Constructing, Adapting, and Validating Admissions Tests in Multiple Languages

Michal Beller

The Open University of Israel*

Naomi Gafni and Pnina Hanani

The National Institute for Testing and Evaluation (NITE)

June 1999

An invited paper presented at the *International Conference on Adapting Tests for Use in Multiple Languages and Cultures*, sponsored by the College Board, Educational Testing Service, and the International Test Commission, May 20-22, 1999, Georgetown University, Washington, DC.

* Department of Psychology and Education
16 Klausner St., Tel-Aviv 61392 ISRAEL
E-mail: michalb@oumail.openu.ac.il

Constructing, Adapting, and Validating Admissions Tests in Multiple Languages

Abstract

This paper focuses on some of the major problems involved in test translation from the perspective of test usage and score interpretation. In particular, it deals with: (a) the extent to which the source language version is translated, adapted or changed; (b) the definition of criteria for evaluating the quality of the translation; and (c) various approaches to calibrating scores on different language versions. These issues are demonstrated using the various language versions of the Psychometric Entrance Test (PET) - a test required by all major institutions of higher learning in Israel for making admissions decisions. Results related to differential guessing patterns, DIF analyses, reliability, construct and predictive validity, and test-bias are presented with respect to the Hebrew and the various translated versions (Arabic and Russian in particular).

Introduction

Translating psychological and educational tests into multiple languages is necessary for cross-cultural comparisons of traits and constructs among members of different cultures. In the past it was believed that culture free measures, such as figural reasoning tests, could be used in cross-cultural assessments (Cattell, 1940). But after many years of experience it has become clear that there is no such thing as a "culture-free" test or task (e.g., Frijda & Jahoda, 1966; Poortinga & Van de Vijver, 1991). Rather, there is a continuum extending from more to less so "culturally specific" tests (Jensen, 1980). Jensen introduced the concept of the degree of "cultural reducedness" of a test as the "cultural distance" over which a test maintains substantially the same psychometric properties of reliability, validity, item-total score correlation, and rank order of item difficulties. Since cultural distance is multi-dimensional, the properties of a particular test may not span the given cultural distance at all levels. For example, a verbal test may span the cultural distance in terms of language, if accurately translated, but it may not span this distance at the conceptual level (due to different connotations in different cultural contexts).

The problem of cross-cultural testing depends on whether the purpose is to determine predictive or construct validity. Demonstrating useful cross-cultural validity for a particular educational or occupational criterion is invariably much easier than establishing construct validity across widely different cultures (Jensen, 1980).

As cross-cultural comparisons are becoming increasingly popular and important, researchers are giving increased attention to the problem of bias and its detection. Van de Vijver and Poortinga (1997) claim that there are three forms of bias: (a) construct bias; (b) method bias; and (c) item bias. They argue that an integrated treatment of all forms of bias is needed to enhance the validity of cross-cultural comparisons. Other specialists and cross-cultural researchers have also stressed the need to ensure that instruments translated or adapted across languages are measuring the same construct (e.g., Geisinger, 1994; Hambleton, 1993, 1994; Van der Vijver & Tanzer, 1998).

The variety of goals for which translated tests are used affects the translation process and the role played by each of the languages involved. For each use a separate discussion of the considerations regarding the appropriateness of the translated versions is needed**.** A familiar use of test translation is applying an already

well-established standard scale, such as IQ tests or personality questionnaires, for research and practical purposes. Such tests (e.g., K-ABC) are used mainly for development of local norms required for making individual decisions, as well as cross-national comparisons for research purposes. In this case the translation process involves only the necessary minimal test adaptation (Poortinga, 1995).

International assessments of education (e.g., TIMMS) are examples of cross-national research in which no particular language or content is considered to be dominant. Rather, all participating countries determine the contents of the assessment together, thereby assuring maximal common ground. The agreed-upon version is translated into all relevant languages. In this case the main purpose of the assessment is cross-national, although in some cases (e.g., Canada) comparisons have been carried out within the country between various sub-groups.

Another goal of test translation is to establish a fair and valid selection procedure for candidates from various language groups who are applying to institutions of higher learning within a specific country and language of instruction. This is typical of countries that serve as targets for large-scale immigration (e.g., USA, Canada, Australia, Israel). In this case, the translation process may seriously affect the validity and fairness of individual high-stakes decisions. Using scores from admissions tests that are administered to all groups in the source language results in confounding of the measured construct and the level of familiarity with the source language. Therefore, there is a need to find ways to reduce the confounding of the two variables, for example, by translating admissions tests into the various languages of the applicants, and measuring mastery of the local language separately.

The goals of translation must be dealt with in the context of the target populations. The status of the target population for translation varies: in some cases, a country may have more than one official language (e.g., Switzerland, Canada, Israel), and, therefore, tests are routinely translated. Even in such cases, countries differ in that some provide their different language groups with a full educational system in their own language (e.g., Switzerland), while others provide only a semi-separate educational system (e.g., Canada and Israel). In Israel, for example, Arabic is a second official language, and the Arabic-speaking population has its own K-12 separate educational system (with the exception of some subjects that are taught in

Hebrew). The lingua franca of higher education, on the other hand, is Hebrew and Hebrew alone; and there is one system for both populations.

Additional target populations for test-translations are immigrants from different countries. Immigrants from a particular country should not be treated automatically as if they were a homogeneous group. They differ in their familiarity with their mother tongue and with the new local language, and in their acquaintance with the local culture and educational system (depending on their age at the time of immigration, the length of time they have spent in the new country, and their level of immersion in the culture of the new country). Therefore, translating a test into a specific language does not assure a valid comparison even among individuals within a given language group. The varied degrees of familiarity of each population with both the source and the target language influence the way in which the test is translated. In some cases, it might be advisable to translate only specific terms in the test rather than translating the test fully. For example, veteran Russian immigrants to Israel (who have already studied in Israel for several years) may prefer to take the admissions exam for higher education in Hebrew with a glossary containing specific terms translated into Russian, rather than in their natural language.

