Evaluating the Effect of Ability Differences between Groups and the Use of a Non-Representative Anchor on Equating in Cross-Lingual Circumstances

Joel Rapp & Avi Allalouf

National Institute for Testing and Evaluation (NITE)

Abstrac

Cross-lingual equating, in which a translated test is equated to its original version, is usually conducted in the “separate monolingual group design.” This design is similar to the well known “common item non-equivalent group” equating design except that in this design source- and target-language versions of the test are administered separately to source- and target-language examinee groups, and a set of translated items, considered to be equivalent across languages, is used as an anchor.

According to the literature, equating in such a design may seriously be affected if there are considerable differences between the ability levels of the language groups being equated or if an unrepresentative anchor item set is used. However, this is the case in many cross-lingual equating circumstances. It is quite common to find ability differences between language groups and to use an anchor that does not represent the whole test properly since many items are non-translatable, or do not retain the same psychometric characteristics following translation. This is especially true for items in which the verbal aspect is critical.

The effect of ability differences and of the use of a non-representative anchor on equating was studied in a typical cross-lingual setting. Data from two versions of the Psychometric Entrance Test (PET) for admission to Israeli universities were used. The equating of the verbal domain subtest using similar vs. dissimilar examinee samples and representative vs. non-representative anchors was compared. In terms of examinee scores, differences found in both comparisons were about one fifth of a standard deviation. It is suggested that the effect on equating of these two factors alone has been overestimated in the literature. Some explanations and implications for cross-lingual equating are discussed.

In the last decade, there has been an increasing interest in cross-lingual assessment. This increase stems from the will to enhance fairness in assessment by allowing examinees to choose the language in which they will be tested and from the desire to facilitate comparative studies across countries, ethnic or cultural groups (Hambleton, 1993; Sireci, 1997). Cross-lingual assessment usually involves translating tests from a source language to a target language. In adapting a test from one language to another we need to develop a common scale for both language test versions. However, establishing a common metric presents a serious problem, as it is very difficult to ensure that the different language test versions measure the same construct - a very important underlying assumption of any co-scaling method (Cook, 2000; Wainer, 1999). Nevertheless, in practice, linking between different language versions is done in many situations where it is necessary to compare the skills and abilities of examinees speaking different languages (Angoff & Modu, 1973; Angoff &

Cook, 1988; Beller, Gafni & Hanani, 1999; Hulin & Mayer, 1986; Schmitt & Dorans, 1999; Woodcock & Munoz-Sandoval, 1993).

A popular design used for cross-lingual linking is the "separate monolingual group design". In this design, source- and target-language versions of a test are separately administered to source- and target-language examinee groups. Items considered to be equivalent across the two language test versions are used to link the tests onto a common scale (see Sireci, 1997, for a detailed description of the design). In fact, this equating design is similar to the familiar "common-item non-equivalent groups" design, except that the translated items or a selected part of them, are used as the common items. One serious problem with employing the separate monolingual group design is the assumption that items that are translations of one another are equivalent. This assumption is quite weak, since the common (translated) items are poor anchor items. However, it is important to note that even in ideal conditions, where items would remain fully equivalent following translation, there would still be other important factors that might affect equating.

According to the literature that deals with practical issues in equating, there are a few factors that may induce equating error and potentially distort the estimated equating relationship in the common-item non-equivalent groups design. Among these factors are (1) ability differences between the groups of examinees taking the alternate test forms; (2) an anchor item set that does not represent properly the linked tests, either in content or in difficulty; and (3) the tests to be equated are not built to the same content or statistical specification (Petersen, Marco & Stewart, 1982; Marco, Petersen & Stewart, 1983; Petersen, Cook & Stocking, 1983; Budescu, 1985; Cook & Petersen, 1987; Kolen & Brennan, 1995).

This study deals with the effect of these factors on equating in typical cross-lingual circumstances. Since cross-lingual equating is usually implemented using the separate monolingual group design, these factors may be problematic in it as well. Moreover, in cross-lingual equating, it is reasonable to expect that these factors will occur more frequently than in the same-language framework, as explained below:

(1) Group ability differences: In many cases, the populations that take the test versions in the different languages differ not only in the language they speak but also in other important cultural and educational characteristics. As a result, the different language groups may also differ with respect to the proficiency measured by the test (for example, see Angoff & Cook, 1988; Beller et.al., 1999).

(2) Non-representative anchor: In most cases, the anchor items used to link the different language tests onto a common scale do not represent the entire test properly. This is especially true in adapting verbal aptitude tests because some item types cannot be translated. Even when items are translatable, it is quite likely that their psychometric characteristics will change following translation (Allalouf, Hambleton & Sireci, 1999). In such cases, the items cannot be considered equivalent across the source- and target- language version and should be considered unique to the different language versions. Typically, cross-lingual linking involves using a Differential Item Functioning (DIF) analyses, used to determine item invariance across languages. Items that display DIF are eliminated from the anchor set (Sireci, 1997). As a result, the anchor used to link between the different language tests in practice is not a miniature of the tests to be equated, as is required for adequate equating.

