*Russian*

Vocabulary items: Four vocabulary items written and piloted in Russian expressly for this purpose.

Letter substitution items: Since items of this type exist in Semitic languages only, they were replaced with two analogies and two sentence completions taken from a Hebrew section in a different test form.

Analogies: Each section included eight analogies (six from the original Hebrew section and another two from a Hebrew section taken from a different test form). It should be noted that in both Russian and Arabic, analogies were the only

item type that posed problems for transadaptation. Certain analogies had to be removed and replaced with analogies from the item pool (i.e., items that had already been transadapted and administered, with sound item analysis statistics). One of the Verbal Reasoning sections used in this study (V1) contained three analogies that posed problems for transadaptation; in the other section (V2), one analogy was problematic for transadaptation. Both teams encountered the same problems, but devised somewhat different solutions. For V1, team RU-T1 replaced all three analogies with analogies from the item pool, while team RU-T2 transadapted all three of them. For V2, team RU-T1 transadapted the problematic analogy, while team RU-T2 replaced it.

Sentence completions: Seven sentence completion items (five from the original Hebrew section and two from a Hebrew section taken from a different test form) were transadapted into Russian.

All five logic items and all six reading comprehension items were transadapted into Russian from the Hebrew.

*Arabic*

In order to adapt the Hebrew Verbal Reasoning subtest for Arabic-speaking examinees, the sections are shortened from 30 to 26 items: one analogy, one sentence completion and two logic items are removed from each section. The items removed are the most difficult of the respective item type. In addition, difficult Hebrew items are replaced with easier items from the Arabic item pool.

Vocabulary items: Four vocabulary items were written and piloted in Arabic.

Letter substitution items: Four letter substitution items were written and piloted in Arabic.

Analogies: Each section included five analogies. In V1, three of the five original Hebrew analogies were too difficult and were therefore replaced with three easier analogies from the item pool. One of the two remaining analogies posed problems for both teams. Team AR-T1 replaced the analogy, while team AR-T2 transadapted it. In V2, one of the five analogies was too difficult, and was replaced with an easier one from the item pool, while the remaining four were transadapted by both teams.

All four sentence completions, all three logic items, and all six reading comprehension items were transadapted into Arabic from the Hebrew.

In both Russian and Arabic, the product of the transadaptation process consisted of the following three groups of items:

- Group 1 - items that were transadapted by both teams;
- Group 2 - identical items selected for the parallel transadapted sections (both original items written in the target language and items taken from the item pool before the section was submitted to the transadapting teams); and
- Group 3 - items that were selected independently by each team to replace those deemed unsuitable for transadaptation.

Table 1 presents the final structure of the Verbal Reasoning sections by language.

**Table 1**
**Number of Items in a Verbal Reasoning Chapter in Hebrew, Russian and Arabic by Item Type**

| Item Type | Hebrew | Russian | Arabic |
|---|---|---|---|
| **Words and Expressions** | **4** | **4** (not translated | **4** (not translated) |
| **Analogies** | **6** | **8** | **5** |
| **Letter Substitution** | **4** | **—** | **4** (not translated) |
| **Sentence Completions** | **5** | **7** | **4** |
| **Logic** | **5** | **5** | **3** |
| **Reading Comprehension** | **6** | **6** | **6** |
| **Total** | **30** | **30** | **26** |

Table 2 presents the transadapted sections, the number of transadapted items (Group 1 items) out of the total number of items in the section, and the number of examinees taking each section, for Russian and Arabic.

**Table 2**
**Transadapted Sections and Number of Transadapted Items (Group 1 Items)**

| Target Language (RU=Russian AR=Arabic) | Hebrew Source (V=Verbal Section Q=Quantitative Section) | Transadapted Sections (T=Team) | Number of Transadapted Items (Group 1 Items) out of Total Items |
|---|---|---|---|
| RU | V1 | RU-V1-T1<br>RU-V1-T2 | 23/30 |
| RU | V2 | RU-V2-T1<br>RU-V2-T2 | 25/30 |
| RU | Q1 | RU-Q1-T1<br>RU-Q1-T2 | 24/25 |
| RU | Q2 | RU-Q2-T1<br>RU-Q2-T2 | 23/25 |
| AR | V1 | AR-V1-T1<br>AR-V1-T2 | 14/26 |
| AR | V2 | AR-V2-T1<br>AR-V2-T2 | 17/26 |
| AR | Q1 | AR-Q1-T1<br>AR-Q1-T2 | 24/25 |
| AR | Q2 | AR-Q2-T1<br>AR-Q2-T2 | 23/25 |

To illustrate: The Hebrew Verbal Reasoning section V2 was transadapted into Arabic by two independent teams, one producing AR-V2-T1 and the other producing AR-V2-T2. Of the 26 items in this section, 17 were transadapted by the two teams and the nine remaining items were common to both sections (eight were originally written in Arabic and one was selected from the Arabic item pool).