This paper focuses on some of the major problems involved in test translation from the perspective of test usage and score interpretation. In particular, it will deal with the extent to which the source language version needs to be translated, adapted or changed; the definition of criteria for evaluating the quality of the translation; and approaches to calibrating scores on different language versions. Methods for dealing with these issues will be discussed and demonstrated using the various language versions of the Psychometric Entrance Test (PET). PET is a scholastic aptitude test constructed and administered by the National Institute for Testing and Evaluation (NITE). It is used, in conjunction with a matriculation certificate, by all Israeli universities and by other major institutions of higher learning for making admissions decisions. The matriculation certificate is based on both school assessment and external nation wide achievement tests[1]. PET measures various cognitive and

[1] For students of foreign origin, the school-based component is either missing or, more often, cannot be compared to the Israeli matriculation scores. Therefore, these candidates are rank-ordered on the basis of their PET score alone. In some universities, admissions decisions are based on a composite score comprised of the PET score and the mean score achieved in preparatory courses which are required of non-Hebrew-speaking candidates before they are admitted to the university.

scholastic abilities, in an attempt to estimate future success in academic studies. It consists of three multiple-choice sub-tests (see Appendix A): Verbal Reasoning (V), Quantitative Reasoning (Q), and English as a second language (E). No correction for guessing is used for scoring the test and examinees are encouraged to guess when they do not know the correct answer. For a detailed description of PET, see Beller (1994).

In establishing admissions policy for the universities in Israel, policy makers and psychometricians have been faced with the problem of finding the best methods for predicting the academic success of non-Hebrew-speaking applicants (along with Hebrew-speakers) in institutions of higher education (where the language of instruction is Hebrew). In other words, the goal is to rank *all* different language speaking examiness on *a* common scale, that best predicts *a* common criterion, within the *same* cultural context. It was decided to administer PET in the language with which the applicant is most familiar, because it was believed that this procedure provides each applicant with the opportunity to perform optimally.

Results from reliability, validity, and item and test-bias studies comparing the translated versions with the Hebrew source version will be presented below. Special attention was given to Arabic, which is Israel's second official language (spoken by about 15% of the population), and to Russian, which is spoken by the largest immigrant population in Israel (about 10% of the population).

Translated Scholastic Aptitude Tests: The Israeli Case

PET is currently translated into the languages spoken by the majority of non-Hebrew-speaking university applicants: Arabic, Russian, French, Spanish, and English[2]. The translation process is an ongoing endeavor: four, two, two, one and one new forms are annually translated into Arabic, Russian, English. French and Spanish, respectively (out of 10-18 new forms in Hebrew).

---

[2] The English version is actually a combined English and Hebrew version in which all of the questions are presented both in English and in Hebrew. It is offered to applicants whose native language is English, as well as to applicants who are not proficient in any of the languages mentioned above. A short dictionary appears at the bottom of each page, which contains a translation of selected key words into the languages most required by these examinees, according to their native language. Currently, these languages are: French, Spanish, German, Hungarian, Rumanian, Italian, Russian, and Amharic.

To familiarize examinees with PET, and to ensure that all persons fully understand the requirements of each type of task involved in the test, NITE publishes an information booklet that includes previously administered tests as well as explanations. This booklet is also translated into the five languages mentioned above. This procedure is particularly important, because the various language groups differ in terms of their previous experience with multiple choice tests. Of the 66,731 examinees to whom PET was administered in 1998, approximately 27% chose to take PET in one of these languages (15% in Arabic, 10% in Russian, and 2% in the other foreign languages). Examinees who take PET in a foreign language are required by some institutions to take an additional Hebrew Proficiency Test (HP), which is scored separately.

The non-Hebrew versions of PET are essentially translations of Hebrew test versions administered to Hebrew-speaking examinees; they thus have a similar structure. The English sub-test of PET is identical for all language versions. The Quantitative Reasoning sub-test is translated from Hebrew. The rationale behind this is that, in general, translated math items are directly comparable to the source. The Verbal Reasoning sub-test is translated only in part. Most of the items are selected from the pool of Hebrew items, but others are specially constructed for the various language versions (e.g., the Word and Expression items).

In the following sections we will discuss both procedural and substantive issues involved in the translation of PET. Special attention will be devoted to the translation of the Verbal Reasoning sub-test.

Selecting a test form for translation: considerations

The tests in the five languages (Arabic, Russian, French, Spanish, and English) are translated from previously-administered Hebrew test forms. This ensures that the items selected for translation are all high-quality items in psychometric terms. The following considerations are taken into account in selecting the Hebrew versions to be translated:

1. *Quality of calibration*: To reduce potential calibration problems, we try to identify a form previously taken by Hebrew-speaking examinees who are relatively similar in distribution of ability to the "other language" examinees (target group).

2. *Reliability*: It has been found that the reliability of the total score on the translated versions is almost as high as that of the original Hebrew version (see Table 2 below). In the past, the reliability of the Verbal Reasoning sub-test for the Arabic-speaking population was lower than the reliability of the other translated sub-tests (see Table 2 for current values). This was caused mainly because this sub-test was extremely difficult for Arabic-speaking examinees. To increase the reliability of the Verbal Reasoning sub-test for this population, an easier test was constructed by selecting the easier half of the items from a Hebrew form (to be used later for calibration) and supplementing it with easy items from the item-bank.
3. *Preservation of frequency of technical terms in reading comprehension texts*: Reading-comprehension texts with an abundance of technical terms (scientific terms, legal language, or psychological jargon) are avoided when selecting texts 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 is often different in different cultural and linguistic contexts; some terms might exist in one language but not in others. A text abounding with "foreign"[3] words will not be translated into Arabic, because Arabic-speaking examinees generally do not encounter such words in elementary school and high school.
4. *Cultural context*: The cultural context of the test must be familiar to all examinees. A reading comprehension text that includes local cultural connotations will not be selected for translation.
5. *Sensitivity reviews*: The tests undergo item sensitivity reviews to avoid choosing items that might be provocative or offensive in their translated version. For instance, an item including the word "uprising" (INTIFADA) in Arabic would not be used because of the political sensitivity of this word. Texts that deal with politics, religion, sex, etc would also not be chosen.

The translation process

There are four stages in the translation process:

[3] In this context, "foreign" refers to words from languages other than Hebrew that are, nonetheless, used by the Hebrew-speaking population (e.g., "technologia").