(3) Test differences: When adapting a test to a different language it is difficult to ensure that it will keep the same content specification and the same difficulty level as that of the source language test. Furthermore, sometimes the translated version is built intentionally with a different difficulty in order to assure it will have other important psychometric characteristics. For instance, when the language groups differ considerably in ability, the reliability of the translated test may be lower. To gain a higher reliability, it is necessary to adapt the level of difficulty of the target language test to the level of ability of the target language group (Beller and Gafni, 1995).

While equating the scores of the Prueba de Aptitud Academica (PAA) and the Scholastic Aptitud Test (SAT), Angoff & Cook (1988) pointed out the problematic character of some of these issues. They claimed that when groups differ greatly in ability, it is not likely that any set of common items, however appropriate, can make adequate adjustments for the differences. This is true even if the two tests were designed for examinees of the same language and culture.

The purpose of this study is to go one step further and to evaluate how equating in typical cross-lingual conditions is affected by such “classic” problematic factors. In particular it will focus on the effect of a large ability difference between the examinee groups taking the different language versions and of the non -representativeness of the anchor test on equating. A central issue is whether equating bias resulting from such problematic factors is acceptable in practice. In other words, is the equating outcome as questionable as Angoff and Cook (1988) and Kolen and Brennan (1995, see below) claimed?

The effect of ability differences between groups and of the non-representativeness of the anchor item set on equating

Kolen and Brennan (1995) claimed that a large difference in ability between groups in the common item equating design leads to failure of the statistical assumptions that hold for any equating method. This can cause significant problems in estimating the equating relationship. In their experience, mean differences between the two groups larger than 0.5 standard deviation units or ratios of group standard deviations greater than 1.2 can be especially troublesome. When there are large differences, the Levine and IRT equating methods might function more adequately than the other methods, provided that the common items and the alternate test forms measure the same construct. However, when the group differences become too large, no method is likely to function well.

Furthermore, considerable group differences in ability level may also enhance equating error that results from inadequate content representation of the anchor. As Cook and Petersen (1987) summarized, the properties of an anchor test can seriously affect conventional equating results. This is especially true as the equating samples become more dissimilar in level and dispersion of ability.

Despite these concerns, a major deviation from the correct equating relationship due to ability differences between groups is not always found. For instance, Petersen et. al. (1982) and Marco et. al. (1983) compared the effectiveness of different equating models for equating the SAT-verbal under several conditions. They reported that the level of similarity between the examinee samples used for equating (i.e., similar v.s. dissimilar in ability) had a relatively small and unsystematic effect on the quality of the equating results. This is true provided that the anchor test is similar in content and in difficulty (or even in difficulty alone). In their studies, most of the linear models gave satisfactory results when dissimilar ability samples were used instead of similar samples. On the other hand, the equating results of all the linear models were biased significantly when the anchor and one of the alternate test forms or when the two test forms differed in difficulty. Thus, ability differences between groups did not seem to be the critical factor in biasing the equating results, but rather the equating method used and the differences existing between the tests.

The current study

The current study is part of a broader attempt to evaluate a cross-lingual equating process that has been used for several years at the National Institute for Testing and Evaluation (NITE) in Israel. It is obvious that the cross-lingual equating process is conducted in far from ideal conditions because (1) a large ability difference exists between the different language examinee groups, (2) it uses a non-representative anchor item set, and (3) the adapted version of the test is intentionally different in level of difficulty. Hence, it was speculated that even if the adaptation from the source language to the target language was perfect and the different language versions were measuring exactly the same construct, the equating outcome would be seriously affected.

Evaluating an equating process is difficult because the true equating relationship between test versions is never known. It is even more difficult and complex in the cross-lingual context where it is impossible to consider the translated items as identical. While there are several procedures which can be used to evaluate equating quality and analyze the sources of equating error for ordinary (same-language) equating, hardly any of them can be applied to the cross-lingual equating case because of the uncertainty that the items are truly common across languages. Indeed, estimating the effect of specific factors on equating has been reported only in cases of regular (same-language) equating but not in cases of cross-lingual equating. This study focuses on the effect of group ability differences and a non-representative anchor on an operational cross-lingual equating process. A method was especially designed in order to explore these effects in a cross-lingual framework. This method allowed us to investigate the influence of a specific factor on equating while controlling the effect of item translation. The assumption that items that are translations of one another are fully equivalent is incorporated in the design.

Method

Instruments

The test that was analyzed was the verbal sub-test of the *Psychometric Entrance Test* (PET) - a high-stakes test used for admissions to universities in Israel (see Beller, 1994). It is a multiple-choice test consisting of three sub-tests: verbal, quantitative and English as a foreign language. PET verbal and quantitative sub-tests are written in Hebrew and translated into five languages: English, French, Spanish, Russian and Arabic. As with most cross-lingual test adaptation, adapting PET from Hebrew to each of the target languages involves specific problems (see Beller et. al., 1999 for a brief overview of the translation problems of PET).

