Evaluating Cross-Lingual Equating[1]

Joel Rapp & Avi Allalouf

National Institute for Testing and Evaluation

[1]This paper was presented at the annual meeting of the American Educational Research Association, New-Orleans, 2002

Abstract

Adapting educational tests from one language to others requires equating across the different language versions. Such equating is usually based on translated items considered to have similar content and psychometric characteristics in both source and target languages. However, because it is not possible to ascertain that these items are really similar in different languages, it is difficult to control and validate the equating outcome. The purpose of this study was to develop a method for evaluating cross-lingual equating and apply it to the Psychometric Entrance Test (PET) used for admission to Israeli universities. This test is written in Hebrew (the source-language, SL) and translated into 5 languages. A cross-lingual equating in a double linking plan was performed in each of 12 forms translated to one target language (TL1), and each of 9 forms translated to another language (TL2). The average difference between the equating results in the two links, indicating the overall instability incorporated in the equating process, was more than 10 times the standard error of equating in the TL1-SL process and about half this size in the TL2-SL process. The significance of the results and the differences found between the two TL's are discussed, as well as the potential displayed by the method for use as a general evaluative tool for cross-lingual equating.

Introduction

In recent years interest in adapting psychological and educational tests from one language to another has increased. This is a result of the spread in comparative studies across national, ethnic and cultural groups and the wish to compare the educational achievement or aptitude of students in different countries and cultures. In many cases, psychometricians who deal with testing across different languages (DL) and who strive to achieve score comparability between the translated version of a test and its source version face serious difficulties. These difficulties are related both to the differences between the languages of the test forms and to the differences in culture between the examinee groups to whom the test forms are administered. Hambleton (1993, 1994), Sireci (1997), Hambleton and Patsula (1998), Cook (2000) and Wainer (1999) review the various methods for linking tests across languages and the difficulties related to these methods. Three general linking designs exist for cross-lingual equating: (1) the bilingual group design; (2) the matched monolingual groups design; and (3) the separate monolingual groups design.

1) The bilingual group design: In this design, a group of bilingual examinees assumed to be equally proficient in both languages with respect to the construct being measured are tested in the two language versions of the test. Any difference in achievement between the two language versions can be attributed to differences in the difficulty of the two versions. Although this method seems promising, its major problem is that in practice it is very difficult to find examinees who are equally proficient in the two languages. Even if such examinees can be found, they probably do not represent in any sense the examinee population for whom comparability is required.

2) The matched monolingual group design: In this design, a group of examinees from each language is selected so that they are matched on particular criteria, such as socio-economic status and level of education. The achievement of these groups in the test is then compared. The main problem with this methodology lies in the need to choose relevant and available criteria for matching.

3) The separate monolingual groups design. This is the most popular design used for cross-lingual equating. In this linking design, source-language (SL) and target-language (TL) versions of a test are administered separately to source- and

target- language examinee groups respectively. In fact, this method is a variation of the familiar "common-item, non-equivalent groups" design, used for "regular" same language equating (Angoff, 1984, Kolen & Brenann, 1995). In this design, a set of items common to the two tests is used to link the scores. While in same-language equating the anchor items are identical, in the cross-lingual equating the anchor items are chosen from the translated items. These items are treated as if they were identical, as if they measure the same construct and as if they have the same psychometrical characteristics. One well-known case in which this equating method was applied is the linking of the Scholastic Aptitude Test (SAT) to its Spanish counterpart, the Prueba de Aptitud Academica (PAA) (Angoff and Modu, 1973, Angoff and Cook, 1988 and Schmitt and Dorans, 1999). Sireci (1997) lists additional examples in which this design was used for educational as well as industrial tests. Indeed, this method seems to be relatively easy to apply, since it does not require examinees with specific characteristics (bilingual, "matching") which might be difficult or impossible to find. However, the separate monolingual groups procedure suffers from a theoretical flaw since it cannot be ascertained that all the translated items used as anchor items are equivalent across languages - a basic requirement for any equating system to hold. In addition, as in all of the cross-lingual equating procedures, it is practically impossible to ensure that the DL test versions measure exactly the same construct. Thus, there is a high risk that a large equating error will accompany implementation of this procedure.

Another difficulty faced by practitioners who wish to link test forms across languages is the lack of suitable methods, criteria or research tools to be used for validating and controlling the outcome of such equating. In other words, not only it is difficult to achieve satisfactory cross-lingual equating, it is also difficult to verify its outcome and confirm its accuracy.