1. *Initial translation*: 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 translation process are discussed with the psychometrician in charge of the entire translation process. Following a recent recommendation by Hambleton (personal communication, 1997), two independent Russian translations are being carried out instead of only one. Experience so far with this more costly procedure indicates that it improves the quality of the review that occurs in the next stage. The cost-effectiveness of this additional procedure is in the process of being evaluated.
2. *Independent reviews*: The translated versions undergo critical reviewing by several bilingual reviewers, some with a solid background in mathematics and logic, and others who are highly competent in verbal reasoning. Both American and British reviewers read the English version, and reviewers from various Spanish-speaking countries read the Spanish version. The reviewers are required to first critique the translated version without looking at the original Hebrew, and only afterwards to compare the translated version with the original Hebrew version. They are then required to pay special attention to the accuracy of the translation as well as to the clarity of the sentences, the difficulty level of the words and the fluency of the text. Each reviewer solves the test items, checking that no changes have resulted in the item's inner logic, that each item still has one, and only one, correct answer, and that the distractors are adequate in terms of their attractiveness. The psychometrician and the translator discuss the reviewers' comments and suggestions, and revisions are made accordingly.
3. *Back translation*: A bilingual expert, who has not previously seen the original Hebrew version, orally translates the translated version back into Hebrew. This stage is carried out orally mainly because it allows for immediate interaction and discussion between the back translator and the psychometrician. The back translation is simultaneously compared with the original Hebrew version, and translated items are revised where necessary.
4. *Final check before initial administration*: The revised version of the translation 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 solve the questions without looking at the original Hebrew, and to ascertain that there is one, and only one, correct answer to each question. The psychometrician evaluates the answers, searching for wrong answers that may derive from translation inaccuracies.

Specific problems in translating the Verbal Reasoning sub-test

The Verbal Reasoning sections are the most problematic to translate because words and concepts in one language do not always have the same meanings, connotations, familiarity or level of difficulty when translated into another language. Idioms and expressions are a typical source of difficulty, as they very often cannot be translated at all. Languages differ in the richness of their semantic fields. For example, Hebrew has a wide range of words relating to agriculture. In English you pick grapes, you pick olives, etc, but in Hebrew there is a different verb for picking grapes, for picking olives, etc. Similarly, there are different words in Hebrew for washing the floor, washing dishes, and washing clothes. English-speakers use the same verb "wash" for all of these activities. English and Hebrew have only one word for camel. In Arabic there are a vast number of words, denoting the different types of camels according to their characteristics.

Translating into Arabic poses enormous problems. On the one hand, written Arabic is the same for all Arabs in all Arab countries. But spoken Arabic, which is very different from written Arabic, varies from country to country, and even from area to area in the same country. A "coat" in written Arabic is "MIATAF" and in spoken Arabic it is "KABBUT". A "hat" in written Arabic is "QUBBA" and in spoken Arabic it is "TAQIUA". In the Arabic translation, an effort is made to avoid words in spoken Arabic, and to use the written words instead, even though they are more difficult. Arabic-speaking examinees that read Arabic literature might encounter these words, but others might be unfamiliar with them.

The following sections will discuss specific translation problems, ranging from item types that cannot be translated at all to those that can be translated directly or that require only slight adaptation.

Letter Exchange Items

Letter exchange items are based on word roots – a feature specific to Semitic languages (see Appendix A). As a result, this item type does not appear in the Russian, French, Spanish and English versions. Since Arabic is a Semitic language, this item type can be used; however, the items cannot be translated and must be written in Arabic. New items of this type are pre-tested before they are used for scoring purposes.

Words and Expressions

Word and expression items cannot be translated from Hebrew, and are written directly in the target language. In the Arabic- language test version it was found necessary and cost-effective to pretest these items before using them for scoring purposes.

Analogies

This item type is the most difficult to translate, as it involves meanings and connotations of single words and the relationships between pairs of words. There are few words that have a precisely equivalent meaning, connotation and level of difficulty in another language. In translating analogies, the relationship between the two words in each pair must be retained as accurately as possible, while at the same time keeping in mind the difficulty level of the vocabulary. The original analogy is often designed to test command of Hebrew vocabulary in addition to analytical ability. In such cases, the translated item is often easier.

Analogies sometimes involve culture-dependent knowledge:

tile : floor –
bead : necklace
grain : crumb
lintel : door
ink : inkstand

Americans generally do not have tiled floors; thus American examinees would not be aware of the relationship between the two words. The above analogy was therefore not translated into English.

Sentence Completion

Sentence Completion items are also difficult to translate. In order to produce a natural sounding, smoothly flowing sentence in the target language, it is often necessary to change the structure of the sentence, and this affects the way the missing words are inserted into the translated sentence. Furthermore, the translator has to ensure that all four distractors produce sentences that are grammatically and syntactically correct, so that choosing the correct answer will depend solely upon internal logic and not upon structural and grammatical "hints." In addition, it is necessary to preserve, as much as possible, the level of the language (everyday, formal, literary etc.), the complexity of the missing words, the number of blanks, etc.

Problems arise, for instance, in Arabic, where every noun has a grammatical gender that is not necessarily the same as its Hebrew counterpart. Arabic also has two plural forms: the plural for more than two items (plural) and the plural for two items (dual), with the verb conjugated accordingly. Moreover, sentences in Arabic usually begin with a verb, unlike Hebrew sentences, which usually begin with a noun. All of these problems call for many alterations in the item's structure and complicate the task of translating the sentence completion items.

When the structure of the sentence in the target language changes, the sentence might contain only three blanks instead of the four blanks in the original Hebrew version. There is no a priori reason not to use such items, but if in the calibration analysis the item is found to be considerably easier than its Hebrew counterpart, it will later be removed from the anchor for calibration.

Logic

Logic items must be translated very carefully and accurately. The translator must try to preserve all of the logical elements of the Hebrew item while adhering to the same structure as existed in the original item. Attention must be paid to whether the context is real or imaginary, and names and measurement units (kilometers versus miles, etc.) must be adjusted, so that the terms used will be equally familiar to all examinees.

In an attempt to preserve the precise structure of the original logic items in the translated version (for example, preserving negatives, double negatives, conjunctions such as "only," "also" etc.), it is sometimes necessary to change the structure so that

the syntax of the target language will be correct. For example, the syntax of the Hebrew structure *"all p's are not q"* is ambiguous in English. Thus, a statement in Hebrew such as: *All birds of prey are not green* (meaning that there is not even one bird of prey that is green) cannot be directly translated into English. In order to preserve the exact Hebrew meaning the structure of the English sentence has to be changed as follows: *No birds of prey are green.* A similar difficulty arises when this sentence has to be translated into French. In French, "all" cannot be followed by a negative. Therefore, the translation has to be: *Aucun oiseau predat n'est vert.*

Another example is: *There are no Japanese-made cars that are not both large and fast.* In Arabic, words that mean "only" and "also" are usually placed at the end of a sentence, and it is not always clear what they are referring to. Therefore, in translating the above sentence into Arabic some information has to be added: *"...which are not large and fast at the same time."*

Reading-Comprehension

In the translation of a text, the emphasis is on the following: translating accuracy; preserving the fluency, richness and style of the language using concepts which are familiar in the target language; and being consistent in the use of the terms appearing in the text.