Each verbal sub-test of PET consists of two parallel sections of 30 items each. In adapting a section from Hebrew into a target language, only about 20 items are translated while the others are specially constructed for the target language versions. Equating between the target language version of a section and the Hebrew version is carried out using the "monolingual separate groups design" and, in most cases, according to the Levine observed score method (Kolen & Brennan, 1995). The common items on which equating is based are selected from the translated items following a DIF (Differential Item Functioning) analysis intended to identify items whose psychometric characteristics have changed following translation (Allalouf, 2000). In general, the final anchor item set for a verbal section consists of 8 to 16 items. Of course, this anchor cannot be considered to be a miniature of the equated sub-tests since it is not composed to the same proportions of item types.

This study focuses on the adaptation of the verbal PET sub-test from Hebrew into Arabic. For many years and over many PET forms it has been repeatedly found that Arabic-speaking examinees taking the test in Arabic differ greatly in ability from the Hebrew-speaking examinees. The typical mean score difference between the two groups is about one standard deviation and the ratio between the standard deviations of the scores is usually higher than 1.2. According to Kolen and Brennan's (1995) rules of thumb, these differences are critical for equating.

The data from two PET forms, form #1 and form #2, were used in the study. Each form was administered to Hebrew-speaking examinees, adapted into Arabic and administered to Arabic-speaking examinees. In each form, the equating relationships between one of the verbal sections in the Hebrew version (section HE1) and its Arabic adapted version (AR1) were analyzed. The other verbal section in each form (HE2) served only for building the special experimental section (see below).

Procedure

Basically, in order to overcome the problem of bias in equating caused by translation distortions, we simulated the equating conditions that exist in the actual Arabic-Hebrew equating framework in a same language context. This was obtained by assembling a special "pseudo Arabic" (PA) section out of Hebrew items, and using it instead of the actual Arabic section AR1. The PA section was built to the same specifications and in a similar way as was section AR1. It consisted of items taken from section HE1which represented the translated items, and supplementary items. The "translated" items were the items that had actually been translated into Arabic in the course of adapting section HE1 into Arabic. The supplementary items were taken from section HE2 and represented the items that were written directly in Arabic when constructing AR1. Since the Arabic version of the Hebrew verbal sections is built to have a somewhat lower level of difficulty[1], the supplementary items chosen from HE2 were the relatively easy items of the section. Figure 1 depicts the equating design used in the study (right schema) and the actual cross-lingual design (left schema). As can be seen, the study design was parallel to the actual equating design in most aspects. The main difference between the two designs is that in the study, the pseudo-Arabic section PA replaced the actual Arabic section AR1. Therefore, the equating implemented in the study was between two Hebrew sections while in the real-life situation it was a cross-lingual equating between an Arabic section and an Hebrew section[2].

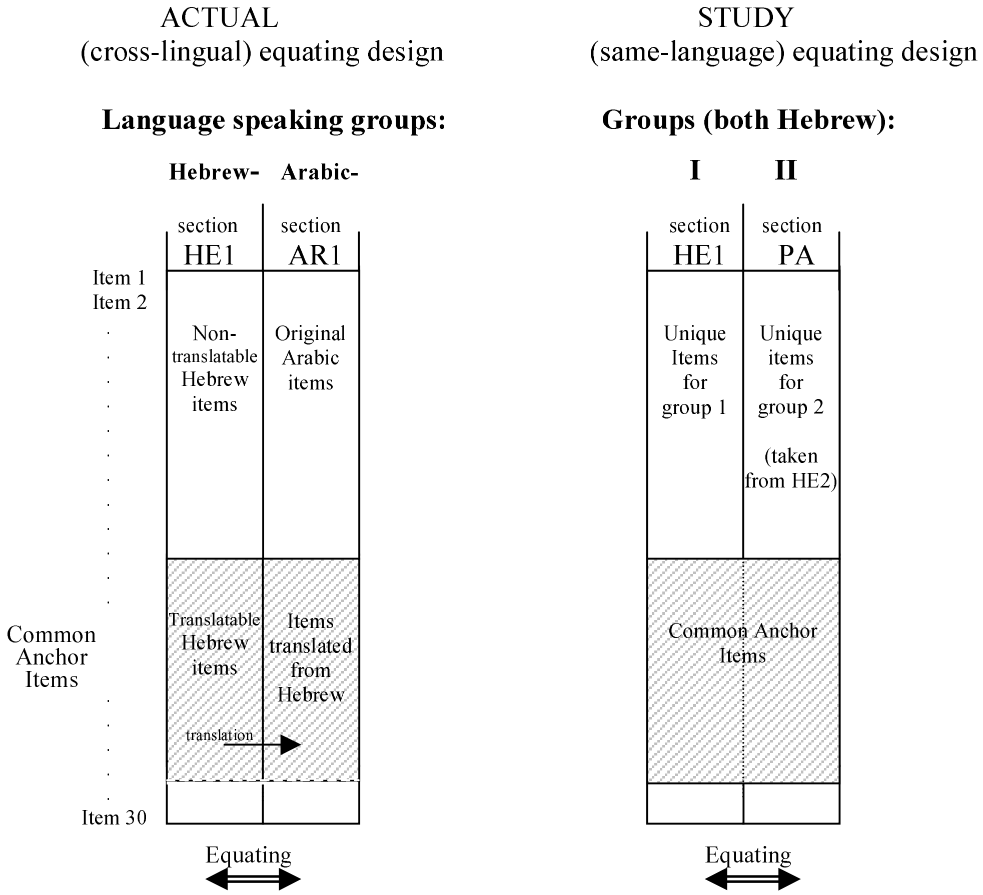
For the equating, two samples of examinees were chosen from the database of the Hebrew-speaking group that was administered the specific Hebrew form. One group (group I) represented the Hebrew-speaking examinee population and the other group (group II) represented the Arabic-speaking examinee population. We treated group I as if it had been administered section HE1 section and group II as if it had