This study arose from the need to evaluate the quality of a cross-lingual equating procedure that has been conducted for many years in a large-scale testing program, namely, the Psychometric Entrance Test (PET), required for entrance into Israeli universities. PET is written and administered in Hebrew and later translated into five languages so that non-Hebrew-speaking can be tested in their native language. Once the translated versions have been administered, an equating procedure using the separate monolingual groups design is applied. The major problem with this

process is that it is based on the untestable assumption that the common items on the SL and TL versions retain the same meaning and psychometric properties following translation. Other inevitable methodological problems with the process exist: (1) the anchor item sets do not properly represent the complete test, and (2) there tend to be wide ability differences between the SL group and some of the TL groups. In the common-item equating design these factors are known to introduce error in equating. Obviously, these factors cause equating errors in the cross-lingual case too.

Since there is no widely used method for evaluating cross-lingual equating, a method was especially developed for this study, one that could eventually be incorporated into the routine equating process. It was hoped that this would result in greater equating stability and a smaller equating error. We believe that the methods and procedure used in this study can easily be adapted to every test where score comparability among DL versions of the test is needed.

Developing a method for evaluating and validating cross-lingual equating

In developing new methods for cross-lingual equating, a reasonable strategy is to adapt procedures that are traditionally used for regular (same-language) equating. Here, the concept of "double linking" was adopted and adapted to examine the quality and stability of the cross-lingual equating process. In the double-linking equating plan, a new test form is independently equated to two old forms. Usually, the two conversion functions are averaged to produce a single conversion. If the two conversion functions differ more than would be expected by chance, it would suggest that a systematic error occurred in at least one of the equating processes. The double linking equating plan is recommended in the literature as "providing a built-in check on equating and leading to greater equating stability" (Kolen and Brennan, 1995, pp. 256-257), but it also introduces more complications into the equating process than does a single link. When using the common-item non-equivalent group design, double linking requires that two sets of common items, which are representative of the test content, be used in the development of new test forms. Applying this principle to the cross-lingual equating context imposes a serious complication on the development of translated forms, and might not be feasible. In fact, new translated forms have to be

assembled with two sets of translated items, each taken from a different source language (SL) form, and each representative of the test in terms of content and statistical parameters. In this study a somewhat similar method was developed, based on the fact that each PET form consists of pairs of parallel sections (see below). In principle, the double link between the TL and SL form was created by equating each of the two parallel sections of a TL form to its respective section in the SL form.

The test

PET is an admissions test used by the universities in Israel and is similar in nature to the Scholastic Assessment Test (SAT) of the Educational Testing Service. It consists of three sub-tests: Verbal Reasoning (V), Quantitative Reasoning (Q) and English as a Foreign Language (E). PET is administered in Hebrew and adapted into five target languages. Of the three sub-tests, only the V subtest is equated across the language forms. The two other sub-tests are considered to be directly comparable and as such, they are not equated across languages. Each V sub-test consists of two parallel sections of 30 items each. In each section, about twenty items are translated from the SL to the TL, while the others are written directly in the TL. This is either because some item types simply cannot be translated, or because the translation of these items results in a grave distortion of their content (see Beller, Gafni & Hanani, 2000 for a detailed overview of the translating process of PET and the problems involved with it).

The cross-lingual equating of PET V sections

The scores of the V sections in the TL version are equated to the score scale of the counterpart SL section using the separate monolingual design and the Levine linear method. The common items used in the equating process are translated items selected according to a DIF (Differential Item Functioning) analysis, intended to identify items whose psychometric characteristics have not significantly changed following translation (see Allalouf, 1999). As mentioned previously, the cross-lingual equating process is based on the following assumptions: (1) The translated items used as common items across the SL and TL language versions are equivalent, i.e., they

retain the same meaning and psychometric properties (e.g. difficulty level) following translation, and (2) The additional items in each TL section, which are written directly in the TL, represent the same content and measure the same psychological construct as the SL verbal section. Taken together, assumptions 1 and 2 imply that the SL and the TL verbal sections measure the same construct. This study examines the PET V subtest cross-lingual equating process.

The study was conducted on several translated forms of PET separately for two target languages, hereafter TL-1 and TL-2. Data from twelve PET forms in TL-1 and nine forms in TL-2 administered between 1997 and 2001, and from the parallel forms in SL were used. TL-1 and the source language belong to the same family of languages and resemble each other in grammar. The same item types can be used in both languages. In contrast, TL-2 does not resemble the source language at all. Nevertheless, the translation of the test into TL-1 is more problematic because this language is characterized by a difference between spoken language and written language. Examinees who are not familiar with literature in this language might find some words in the test unfamiliar and difficult. In addition, the examinee groups to whom the adapted tests in TL-1are administered differ from the SL group in many relevant and important aspects such as: educational background, socio-economic level, culture, motivation, within-group diversity and so on.