One of the criticisms of cross-cultural testing is that a translated text cannot convey the same meaning and preserve the same level of difficulty as the original text. Therefore, at NITE, a special team was recently established for the purpose of finding texts for Arabic speakers, the largest group of non-Hebrew-speaking examinees. The texts are written in Arabic, and adapted to NITE's test requirements. One of the questions that will have to be answered is whether scores that are obtained from a test that uses comprehension texts written originally in Arabic are comparable to scores derived from test versions that use texts originally written in Hebrew.

Scoring the language versions

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

score is a weighted sum of the scores on the three sub-tests (2V, 2Q, and E), with a mean of 500 and a standard deviation of 100. For a more detailed description of PET, see Beller (1994).

The same parameters are applied for scoring the English and the Quantitative sub-tests in all language versions (assuming that translation does not alter the meaning of the quantitative questions). A calibration procedure similar to the one described by Angoff and Modu (1973) is used in scoring the Verbal Reasoning sub-test. An anchor is established between the Hebrew version and each of the other language versions. This is done by selecting items that have similar psychometric indices and a similar order of difficulty (using delta-plot techniques) for the two groups of examinees. Once an anchor is established, linear equating methods (Tucker or Levin) are applied.

Table 1 presents the means and standard deviations for the various language versions of PET and its sub-tests for the academic year 1997/8. The French-, Spanish-, and English-speaking groups are very small; therefore, these results cannot be generalized beyond this specific context. The Arabic- and Russian-speaking groups are large and stable enough across the years to be representative of the two largest minority groups applying to institutions of higher education in Israel. Consequently, most of the following analysis and discussion will be based on the Hebrew, Arabic, and Russian versions. As mentioned earlier, the Arabic-speaking population in Israel has a separate educational system, whose language of instruction is Arabic. In general, this educational system is less developed than the Hebrew educational system. The disparity between the Hebrew- and the Arabic-speaking populations is already evident in the early grades, as has been found in several national assessments of the educational system (e.g., Aviram, Kfir, & Ben-Simon, 1996; Ben-Simon & Hamza, 1996). The somewhat higher level of performance of Arabic-speaking examinees in mathematics, relative to verbal achievement, is also evident at an early age.

Table 1

Means and Standard Deviations for the Various Language Versions of PET and Its Sub-tests in 1997/8

		Total Score		Verbal Reasoning		Quantitative Reasoning		English	
Language	**N**	**M**	**SD**	**M**	**SD**	**M**	**SD**	**M**	**SD**
Hebrew	48,897	554	101	108	20	111	19	111	23
Arabic	9,949	431	85	86	16	92	19	82	16
Russian	6,366	512	101	92	18	106	18	95	22
French	511	521	84	99	16	104	19	112	17
Spanish	363	480	82	90	14	96	17	108	22
English	645	552	106	100	21	107	21	131	23

It is interesting to note that the greatest difference in performance between the Arabic- and the Russian-speaking groups and the Hebrew group is on the non-translated English sub-test.

Quality of the translated versions

In addition to the meticulous process of translation described above, the following quantitative criteria are also used to assess the quality of the translated versions: differential effect of guessing, item analysis, differential item functioning (DIF), reliability, construct equivalence, validity, and bias in prediction of criterion scores. These criteria are affected by translation as well as various other inseparable cultural group-distribution factors.

Differential effect of guessing

A study conducted by Gafni and Melamed (1990) investigated the following phenomenon: despite being instructed to guess when they did not know the correct answer, only 75% to 93% of the examinees (depending on the specific sub-tests) responded to all the items on PET. It was postulated that different language groups might manifest different guessing behaviors. For example, it was expected that the

English-speaking group would be more familiar with multiple-choice tests and would, therefore, be more likely to follow the test instructions closely. On the other hand, the Russian-speaking group, being less acquainted with this type of test, might be less inclined to guess. It was also hypothesized that the degree of familiarity of the general public with multiple-choice testing might have an effect.

The results suggested that people with differing cultural backgrounds differ in their tendency to guess. A language-group effect and a familiarity effect were found. In 1984, Russian-, Arabic- and French-speaking examinees tended to omit more items than Hebrew-, English- and Spanish-speaking examinees; in 1987 (after four years of PET administration), Russian-speaking examinees tended to omit more items than all other groups. The proportion of omitted items has dropped significantly for all groups, as the test has become more familiar and test preparation more prevalent. The 1990 study recommended that the importance of test instruction be emphasized, in particular among members of groups with a greater tendency to avoid guessing.

Item analysis and DIF

The quality of each translated item is examined (in terms of its level of difficulty and degree of discrimination). In addition, the differential item functioning (DIF) of each translated item is examined, comparing Hebrew- and non-Hebrew-speaking examinees (DIF refers to the simple observation that an item displays different statistical properties for different groups, after controlling for differences in the abilities of the groups). If the statistical properties of certain translated items are poor, those items are reviewed, and possible (*post hoc*) reasons for their failure are raised. A decision is made regarding each item whether to include the item in the scoring of the translated version, and if so, whether to include it in the anchor used for calibration.

Gafni and Canaan-Yehoshafat (1993) examined DIF on the Verbal Reasoning sub-tests of three Russian forms of PET using a delta-plot technique proposed by Angoff (1972). The greatest DIF was found for analogies, and the smallest DIF for the logic and sentence completion items. These results are similar to those found by Angoff and Cook (1988) for English- and Spanish-speaking examinees taking the SAT, with the exception of the logic items, which are not included in the SAT verbal

sections. The reading comprehension items showed relatively greater DIF than in the Angoff and Cook study.

Allalouf, Hambleton, and Sireci (1999) investigated the relationship of DIF (using the Mantel-Haenszel method) to item type and hypothesized sources of DIF. They analyzed three forms of the Hebrew and Russian versions of PET. The results reflect the extent of the problems involved in translating the different verbal item types, as described in the previous section. It was found that 42 out of 125 items (34%) functioned differentially across languages. The analogies were the most problematic, with 65% of them exhibiting DIF. On most of these items, the Russian-speaking examinees performed better than the Hebrew speaking examinees. A large proportion (45%) of the sentence completion items also exhibited DIF, but in this case, neither group performed better than the other did. A panel of translators was asked to speculate on possible causes of DIF for each item. The main causes suggested were changes in word difficulty, differences in cultural relevance and changes in content.

In a follow-up study that is presently in progress (Allalouf, 1999), a panel of translators is revising the above-mentioned 42 DIF items, taking into account the specific reasons that were presumed to have caused the DIF. It is hoped that the revision will reduce the amount of DIF found for these items. Results are not yet available.