[1] This is done in order to increase the reliability of the Arabic test version (Beller et.al.1999)

[2] Equating in the study was implemented using the Levine observed score method which is the method usually used to equate between Arabic and Hebrew PET verbal sections.

been administered section PA. Since both groups were sampled from the same population, it could be assumed that they were similar in many important aspects. Hence, it was postulated that no interaction between specific content or item types and group performance existed. And, most important, it was assumed that the two sections to be equated in the study, HE1 and PA, measured exactly the same construct in both groups.

FIGURE 1. Graphical depiction of the actual equating design and the parallel study design.



HE1	Operational verbal section in a given Hebrew PET form
HE2	Other operational verbal section in the same PET form
AR1	Operational verbal section in a given Arabic PET form. It is an adapted version of HE1. It includes some translated items from HE1 and supplementary items written directly in Arabic.
PA	Pseudo-Arabic section: An experimental verbal section built to the same specification as AR1 but in Hebrew. It was especially assembled for the study and included some items from HE1and supplementary items from HE2.

1) Evaluating the effect of the group ability differences.

Equating between PA and HE1 was implemented using different samples of examinees (similar vs. dissimilar) and using different anchor item sets (representative vs. non-representative). In the “similar samples” condition both samples had the same score distribution as the actual Hebrew-speaking examinee group had on HE1. In the dissimilar condition, one sample had the score distribution as the actual Hebrew-speaking examinee group and the other sample had the same score distribution as the actual Arabic-speaking examinee group had on AR1. The dissimilar condition imitated the actual situation in the sense that the groups differed in ability just as the actual Arabic- and Hebrew-speaking examinee groups differed. In contrast, the similar condition represented an ideal equating setting, in which groups do not differ in ability. Comparing the equating functions in the two conditions would provide an indication of the equating bias due solely to group differences in level and dispersion of ability[3].

2) Evaluating the effect of the non-representativeness of the anchor.

The equating relationships were calculated using two anchor sets: a representative and a non-representative one. In the non-representative condition the anchor item set consisted of the items that were actually translated into Arabic and actually served as the anchor item set between the HE1 and the AR1 sections. In one of the PET forms analyzed in the study, the item set common to HE1 and AR1 consisted of only 8 translated items, and in the other PET form it consisted of 16 items. In the representative anchor condition, an “ideal” anchor set that kept the same proportionality of item types as in the entire section was assembled. This anchor

[3] In an additional analysis that focused on the actual Arabic language- and Hebrew language-sections, the AR1 section was equated to HE1 using similar examinee samples and compared to the actual equating relationships. An Arabic-speaking examinee sample that had a score distribution on AR1 similar to the score distribution of the Hebrew-speaking examinee group on HE1 was used for the “similar sample condition”. Alternatively, a Hebrew-speaking examinee sample that had a score distribution on HE1 similar to the score distribution of the actual Arabic-speaking examinee group on AR1 was used. The results of this analysis were consistent with the findings of the principal analysis.

included item types that are usually not translated into Arabic. In one case it consisted of a third of the section items (10 items out of 30), and in the second case it consisted of half of the section items (15 items out of 30). The distance between the equating functions using an ideal vs. actual anchor would provide an indication of the bias in equating resulting from using a non-representative anchor item set.