In the Quantitative Reasoning sections, only items that contained text were transadapted and compared (24 and 23 items in Q1 and Q2 respectively).

**Criteria and Procedures Used to Assess the Variance between the Transadapted Sections**

The parallel transadapted test sections were compared, within languages, according to both quantitative and qualitative criteria. This paper will focus on the quantitative criteria which included:

1. A comparison of the reliability coefficients (KR-20) and SEM's for parallel sections (including both transadapted items and those originally written in the target language and used in both sections). This level of analysis refers to the product of the transadaptation process in its entirety.
2. An analysis of the different scores obtained on the two transadapted sections, using the GLM procedure, with the score on the respective operational sections (Verbal or Quantitative Reasoning) serving as the covariate. Since allocation of the parallel section was not entirely random, this procedure enabled us to control for initial differences in the measured ability between groups, and yielded adjusted means. This analysis was employed for Group 1 items only and did not include Group 2 or Group 3 items (Group 2 items were identical and common to the two parallel sections and would hence artificially reduce the transadaptation effect).
3. DIF detection by application of DIF analysis (Mantel-Haenszel procedure) to the pairs of transadapted sections (Rogers & Swaminathan, 1993; Holland & Thayer, 1988). The DICHODIF computer program (Rogers, Swaminathan & Hambleton, 1993) was used. DIF classification rules used in this study were based on the DIF classification rules of the Educational Testing Service (ETS) (Dorans & Holland, 1993). Two categories of DIF were defined: (1) Large – an absolute MH D-DIF value of at least 1.5; and (2) Moderate – an absolute

MH D-DIF value of at least 1.0.

## Results

### Reliability and SEM

Tables 3 and 4 present the raw mean score, standard deviation (SD), mean percent correct (P), reliability coefficient (KR-20), and standard error of measurement (SEM) for the transadapted sections, by sub-test, for the Russian and Arabic sections respectively. These statistics relate to whole sections.

**Table 3**
**Number of Examinees (N), Mean, SD, P, KR-20, and SEM of Sections Transadapted into Russian**

| Section | N | Mean | SD | P | KR- 20 | SEM |
|---|---|---|---|---|---|---|
| V1 | | | | | | |
| Russian T1 | 547 | 19.0 | 4.9 | 63 | .76 | 2.4 |
| Russian T2 | 520 | 18.1 | 4.8 | 60 | .75 | 2.4 |
| V2 | | | | | | |
| Russian T1 | 564 | 16.8 | 4.0 | 56 | .65 | 2.4 |
| Russian T2 | 774 | 16.6 | 4.3 | 55 | .69 | 2.4 |
| Q1 | | | | | | |
| Russian T1 | 287 | 13.7 | 4.6 | 55 | .78 | 2.2 |
| Russian T2 | 255 | 13.3 | 4.5 | 53 | .77 | 2.2 |
| Q2 | | | | | | |
| Russian T1 | 217 | 13.7 | 5.3 | 55 | .83 | 2.2 |
| Russian T2 | 441 | 13.8 | 5.1 | 55 | .82 | 2.2 |

**Table 4**
**Number of Examinees (N), Mean, SD, P, KR-20, and SEM of Sections Transadapted into Arabic**

| Section | N | Mean | SD | P | KR- 20 | SEM |
|---|---|---|---|---|---|---|
| | | | | | | |
| V1 | | | | | | |
| Arabic T1 | 1043 | 14.2 | 4.2 | 54 | .71 | 2.3 |
| Arabic T2 | 1105 | 14.2 | 4.3 | 55 | .73 | 2.2 |
| V2 | | | | | | |
| Arabic T1 | 1014 | 13.9 | 4.4 | 53 | .74 | 2.2 |
| Arabic T2 | 1028 | 12.6 | 4.3 | 48 | .72 | 2.3 |
| Q1 | | | | | | |
| Arabic T1 | 307 | 9.7 | 4.1 | 39 | .72 | 2.2 |
| Arabic T2 | 268 | 10.1 | 4.3 | 41 | .74 | 2.2 |
| Q2 | | | | | | |
| Arabic T1 | 283 | 9.6 | 4.7 | 38 | .78 | 2.2 |
| Arabic T2 | 797 | 8.7 | 4.5 | 35 | .77 | 2.2 |