Reliability

The internal reliability of each sub-test, as well as that of the total score, is routinely estimated for each language version. Table 2 presents the median internal consistency coefficients (KR-20) for the three sub-tests and for the total score of the various language versions of PET. These reliabilities are relatively high, both for the Hebrew version and for the other language versions. The somewhat lower reliabilities in the foreign language versions may be explained by translation-related problems. However, internal reliability is not determined solely by the quality of the test items and the quality of the translation, but also by the true variance within the group of examinees. From experience gained at NITE, it appears that in many cases the quality of the translation is confounded with differences in performance. When two groups differ in ability, this in and of itself might create differences in reliability,

comparability, and item-DIF. When items are too difficult for a certain group, the reliability of the test for that group is relatively low. For example, the median reliability for the Verbal Reasoning sub-test of the first five Arabic forms, constructed between 1984-1989, was 0.68 (Beller & Gafni, 1995). To raise reliability, a Verbal Reasoning test, which included much easier items, was specially constructed for the Arab version. The median reliability obtained for 23 forms constructed in this way increased to 0.84. While the reliability of this new sub-test is higher, it probably introduces a larger equating error than that of the previous sub-test.

Table 2

Median Reliability Coefficients (KR-20) of PET Sub-tests and of the Composite Total Score for Each Language Version Administered During 1990 – 1997 (the number of test forms appears in parentheses)

Language	V	Q	E	PET
Hebrew (65)	0.90	0.90	0.94	0.96
Arabic (23)	0.84	0.86	0.82	0.93
Russian (18)	0.86	0.88	0.92	0.94
French (9)	0.82	0.87	0.90	0.93
Spanish (9)	0.82	0.87	0.93	0.94
English (15)	0.90	0.90	0.96	0.96

As important as these results are, reliability is only a necessary, not a sufficient, condition for test development. It is the validity of various translated versions that provides the most important justification for using the scores obtained on them.

Construct equivalence

To ensure that the translated versions of PET are measuring the same construct as the original Hebrew version, Allalouf, Bastari, Hambleton, & Sireci (1997) used exploratory factor analysis, multidimensional scaling and confirmatory factor analysis to evaluate the structural equivalence of the Verbal Reasoning sub-test in two Hebrew and Russian versions. Specifically, they analyzed four of the five content areas: analogies, logic, reading comprehension, and sentence completion. A total of 41 items were included in the analysis. In the analyses performed on the two versions,

the structure of PET was found to be similar across the two language versions for the subset of items.

Validity

At NITE, the validity of the selection procedure is routinely tested by examining the predictive validity of PET against the criterion of GPA at the end of the first year of university studies and at the completion of undergraduate studies.

Predictive validity of the Arabic vs. the Hebrew versions

Validity coefficients have been recently computed for each of the six research universities in Israel. The predictors in this validity study were PET (total score as well as the three sub-test scores), the average score on the matriculation certificate obtained at the end of high school (Bagrut), and the composite score (Adm). The composite score is used in the admissions process and is typically based on equal weighting of the PET and Bagrut scores. The criterion was the first year GPA from the respective universities for those students who began their studies between 1987 and 1996. The analysis was carried out for each department within each cohort, provided that it included at least five students from each language group. Results are reported for 538 departments that met the above condition. The number of students, means, and standard deviations of the various predictors and the criterion are presented in Table 3.

Table 3

Means and Standard Deviations (in parentheses) of the Predictor and Criterion Scores (for Hebrew and Arabic)

Language	PET	V	Q	E	Bagrut	Adm	GPA
Hebrew	581	115	114	114	90	50	79
N=56,009	(61)	(13)	(13)	(15)	(7.2)	(7.2)	(8.8)
Arabic	505	101	104	94	91	46	69
N=9,071	(46)	(12)	(12)	(11)	(5.9)	(5.3)	(10.1)
Both	571	113	112	111	90	49	78
N=65,080	(63)	(14)	(14)	(16)	(7.2)	(7.1)	(9.6)

The validity coefficients (Pearson correlations) for the two language groups are presented in Table 4. In general, the validities of the admissions score and all the other predictor scores were higher for the Hebrew-speaking examinees than for the Arabic-speaking examinees. These results can be explained to some extent by the restricted variance that characterizes the predictor scores of the Arabic-speaking examinees. The validities of the Q and V sub-tests for Hebrew-speaking examinees were fairly similar, while the validity of the E sub-test was lower. A different pattern was found for Arabic-speaking examinees, with a low validity for V as well.

Table 4

Predictive Validities (correlations corrected for range restriction) of PET, Bagrut and the Admissions Score For the Criterion (First-year GPA) for Arabic- and Hebrew- Speaking Examinees (observed correlations appear in parentheses)

Language	PET	V	Q	E	Bagrut	Adm
Hebrew	.34 (.28)	.25 (.23)	.29 (.26)	.36 (.13)	.36 (.31)	.45 (.37)
Arabic	.26 (.19)	.12 (.10)	.23 (.19)	.13 (.09)	.31 (.24)	.39 (.28)
Both	.40 (.34)	.30 (.27)	.31 (.28)	.25 (.23)	.26 (.23)	.45 (.36)

Although within each language group the validity of Bagrut was higher than that of PET, the opposite pattern was obtained in the combined sample. This is due to the fact that unlike PET, the Bagrut scores are obtained from two separate educational systems - the Hebrew and the Arabic - and are not calibrated. Therefore, it is not surprising that they do not reflect the difference between the group means on the criterion (see Table 3). It is worth noting that the validity of the combined score for both groups and across groups was higher than each of the component scores taken separately.

The differences in the validity coefficients of the two groups (e.g., on the V sub-test) cannot be attributed solely to translation effects, due to the large differences between the two language groups on PET as well as on the criterion.

Predictive validity of the Russian vs. the Hebrew versions

The predictive validity of the translated Russian version is less affected by factors such as the large differences in ability between the Arabic- and Hebrew-speaking groups. Therefore, research regarding the Russian translation will be discussed more extensively than research on the Arabic version.

In a study conducted recently by Gafni and Bronner (1998), the predictive validity of the PET score was calculated for the Russian-speaking group and compared with that of the Hebrew-speaking group. The predictors in this study were PET, its three sub-tests (V, Q, and E), and the admissions score (Adm). The admissions score in this study was based on equal weights of PET and an achievement score obtained either in high school (Bagrut) or in a preparatory program for applicants who did not study in an Israeli high school[4]. This study included an additional predictor – the score on a Hebrew Proficiency test (HP)[5] administered to all non-Hebrew-speaking examinees. Validity coefficients for the various predictors for students who began their university studies between 1992-1996 were computed for two criteria: first-year GPA (FGPA), and third-year GPA (TGPA). The analyses were conducted for each department within each cohort, provided that it included at least five students from each language group. Results are reported across 463 departments that met the above condition for FGPA, and 83 departments that met this condition for TGPA. Table 5 presents the number of students, the mean, and the standard deviation of the various predictors and criteria.