Results

Tables 1a and 1b (for the two PET forms respectively) present statistics of the raw scores obtained by the two samples on the respective sections and on the anchor item set, for the similar and dissimilar samples condition. In both forms, the anchor item set was based on the anchor items used in the actual cross-lingual equating (and hence was a non-representative anchor). As can be seen from the tables, when the two samples were similar, the mean scores in the anchor set were similar for both groups. However, the mean score was considerably higher in the PA section than in the HE1 section (20.7 vs. 19.8 in form #1 and 19.7 vs.17.4 in form #2). This indicates that the PA section was easier than the HE1 section. When the samples were dissimilar, the mean differences in the anchor set between the pseudo-Arabic group and the Hebrew-speaking group were almost one standard deviation. They were 5.7 vs. 4.1 raw score points (S.D.=1.9) in form #1 and 9.5 vs. 6.6 raw score points (S.D=3.3) in form #2. In addition, in form #2, the ratio between the standard deviations in the dissimilar groups condition was considerably higher than 1.2 (3.3/2.4=1.375). These differences are considered large in relation to any equating criteria reported in the literature.

Another important feature demonstrated in the tables is that the reliability of the PA section drops considerably when the examinee sample to which it is administered is low in ability (it drops from 0.8 to 0.76 in form #1 and from 0.84 to 0.65 in form #2). As explained previously, this demonstrates how the problem of reliability is related to the issue of ability differences between groups and has to be taken into consideration.

Following Tables 1a and 1b, Graphs 1a and 1b show the equating functions obtained in the similar- and dissimilar-samples conditions. The distances between the lines in the two forms were about 0.5 to 1 raw score points in the center of the scale and 1 to 2 raw score points at the ends. These differences are smaller than expected and do not seem to be critical for equating. Furthermore, in the two PET forms analyzed, although the dissimilarity between the groups was in the same direction, the equating bias was manifested in opposite directions. This finding refutes the idea that dissimilarity between groups causes a systematic and directional bias in equating.

The next graphs, 2a and 2b (for the two PET forms respectively), show the equating functions obtained in the representative vs. non-representative anchor condition for dissimilar examinee samples. The distances between the equating functions were about one raw score point along most of the raw score scale in form #1 (see Graph 2a) and between -1 to +0.7 in form #2 (see Graph 2b).

These differences were reduced considerably when the same procedure was carried out with similar samples of examinees instead of dissimilar samples (not shown on the graphs). In form #1 the distance dropped to about half a raw score, and in form #2 the distance dropped to almost zero. Thus, the effect on equating of the non- representativeness of the anchor is smaller when the ability difference between the examinee groups taking the test version is small. Similarly, the ability difference effect was smaller when the anchor was built representatively, i.e. in form #2 we found that the ability difference effect reported above was reduced by about half when the anchor set was representative instead of non-representative.

Table 1a

Statistics of the samples' raw scores on the Hebrew and pseudo-Arabic test sections and on the anchor - form #1 (non-representative anchor, anchor length: 8).

		Test sections (30 items)			*Anchor (8 items)*			
Section	N	Mean	SD	r_{vv}	Mean	SD	r_{uu}	r_{vu}
			Similar samples					
Hebrew (HE1)	3733	19.8	5.3	0.80	5.7	1.8	0.59	0.79
Pseudo-Arabic (PA)	3706	20.7	5.2	0.80	5.7	1.8	0.60	0.78
			Dissimilar samples					
Hebrew (HE1)	3733	19.8	5.3	0.80	5.7	1.8	0.59	0.79
Pseudo-Arabic (PA)	1869	15	5.2	0.76	4.1	1.9	0.55	0.75

Graph 1a

Linear equating relationship between Hebrew and pseudo-Arabic test sections using similar vs. dissimilar samples- form #1

(The distance between the two lines at various score-points is shown).