In general, it was found that the reliabilities of the transadapted pairs of sections as well as the SEMs were similar to each other, indicating that there was no strong effect of transadaptation team, both in Russian and in Arabic.

**Means and SD's of Parallel Sections**

Tables 5 and 6 present the unadjusted and adjusted mean scores on each of the parallel transadapted forms, for Russian and Arabic respectively. The statistics refer only to Group 1 items. Group 2 and Group 3 items are not relevant to this analysis. The tables also show the Verbal Reasoning score used as covariate and the transadaptation effect size (TES). TES was defined as the difference between the two adjusted means divided by the mean SD of the two transadapted sections.

**Table 5**
**Number of Examinees (N), Unadjusted Mean Score, Adjusted Mean Score, Standard Deviation (SD), and Transadaptation Effect Size (TES) for the Two Parallel Transadaptations into Russian**

| | T1 | | | T2 | | | |
|---|---|---|---|---|---|---|---|
| **Variable** | **N** | **Mean** | **SD** | **N** | **Mean** | **SD** | **TES** |
| *Russian Verbal* | | | | | | | |
| Unadjusted Mean **V1** | 546 | 14.5 | 4.0 | 519 | 13.8 | 4.0 | |
| Adjusted Mean **V1** | | **14.2** | | | **14.0** | | 0.05 |
| *Verbal Reasoning (Covariate)* | | *102.2* | *16.6* | | *99.9* | *15.3* | |
| Unadjusted Mean **V2** | 561 | 12.8 | 3.5 | 767 | 12.6 | 3.8 | |
| Adjusted Mean **V2** | | **12.8** | | | **12.6** | | 0.05 |
| *Verbal Reasoning (Covariate)* | | *100.3* | *15.3* | | *99.8* | *16.3* | |
| *Russian Quantitative* | | | | | | | |
| Unadjusted Mean **Q1** | 286 | 13.4 | 4.4 | 255 | 13.1 | 4.4 | |
| Adjusted Mean **Q1** | | **13.2** | | | **13.3** | | 0.02 |
| *Quantitative Reasoning (Covariate)* | | *102.2* | *17.5* | | *100.1* | *17.0* | |
| Unadjusted Mean **Q2** | 217 | 12.3 | 5.0 | 440 | 12.4 | 4.8 | |
| Adjusted Mean **Q2** | | **12.3** | | | **12.4** | | 0.02 |
| *Quantitative Reasoning (Covariate)* | | *102.4* | *18.1* | | *102.1* | *18.2* | |

**Table 6**
**Number of Examinees (N), Unadjusted Mean Score, Adjusted Mean Score, Standard Deviation (SD), and Transadaptation Effect Size (TES) for the Two Parallel Transadaptations into Arabic (an asterisk indicates a significant difference between the scores of the two groups taking the two transadapted versions)**

| | T1 | | | T2 | | | |
|---|---|---|---|---|---|---|---|
| **Variable** | **N** | **Mean** | **SD** | **N** | **Mean** | **SD** | **TES** |
| *Arabic Verbal* | | | | | | | |
| Unadjusted Mean **V1** | 1038 | 6.7 | 2.8 | 1099 | 6.4 | 2.5 | |
| Adjusted Mean **V1** | | **6.6** | | | **6.5** | | 0.04 |
| *Verbal Reasoning (Covariate)* | | *92.4* | *18.3* | | *90.9* | *18.3* | |
| Unadjusted Mean **V2** | 1010 | 8.0 | 3.1 | 1028 | 7.1 | 3.0 | |
| Adjusted Mean **V2** | | **7.8** | | | **7.3*** | | 0.16 |
| *Verbal Reasoning (Covariate)* | | *93.3* | *18.3* | | *90.5* | *17.8* | |
| *Arabic Quantitative* | | | | | | | |
| Unadjusted Mean **Q1** | 306 | 9.4 | 4.0 | 265 | 10.0 | 4.2 | |
| Adjusted Mean **Q1** | | **9.5** | | | **10.0*** | | 0.05 |
| *Quantitative Reasoning (Covariate)* | | *90.8* | *14.0* | | *91.2* | *14.5* | |
| Unadjusted Mean **Q2** | 283 | 8.5 | 4.2 | 793 | 7.7 | 4.0 | |
| Adjusted Mean **Q2** | | **8.2** | | | **7.9*** | | 0.07 |
| *Quantitative Reasoning (Covariate)* | | *90.4* | *15.4* | | *88.5* | *14.5* | |