[4] The achievement score was not available separately.

[5] The Hebrew Proficiency Test (HP) is comprised of multiple-choice items (67%) and an essay (33%). It is scored separately with a mean of 100 and a standard deviation of 20.

Table 5

Means and Standard Deviations (in parentheses) of the Predictor and Criterion Scores (Hebrew and Russian) for FGPA and TGPA

Language	PET	V	Q	E	Adm	HP	FGPA	TGPA
FGPA								
Hebrew	600	118	117	118	101		80	
N=55,434	(60)	(13)	(13)	(16)	(6.8)		(8.8)	
Russian	561	111	116	100	99	93	73	
N=7,313	(52)	(12)	(12)	(16)	(6.2)	(16)	(11.9)	
TGPA[6]								
Hebrew	590	116	116	116	101		82	84
N=6,612	(57)	(12)	(13)	(15)	(6.2)		(6.7)	(6.9)
Russian	540	108	112	96	98	87	76	81
N=1,011	(54)	(13)	(12)	(16)	(6.1)	(14.4)	(8.4)	(8.3)

The largest difference between the two groups was found for E, with Hebrew-speaking examinees exhibiting better performance; no difference was found for Q. The difference for V was somewhere between the two. A slight difference in favor of the Hebrew-speaking examinees was found for Adm, implying a reverse pattern of differences on the achievement score (which is not calibrated) compared with PET. The difference on the FGPA criterion was similar to that on PET. It is interesting to note that the difference on TGPA decreased compared with that on FGPA.

The validity coefficients for the two language groups are presented in Table 6, averaged across all departments. The observed correlations (in parentheses) are corrected for range restriction. The average validity coefficients of both the admissions score and PET for the FGPA group were similar for both the Russian and

[6] The samples for TGPA are much smaller than those for FGPA. This is partially due to attrition, but is mainly due to the fact that most students had not yet completed their third year of study at the time the study was conducted.

Hebrew groups across all fields of study. However, the pattern of validity coefficients for PET sub-tests was different for the two language groups. Whereas Q had the highest validity for the Hebrew-speaking group, the most valid test for Russian-speaking examinees was E. V had the lowest validity for the Russian-speaking group, and this may indicate that V does not measure exactly the same construct in both languages, either due to the translation and adaptation of the test, or because of the specific test content, which was originally chosen for the Hebrew-speaking group. Apparently, numerous factors determine the validity of a test within a group, and the quality of the translation is only one of them.

Table 6

Predictive Validities (correlations corrected for range restriction [7]) of PET, Admissions Score (Adm) and the Hebrew Proficiency Test (HP) for GPA at the End of the Freshman Year (FGPA) and Senior Year (TGPA) for Russian - and Hebrew- Speaking Examinees (raw correlations appear in parentheses)

Language	PET	V	Q	E	Adm	HP
FGPA						
Hebrew	.39	.32	.36	.24	.48	
	(.26)	(.21)	(.26)	(.12)	(.37)	
Russian	.35	.26	.30	.29	*	*
	(.27)	(.16)	(.21)	(.24)	(.38)	(.23)
TGPA						
Hebrew	.44	.36	.37	.29	.54	
	(.20)	(.16)	(.17)	(.12)	(.28)	
Russian	.45	.35	.38	.35	*	*
	(.26)	(.17)	(.19)	(.22)	(.33)	(.20)

- The variances of the unrestricted populations were not available

[7] The corrected correlations are estimates based on a similar set of data where the population variances were provided for the non-restricted sample.

The relatively high validity of E for the Russian-speaking group might be attributed to moderating variables not investigated in this study. For example, those students who immigrated to Israel from a large city with a good educational system might have had a better opportunity to learn English than immigrants who came from some remote town without a well-developed, modern educational system. It is also possible that some of the Russian-speaking examinees immigrated to Israel several years before taking PET and had the opportunity to study within the Israeli educational system. This could be reflected both in their English score and in their criterion score.

Test bias

NITE has conducted research to detect whether there is test bias for the Arabic-speaking (Bronner, Allalouf, & Oren, 1996) and the Russian-speaking examinees (Gafni & Bronner, 1998). The term "bias" refers to systematic errors in the predictive validity or construct validity associated with an examinee's group membership. The methods that follow from the definitions given in Darlington (1971) and the discussion by Linn (1984) were used to detect bias in the various predictors (for a brief description of the method see Appendix B). Results regarding single predictors should be viewed with caution, due to the effect of excluding a predictor from a regression equation on which there are pre-existing group differences (Linn & Werts, 1971).

Bias in testing Arabic-speaking examinees

The sample consisted of 9,999 Hebrew-speakers and 1,624 Arabic-speakers who began their studies in 1992 and 1993, and whose PET and Bagrut scores as well as their first-year GPA were available. Six Israeli universities were included in the study, each with several faculties consisting of several departments, resulting in a total of 99 departments (Bronner, Allalouf, & Oren, 1996). Table 7 presents the number of significant cases of bias detected across all departments. In examining the predictors separately, hardly any bias was detected for PET as a single predictor (bias against the Arabic-speaking group was detected in only 3% of the departments). Similar results were found for V and E, taken separately. For the Q sub-test, bias was found in 8% of the cases – 3% against and 5% favoring the Arabic-speaking group. On the other

hand, the Bagrut was biased in favor of Arabic-speaking examinees in 60% of the departments. This latter result is consistent with the previous observation that the score scales of the Bagrut for the two groups are not calibrated. Therefore, it is not surprising that a clear indication of bias favoring the Arabic-speaking examinees was detected for the admissions score (Adm.) in 14% of the departments.