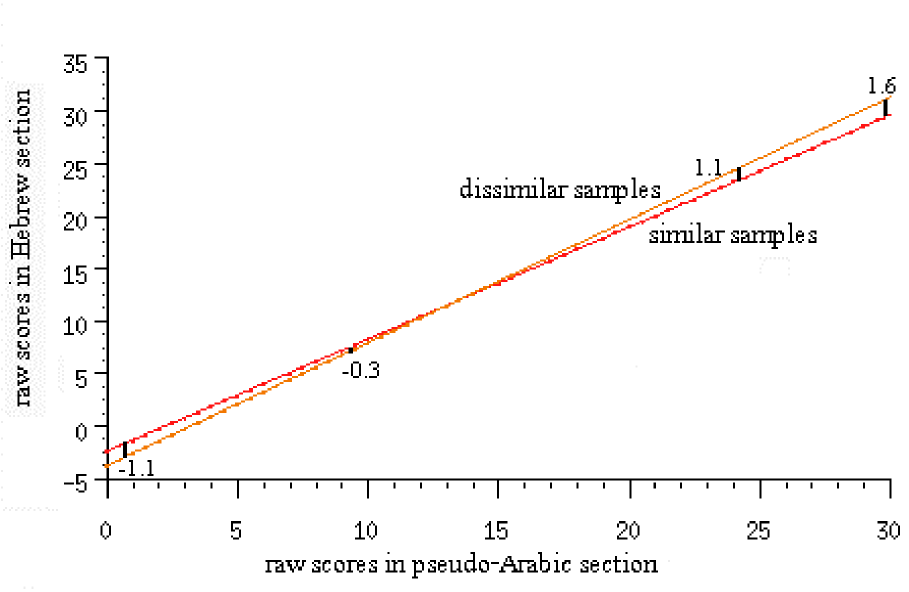


Table 1b

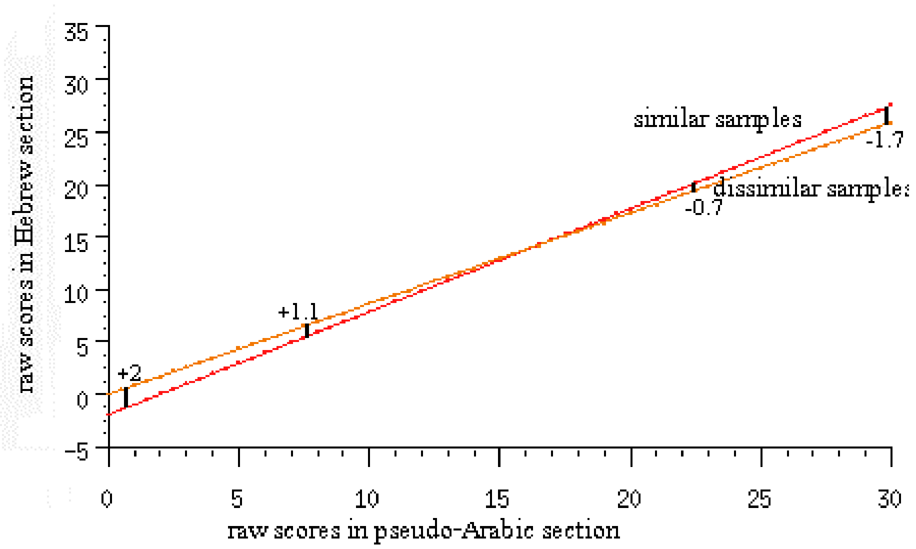
Statistics of the samples' raw scores on the Hebrew and pseudo-Arabic test sections and on the anchor - form #2 (non-representative anchor, anchor length: 16)

		Test sections (30 items)			*Anchor (16 items)*			
Section	N	Mean	SD	r_{vv}	Mean	SD	r_{uu}	r_{vu}
				Similar samples				
Hebrew (HE1)	3772	17.4	5.5	0.81	9.5	3.3	0.73	0.92
Pseudo-Arabic (PA)	3798	19.7	5.7	0.84	9.5	3.4	0.74	0.92
				Dissimilar samples				
Hebrew (HE1)	3772	17.4	5.5	0.81	9.5	3.3	0.73	0.92
Pseudo-Arabic (PA)	1323	14	4.3	0.65	6.6	2.4	0.65	0.82

Graph 1b

Linear equating relationship between Hebrew and pseudo-Arabic test sections using similar vs. dissimilar samples - form #2

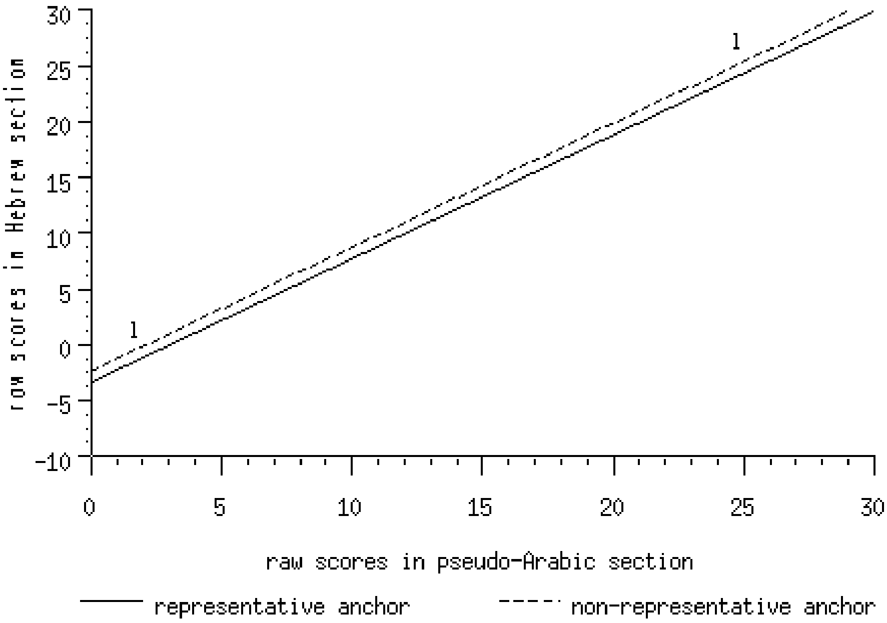
(The distance between the two lines at various score-points is shown).