Tables 1 and 2 summarize data concerning the cross-lingual equating process in TL-1 and in TL-2 respectively. For each PET form the following variables are presented for the SL and TL examinee groups: (1) the number of anchor items used in practice for the cross-lingual equating; (2) Mean raw scores and SD's for V1 and V2 (the first ordered and second ordered verbal section within a form) and for the anchor item sets; (3) the reliabilities of the anchor item sets, and (4) correlation-coefficients (r) between the raw scores of the anchor and the complete section. The last row shows the averages across the various PET forms. Note that since the anchor length is different in each section, the anchor raw score varies in meaning. An anchor raw

Table 1: Raw Score Statistics for PET Verbal Section Tests and Anchors in the SL and TL-1 Examinee Groups

Form	Section	No. of Anchor Items	Source-Language (SL)					Target-Language 1 (TL-1)				
			N	**Test** Mean (SD)	**Anchor** Mean (SD)	Rel.[1]	Cor.[2]	N	**Test** Mean (SD)	**Anchor** Mean (SD)	Rel.[1]	Cor.[2]
X4	V1	12	6459	*18.7 (5.5)*	*8.0 (2.6)*	*.67*	*.85*	1771	*14.4 (4.6)*	*4.8 (2.1)*	*.43*	*.78*
	V2	15		*18.8 (5.7)*	*9.8 (3.1)*	*.71*	*.91*		*13.5 (4.6)*	*5.8 (2.4)*	*.46*	*.83*
A5	V1	8	7433	*19.8 (5.3)*	*5.7 (1.8)*	*.59*	*.79*	1875	*15.0 (5.2)*	*3.5 (1.8)*	*.46*	*.76*
	V2	9		*19.8 (5.5)*	*6.3 (2.1)*	*.65*	*.81*		*15.5 (5.4)*	*3.8 (1.9)*	*.48*	*.78*
A7	V1	16	5056	*20.9 (5.8)*	*11.1 (3.4)*	*.77*	*.94*	2007	*14.1 (4.6)*	*6.8 (2.7)*	*.55*	*.87*
	V2	19		*20.9 (5.9)*	*12.7 (3.9)*	*.80*	*.96*		*13.5 (4.6)*	*7.1 (2.9)*	*.56*	*.91*
A8	V1	14	7861	*19.0 (6.2)*	*8.9 (3.2)*	*.76*	*.92*	2645	*15.8 (5.1)*	*6.4 (2.7)*	*.59*	*.88*
	V2	12		*19.1 (6.1)*	*7.7(2.7)*	*.70*	*.88*		*15.2 (5.0)*	*5.5 (2.3)*	*.50*	*.83*
D2	V1	16	7141	*20.2 (5.9)*	*10.2 (3.5)*	*.76*	*.93*	2327	*16.1 (4.7)*	*6.3 (2.8)*	*.57*	*.87*
	V2	17		*21.3 (5.4)*	*12.4 (3.4)*	*.77*	*.93*		*17.1 (4.9)*	*7.9 (2.9)*	*.59*	*.89*
D4	V1	15	7122	*18.6 (5.8)*	*9.3 (3.2)*	*.73*	*.92*	2821	*16.7 (4.9)*	*7.3 (2.7)*	*.58*	*.88*
	V2	11		*19.3 (5.6)*	*7.9 (2.3)*	*.67*	*.88*		*15.7 (4.8)*	*5.2 (2.1)*	*.50*	*.81*
Z6	V1	10	4271	*18.6 (5.8)*	*6.0 (2.2)*	*.59*	*.86*	1860	*15.2 (4.6)*	*4.2 (1.9)*	*.36*	*.79*
	V2	8		*18.9 (6.2)*	*5.6 (1.9)*	*.62*	*.85*		*15.9 (5.4)*	*3.5 (1.8)*	*.50*	*.79*
D6	V1	15	7566	*17.5 (5.5)*	*9.5 (3.3)*	*.73*	*.92*	1324	*14.0 (4.3)*	*6.3 (2.3)*	*.42*	*.81*
	V2	13		*18.9 (6.0)*	*7.9 (2.9)*	*.71*	*.90*		*13.8 (4.3)*	*5.3 (2.0)*	*.36*	*.76*
N9	V1	10	7306	*19.7 (6.2)*	*7.1 (2.3)*	*.70*	*.87*	3097	*15.7 (4.5)*	*4.0 (1.9)*	*.41*	*.78*
	V2	11		*19.6 (6.4)*	*7.5 (2.6)*	*.72*	*.91*		*16.4 (4.9)*	*4.8 (2.0)*	*.44*	*.78*
Z3	V1	13	3848	*19.8 (6.0)*	*9.3 (3.0)*	*.77*	*.91*	2980	*16.5 (5.3)*	*6.0 (2.5)*	*.55*	*.85*
	V2	12		*19.5 (5.7)*	*8.4 (2.6)*	*.69*	*.88*		*17.0 (5.5)*	*5.9 (2.5)*	*.57*	*.86*
Z0	V1	11	5002	*18.4 (5.7)*	*7.4 (2.4)*	*.66*	*.86*	3072	*17.8 (5.2)*	*5.7 (2.2)*	*.53*	*.81*
	V2	12		*19.0 (5.7)*	*8.1 (2.6)*	*.68*	*.88*		*17.2 (4.9)*	*5.7 (2.2)*	*.45*	*.81*
M8	V1	11	7143	*18.9 (6.2)*	*6.8 (2.7)*	*.74*	*.90*	2878	*15.7 (4.7)*	*4.5 (1.9)*	*.40*	*.77*
	V2	13		*19.8 (5.9)*	*8.9 (2.9)*	*.74*	*.90*		*15.9 (5.1)*	*5.3 (2.6)*	*.60*	*.81*
AVERAGE		12.6		*19.4*		*.71*	*.89*		*15.6*		*.49*	*.82*