Table 7

Relative Frequency (%) of Significant Cases of Test Bias for Arabic-Speaking Examinees with FGPA as the Criterion

Predictor	Bias Against Arabic-Speakers	Bias Favoring Arabic-speakers
Adm.	2	14
Bagrut	3	60
PET	3	0
V	3	0
Q	3	5
E	4	0

Bias in testing Russian-speaking examinees

The first sample consisted of 55,434 Hebrew-speaking examinees and 7,313 Russian-speaking examinees who began their studies in one of the years 1992 – 1996, and whose PET scores and FGPA scores were available. Six Israeli universities were included, for a total of 463 departments. The admissions score was only available for a sub-sample consisting of 26,875 Hebrew-speakers and 3,478 Russian-speakers. For this predictor, the analysis was conducted on only 259 departments (Gafni & Bronner, 1998). Table 8 presents the number of significant cases of bias detected across all departments, with PET and the admissions score (Adm) as predictors. Hardly any clear bias was found for PET as a single predictor: in 3% of the 463 cases, there was a clear indication of bias against the Russian-speaking examinees, and in 2% of the cases, the bias was in their favor. Similar results were found for V and Q (the

translated sub-tests), with a tendency to overpredict FGPA for the Russian-speaking group. A reverse tendency was found on the English sub-test. For the admissions score, in about 10% of the 259 departments a clear indication of bias was detected, mostly in favor of the Russian speaking examinees.

Table 8

Relative Frequency (%) of Significant Cases of Test Bias for Russian-Speaking Examinees, with FGPA and TGPA as the Criteria

Predictor	Bias Against Russian-Speakers	Bias Favoring Russian-speakers
FGPA		
PET	3.0	2.0
V	0.5	1.5
Q	1.0	7.0
E	1.5	0.0
Adm.	1.0	9.0
TGPA		
PET	8.0	5.5
V	2.7	5.5
Q	2.7	5.5
E	11.0	0
Adm.	2.7	0

The above results do not provide any clear cut answers as to the possible prediction bias resulting from translation: on the one hand, E, which is not translated at all, exhibited bias against the Russian-speaking examinees, but on the other hand, Q, which is also relatively unaffected by translation, exhibited bias in favor of this group. The verbal reasoning sub-test, which is most affected by translation, produced results that do not indicate evident translation problems. It may be concluded that, if the translation process succeeds in more or less preserving the same difficulty level of

the two language versions, and if the meaning of what is measured is as similar as possible, then no bias should be expected as a result of the translation per-se.

One of the main criticisms of conventional bias studies is that they often overpredict the criterion scores of minority groups. The reason for this is that the transition to a college in which the student body is predominantly the majority population is initially more demanding for the minority students than for the majority students. Therefore, one might expect the overprediction to disappear in the third year of college. To test this assumption, a sub-sample consisting of 6,612 Hebrew-speaking examinees and 1,011 Russian-speaking examinees whose scores on PET and third-year GPA (TGPA) were available was examined. A total of 83 departments were included in this study. The Adm score was only available for a sub-sample consisting of 2,687 Hebrew- speakers and 338 Russian-speakers. For this predictor, the analysis was available for only 37 departments. Table 8 presents the number of departments in which clear bias was detected against and in favor of Russian-speaking examinees. In general, when TGPA served as the criterion, a pattern similar to FGPA emerged, with a slight decrease in the tendency of the predictors to be biased in favor of the Russian-speaking group, as expected.

Summary

The issue of test translation should be approached while keeping in mind various dimensions, such as the goal of the translation, the target population, and the type and content of the test. A test can be translated for research purposes, where mainly group differences are of interest, and it can be used for individual high-stakes decision-making purposes, such as admissions to universities. The requirements for quality translation are higher in the latter than in the former.

The translation process of PET, a test used for admissions to higher education in Israel, from Hebrew to Arabic, Russian, French, Spanish and English was described in detail, exemplifying inherent translation problems (in particular for the verbal sections). The quality of the translations was checked by applying various qualitative and quantitative methods, thus highlighting different aspects of the translatability of the various language versions. Several steps were taken to ensure the quality of the

translation: (1) investing substantial effort in the qualitative check of the translations, in some cases using two independent translators; (2) examining response patterns and differential tendency of examinees to guess on multiple-choice items; (3) examining item analysis and DIF; (4) checking reliability and its relation to the groups' ability level; (5) investigating predictive validity for two criteria (first- and third-year GPA); and (6) analyzing predictive test bias for the various sub-groups.

The extensive set of analyses presented in this paper regarding PET and its translated versions provides a great deal of information regarding the complexity of the issue of comparability and equivalence of translated admissions tests. It is argued that when assessing the quality of translated tests in a context of individual high-stakes decisions, a broader view of test fairness and equivalence of versions should be adopted. An examination of predictive validity and test bias should be carried out in addition to the more common DIF-like analyses. The issue of validity and that of test-bias are both necessary to establish the overall justification for using the translated test results for admissions (high-stakes) purposes. In the context of admissions tests, the criterion against which the test is validated is not of less importance than the test itself.

The results presented about PET demonstrate that when a proper translation process is applied, it can produce a set of translated tests that are construct-equivalent, reliable, and relatively valid and fair. However, even when all steps are taken to ensure the quality of translation and the comparability of scores, it is still not possible to assure that the original and translated versions are indeed fully equivalent. Yet the alternative of testing non-Hebrew speakers in Hebrew would seem to constitute a much less fair solution. Moreover, parameters related to cost-effectiveness and overall expected utility gains should be considered as well (e.g., weighing the costs associated with further improving the current process).

References

Allalouf, A. (1999). *Reducing DIF of translated verbal items by revising DIF items*. A paper to be presented in the ITC conference, Gratz.

Allalouf, A., Bastari, B., Hambleton, R. K., & Sireci, S. G. (1997). Comparing the dimensionality of a test administered in two languages. *Laboratory of Psychometric and Evaluative Research Report no. 319*. Amherst, MA: University of Massachusetts, School of Education.

Allalouf, A., Hambleton, R. ,& Sireci, S. G. (1999). Identifying the causes of DIF in Translated Verbal Items. *Journal of Educational Measurement* (in print).

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 scales of the Prueba de Aptitud Academica and the Scholastic Aptitude Test.* Research Report 3. New York: College Entrance Examination Board.

Aviram, T., Kfir, R., & Ben-Simon, A. (1996). *The Israeli National Assessment of Educational Progress: Achievements in Math for eighth graders*.

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: Ben-Shakhar, G. & Lieblich, A. (Eds.). *Studies in Psychology*, 202-219. The Magnes Press, The Hebrew University, Jerusalem.

Ben-Simon & Hamza (1996). *The National Assessment of Educational Progress*: *Achievements in language studies in the fourth grade*.

Bronner, S., Allalouf, A., & Oren, C. (1996). Fairness of using the Psychometric Entrance Test for selection of Arabic applicants to universities in Israel. Jerusalem, Israel: National Institute for Testing and Evaluation, Technical Report No. 223.

Cattell, R. B. (1940). A culture-free intelligence test: Part I. *Journal of Educational Psychology, 31*, 161-179.

Darlington, R. B. (1971). Another look at "cultural fairness". *Journal of Educational Measurement, 8*, 71-82.