Graph 2a

Linear equating relationship between Hebrew and pseudo-Arabic test sections using a representative vs. non-representative anchor - form #1

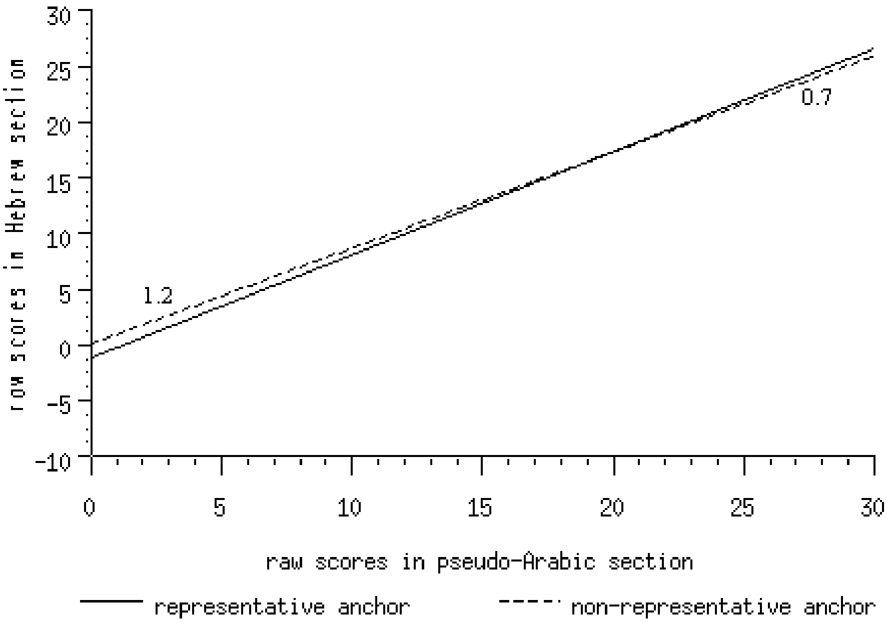
(dissimilar samples were used for the equating)



Graph 2b

Linear equating relationship between Hebrew and pseudo-Arabic test sections using a representative vs. non-representative anchor - form #2

(dissimilar samples were used for the equating)



Discussion

This study had attempted to evaluate how two factors, typical to cross-lingual test linking, affect equating. The difference in ability level and dispersion existing between the two examinee groups taking the test in different language and the fact that a non-representative anchor item set is used for equating. Aside for the findings, the importance of the study lays in the method used, a method that was especially designed for examining the issues in question.

The method was designed to imitate the conditions in which cross-lingual equating is conducted while assuming a perfect translation process that assures full equivalency of the common items across the different language versions. Another assumption incorporated in the method was that the different language tests measure exactly the same construct. Since the samples representing the different language groups were similar in all aspects except for the ability measured in the test, it implied that no interaction existed between the language groups and important variables such as item contents or item type. The only problematic features that were retained in the experimental section as in the actual target language section were the difficulty difference between the sections and the non-representativeness of the anchor item set. Thus, it was possible to extract the systematic error incorporated in cross-lingual equating stemming from the factors in question.

This innovative method can be useful for practitioners who are dealing with cross-lingual equating. The presented method can be used to control and improve the outcome of linking process carried out between different language versions of tests. As in the current study, it can be used to evaluate the effect of specific factors on cross-lingual equating. Alternatively, it can be used to compare among several equating methods in a cross-lingual setting.

The range of equating error that was attributed to the ability difference between groups was not critical and not even directional. That is, even if there were directional differences between some of the language groups, they did not distort the equating outcome in a specific direction and did not disfavor one language group over the other. It seems that some additional variables or random elements contribute to the equating bias. With regards to the error that could be attributed to the non-representativeness of the anchor item set, it seems that this factor, alone or combined with the ability difference factor, was also not critical to equating. In general, it could be postulated that these two factors did not affect equating to an unacceptable extent. Moreover, it is important to note that in practice, each equating procedure is implemented independently between one PET target language and one source language version. Thus, the equating error does not accumulate over several equating processes and the damage is restricted only to the specific equating procedure.

However, these results were not consistent with our expectations and with the literature. According to Kolen & Brennan ‘s (1995) summary of the issue, when ability differences as large as these exist and when the anchor item set is not built to represent properly the entire test, it can cause serious problems for any equating method to hold. The current findings are unlike the findings of studies in which considerable group differences affected equating dramatically (e.g. Cook & Petersen, 1987). On the other hand, they are similar to the findings of some papers, in which the differences between equating using similar and dissimilar samples of examinees tend to be moderate (e.g. Petersen et.al., 1982 and Marco et.al.,1983). This inconsistency suggests that these effects are test specific or situation specific. It is also possible that the occurrence of the effect depends on the presence of other variables that interact with performance, for example the recency of coursework (Cook, 1984, reported in Cook & Petersen, 1987). However, when the difference between the groups is restricted to ability per se, the effect may be weaker than was thought. This situation rarely occurs in practice, neither in cross-lingual equating nor in regular equating, but it has important theoretical implications. It suggests that, as far as the different language test versions measure the same construct, and the group differences are restricted to that specific measured construct, equating across languages could be quite valid. Hence, practitioners who deal with cross-lingual testing should put the greatest emphasis on proper translation in order to achieve this goal.