With the exception of one case (a verbal section in Arabic), the TES estimates were fairly small; they were larger for verbal sections than for quantitative ones, and smaller in Russian than in Arabic. TES was, on average, across target languages and domains, about 6% of the observed score SD. The difference between the two languages was unexpected and might be explained by the larger variability of dialects in Arabic.

**DIF**

DIF analysis was conducted on each pair of transadapted sections. Table 7 presents the percentage of items showing large DIF values (MH, $\Delta \geq 1.5$) and the percentage of items showing moderate DIF values (MH,$1.5 > \Delta \geq 1.0$).

**Table 7**
**Percentage of Items with Large and Moderate DIF Values**

| | **Large DIF** | | **Moderate DIF** | |
|---|---|---|---|---|
| | **V1, V2** | **Q1, Q2** | **V1, V2** | **Q1, Q2** |
| **Russian** | 9%, 8% | 0%, 0% | 13%, 8% | 4%, 4% |
| **Arabic** | 7%, 18% | 0%, 0% | 14%, 14% | 0%, 13% |

It was found that the verbal sections contained more items with large DIF than the quantitative sections. Also, somewhat more items with DIF were found in Arabic than in Russian. It should be noted that each item constitutes about 4% of the transadapted items. Considering that the transadapted items were derived from identical Hebrew items, no DIF was expected. However, as the table shows, at least two items in the Verbal Reasoning sections had large DIF, in both the Russian and Arabic transadapted sections.

## Discussion

The process of transadapting psychological and educational tests embraces two objectives which are potentially conflicting. The first is to obtain maximal accuracy in translation while ensuring that the difficulty of the transadapted item approximates the difficulty in the source language as closely as possible. The second is to have the transadapted items read as fluently and naturally as they do in the source language. The extent to which those two objectives are achieved depends on several factors:

1. the purpose and use of the test (e.g., whether the test is used to make high-stakes decisions)
2. the content of the test material (e.g., quantitative reasoning vs. verbal reasoning)
3. the expertise and experience of those involved in the process (translators as well as reviewers)
4. the strength of the linguistic relationship between the source and the target language
5. the extent to which the two populations taking the test in the source and target language are alike with respect to the assessed ability.

The findings of this study support the conclusion that in high-stakes tests transadapted by teams of experts, the amount of variance that can be attributed to the transadaptation process is about 6% of the standard deviation of the scores. It is larger for more verbally loaded tests than for tests that are less verbally loaded. Moreover, transadaptation variance is not necessarily smaller when the target and source languages are closely related (as in the case of Hebrew and Arabic.

## References


Beller, M. (1994). Psychometric and social issues in admissions to Israeli universities. *Educational Measurement: Issues and Practice, 13(2)*, 12-20.

Beller, M., & Gafni, N. (1995). Translated scholastic aptitude tests. In G. Ben-Shakhar & A. Lieblich (Eds.), *Studies in psychology* (pp. 202-219). Jerusalem: The Hebrew University, The Magnes Press.

Beller, M., Gafni, N., & Hanani, P. (2005). Constructing, adapting, and validating admissions tests in multiple languages: the Israeli case. In: Hambleton, R. K., Merenda, P. F., and Spielberger, C. D. (Eds.). *Adapting educational and psychological tests for cross-cultural assessment* (pp. 3-38). Lawrence Erlbaum Associates, Publishers.

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, New Jersey: Lawrence Erlbaum Associates. 35-66.

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

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 129 - 145

Rogers, S. J., & Swaminathan H. (1993). A Comparison of Logistic Regression and Mantel-Haenszel Procedures for Detecting Differential Item Functioning *Applied Psychological Measurement,* 17, 105-116.