Frijda, N. & Jahoda, G. (1966). On the scope and methods of cross-cultural research. *International Journal of Psychology, 1*, 109-127.

Gafni, N., & Bronner, S. (1998). *An examination of criterion-related bias for Hebrew- and Russian-speaking examinees in Israel.* Paper presented at the annual meeting of the American Educational Research Association. San-Diego.

Gafni, N., & Canaan-Yehoshafat, Z. (1993). *An examination of deferential item functioning for Hebrew and Russian-speaking examinees in Israel*. Paper presented at the 24th annual conference of the Israeli Psychological Association, Ramat-Gan.

Gafni, N. & Melamed, E. (1994). Differential tendencies to guess as a function of gender and lingual-cultural reference group. *Studies in Educational Evaluation, 20,* 309-319.

Geisinger, K. F. (1994). Cross-cultural normative assessment: translation and adaptation issues influencing the normative interpretation of assessment instruments. *Psychological Assessment, 6,* 304-312.

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

Hambleton, R. K. (1993). Translating achievement tests for use in cross-national studies. *European Journal of Psychological Assessment, 9,* 57-68.

Jensen, A. R., (1980). *Bias in mental testing*. London: Methuen; New-York: Free Press.

Linn, R. L. (1984). Selection Bias: multiple meanings. *Journal of Educational Measurement, 21*, 33-47.

Linn, R. L., & Werts, C. E. (1971). Considerations for studies of test bias. *Journal of Educational Measurement, 8(1)*, 1-4.

Poortinga, Y. H. (1995). Cultural bias in assessment: Historical and thematic issues. *European Journal of Psychological Assessment, 11 (3),* 140-146.

Poortinga, Y., H., & Van de Vijver, F. J. R. (1991). Culture-free measurement in the history of cross-cultural psychology. *Bulletin of the International Test Commission, 18*, 72-87.

Sireci, S. G., Bastari B., & Allalouf, A. (1998). *Evaluating construct equivalence across adapted tests.* Paper presented at the annual meeting of the American Psychological Association, San Francisco, CA.

Van de Vijver, F. J. R. & Poortinga, Y. H. (1997). Toward an integrated analysis of bias in cross-cultural assessment. *European Journal of Psychological Assessment, 13 (1),* 29-37.

Van de Vijver, F. J. R & Tanzer, N. K. (1998). Bias and equivalence in cross-cultural assessment: An overview. *European Review of Applied Psychology, 47,* 263-279.

Appendix A

Description of PET

All items are given in multiple-choice format.

1. **Verbal Reasoning (V):** this section consists of 60 items which focus 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 the meaning of words and concepts. The verbal section includes words and expressions, analogies, sentence completion, logic, reading comprehension and letter exchange items.

   Words and expressions: These items are intended to directly assess the vocabulary of examinees. They appear in a number of forms: items dealing directly with the meaning of words or expressions; sentence completion items with one blank; antonyms; and items in which examinees have to choose one option whose meaning is distinct from that of the other options.

   Analogies: these items assess the ability to define the relationship between two concepts (a lower-order mapping between A and B) and to recognize the similarity between two relationships (a higher-order mapping between A-B and C-D pairs).

   Sentence Completion: these items entail understanding of the logical and semantic relationships within a complex sentence. There are different types of connections between different parts of the sentence: one part could specify another part, explicate it, exemplify it, negate it, etc. In decoding the nature of the relationship between two or more parts of the sentence, one must pay particular attention to the prepositions in the sentence, because their meaning establishes the type of connection. After filling in the blanks, the entire sentence should constitute an argument that must be coherent in order for it to be the right answer. Thus, the ability to analyze and understand arguments is needed for solving these items.

   Reading-Comprehension: these items reflect the conception of the skilled reader as one who constructs meaning from a text, as opposed to simply decoding what is on the page. The items reflect the process that a reader goes through while deriving meaning from the text. They assess the test taker's ability to interpret,

synthesize, analyze, and evaluate the reading material, and thus measure higher order analytical and evaluative skills.

Letter-Exchange: these items are based upon a morphological feature of Semitic languages not shared by Indo-European languages, namely, the fact that most of the vocabulary in Hebrew – all verbs and most nouns and adjectives – can be characterized as a combination of Root + Pattern. The root is most typically composed of three consonants, and it carries the semantic core of the words formed by it; the patterns take the form of vocalic and syllabic additions to the root, and they modify the core meaning of the root. The Letter Exchange items are composed of four sentences. In each sentence one word is altered by changing its root letters into a standard template (the letters **p.t.l**). In three of the four sentences the standard template stands for the same three letters. In the remaining sentence the template replaces another root. The examinees have to identify this sentence.

2. **Quantitative Reasoning (Q)**: this section consists of 50 items which focus on the ability to use numbers and mathematical concepts (algebraic and geometrical) to solve quantitative problems, and the ability to analyze information presented in the form of graphs, tables, and charts. Solving problems in this area requires only basic knowledge of mathematics – the math level acquired in the ninth or tenth grades in most high schools in Israel. Formulas and explanations of mathematical terms that may be needed in the course of the exam appear in the test booklet.

3. **English as a Foreign Language (E):** this section consists of 54 items designed to test command of the English language (reading and understanding texts) at an academic level. The English sub-test contains three types of items: Sentence Completions, Restatements, and Reading Comprehension. This sub-test serves a dual purpose: it is a component of the PET total score, and it is also used for placement of students in remedial English classes.

Appendix B

Linn (1984) demonstrated that if group membership is coded so that there is a positive correlation between group and qualifications, then there should be no relationship between group membership and both X (the predictor) and Y (the criterion) for a given qualification level. In order to detect bias, Linn presented two boundary conditions on the regression coefficients that imply clear bias:

1. $\beta_{YC.X} < 0,$ or
2. $\beta_{XC.Y} < 0.$

where C is the language group category. The higher value of C is given to the group with the higher mean value for X. Linn's first boundary condition states that in the regression of Y on X, bias occurs when the group with the higher value for X has the lower regression line. Note that in this case, the average regression underpredicts the criterion for the group with the lower value for X, and is therefore indicative of bias against this group. Linn's second boundary condition states that in the regression of X on Y, bias occurs when the group with the higher value for X has the higher regression line. Note that in this case, the average regression overpredicts the criterion for the group with the lower value for X, and is therefore indicative of bias in favor of this group.

Authors Note

We wish to thank Maya Bar-Hillel, Gershon Ben-Shakhar, Ruth Fortus, Chava Kassel and Jerry Levinson for their insightful comments, and Shmuel Bronner for his assistance in the data analysis.