As to the effect of the non-representativeness of the anchor item set on equating, it was, as expected, larger when the samples of examinees were dissimilar rather than similar. But although it was quite consistent along the entire score scale and it was relatively restricted and not meaningful. One possible explanation for this finding is that the non-representativeness of the anchor in the current case was expressed in the proportion of item types, not in their content. It is likely that although the anchor set did not adequately represent all types of items, it still represented the same content and measured the same constructs as do the sections to be equated.

In conclusion, the present study has two important implications. First, it presents a serious attempt to approach the unattainable goal of evaluating the process cross-lingual equating and offer a new tool to be used by practitioner who deal with cross-lingual equating. Secondly, it has been demonstrated that factors considered to be critical to equating are not necessarily so critical as they were thought to be. It seems that by properly extracting the influence of these factors on equating, a lesser effect is found. Finally, it demonstrates how the robustness of an equating process in face of aberrant factors is test and situation specific.

References:

Allalouf, A. (1999). Scoring and equating at the National Institute for Testing and Evaluation (Research report 269). Jerusalem: National Institute for Testing and Evaluation.

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

Angoff, W.H. & Cook, L.L. (1988). Equating the scores of the Prueba de Aptitud Academica and the Scholastic Aptitude Test (College Board Report No. 88-2). New-York : College entrance Examination Board.

Angoff, W.H. & Modu, C.C. (1973). Equating the scores of the Prueba de Aptitud Academica and the Scholastic Aptitude Test (College Board Report No. 88-2). New-York : College entrance Examination Board.

Beller, M. (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.

Beller, M., Gafni, N. & Hanani, P. (1999). Constructing, adapting and validating admissions tests in multiple languages. Paper presented at the international conference on adapting test for use in multiple languages and cultures, Georgetown University, Washington, DC; and to appear in Hambleton, R. K., Merenda, P. & Spielberger, C. (Eds.). (in press). Adapting educational and psychological tests for cross-cultural assessment. Hillsdale, NJ: Erlbaum.

Budescu, D. (1985). Efficiency of linear equating as a function of the length of the anchor test. Journal of Educational Measurement, 22, 13-20.

Cook, L.L. (2000). Factors affecting the validity of scores obtained on tests given in different languages to examinees of different cultural backgrounds. Paper presented at the annual meeting of the International Association for Educational Assessment, Jerusalem.

Cook, L.L., & Petersen, N.S. (1987). Problems related to the use of conventional and item response theory equating methods in less than optimal circumstances. Applied Psychological Measurement, 11, 225-244.

Harris, D.J. (1991). Equating with nonrepresentative common item sets and nonequivalent groups. Paper presented at the annual meeting of the American Educational Research Association, Chicago.

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

Hulin, C. L., & Mayer, L. J. (1986). Psychometric equivalence of a translation of the Job Descriptive Index into Hebrew. Journal of Applied Psychology, 71, 83-94.

Kolen, M. J., & Brennan, R. L. (1995). Test equating, methods and practices. New-York: Springer-Verlag.

Marco, G.L., Petersen, N.S., & Stewart, E.E. (1983). A test of the adequacy of curvilinear score equating models. In D. Weiss (Ed.), New horizons in testing, 147-176. New York: Academic Press.

Petersen, N.S., Marco, G.L., & Stewart, E.E. (1982). A test of the adequacy of linear score equating models. In P.W. Holland and D.B. Rubin (Eds.), Test equating, 71-135. New York: Academic Press.

Petersen, N.S., Cook, L.L., & Stocking, M.L. (1983). IRT versus conventional equating methods: A comparative study of scale stability. Journal of Educational Statistics, 8, 137-156.

Sireci, S.G.(1997). Problems and issues in linking assessments across languages. Educational Measurement: Issues and Practice, 16 (1), 12-19, 29.

Schmitt, A., & Dorans, N.J. (1999). Linking scores from tests of similar content given in different languages: The case of the spanish Language PAA and the English Language SAT I. Paper presented at the Annual meeting of the National Council on Measurement in Education, Montreal.

Wainer, H. (1999). Comparing the incomparable: an essay of the importance of big assumptions and scant evidence. Educational Measurement: Issues and Practice, 18 (4), 10-16.

Woodcock, R.W., & Munoz-Sandoval, A.F. (1993). An IRT approach to cross-language test equating and interpretation. European Journal of Psychological Assessment, 9, 233-241.