[1] KR20 Reliability [2] Pearson correlation between test raw score and anchor raw score

Table 2: Raw Score Statistics for PET Verbal Sections Tests and Anchors in the SL and TL-2 Examinee Groups

Form	Section	No. of Anchor Items	Source-Language (SL)					Target-Language 2 (TL-2)				
			N	**Test** Mean (SD)	**Anchor** Mean (SD)	Rel.[1]	Cor.[2]	N	**Test** Mean (SD)	**Anchor** Mean (SD)	Rel.[1]	Cor.[2]
X4	V1	15	6459	*18.7 (5.5)*	*9.3 (3.0)*	*.69*	*.90*	2628	*17.7(4.6)*	*7.7 (2.5)*	*.53*	*.88*
	V2	21		*18.8 (5.7)*	*13.8 (4.2)*	*.78*	*.95*		*18.1 (4.7)*	*11.9 (3.6)*	*.70*	*.95*
X1	V1	14	3945	*19.8 (5.2)*	*8.7 (2.8)*	*.65*	*.88*	4223	*18.5 (4.6)*	*8.1 (2.7)*	*.69*	*.88*
	V2	14		*20.1 (5.4)*	*9.2 (2.9)*	*.69*	*.89*		*18.3 (5.1)*	*7.8 (2.8)*	*.63*	*.89*
A7	V1	18	5056	*20.9 (5.8)*	*12.3 (3.7)*	*.78*	*.95*	3858	*19.2 (5.2)*	*11.1 (3.6)*	*.74*	*.94*
	V2	19		*20.9 (5.9)*	*12.8 (4.0)*	*.80*	*.96*		*19.8 (5.4)*	*11.6 (3.7)*	*.74*	*.95*
A8	V1	13	7861	*19.0 (6.2)*	*8.0 (3.0)*	*.73*	*.90*	2503	*18.6 (4.9)*	*6.9 (2.8)*	*.65*	*.89*
	V2	18		*19.1 (6.1)*	*11.1 (3.9)*	*.78*	*.94*		*19.3 (4.8)*	*9.8 (3.6)*	*.72*	*.95*
D2	V1	19	7141	*20.2 (5.9)*	*13.0 (3.9)*	*.79*	*.94*	2501	*17.1 (4.8)*	*10.4 (3.6)*	*.71*	*.95*
	V2	18		*21.3 (5.4)*	*13.2 (3.5)*	*.77*	*.93*		*20.1 (4.8)*	*10.6 (3.5)*	*.71*	*.95*
Z6	V1	20	4271	*18.6 (5.8)*	*11.7 (4.0)*	*.75*	*.95*	3775	*17.3 (5.0)*	*10.6 (3.5)*	*.67*	*.94*
	V2	19		*18.9 (6.2)*	*11.7 (4.1)*	*.79*	*.95*		*16.7 (4.9)*	*9.7 (3.4)*	*.66*	*.93*
N9	V1	17	7306	*19.7 (6.2)*	*11.0 (3.7)*	*.78*	*.94*	2164	*15.7 (4.4)*	*8.6 (3.2)*	*.65*	*.94*
	V2	20		*19.6 (6.4)*	*12.8 (4.5)*	*.83*	*.96*		*16.2 (4.9)*	*9.8 (3.8)*	*.72*	*.96*
Z3	V1	20	3848	*19.8 (6.0)*	*12.9 (4.3)*	*.81*	*.96*	4338	*17.9 (4.8)*	*11.1 (3.8)*	*.73*	*.96*
	V2	15		*19.5 (5.7)*	*12.4 (3.8)*	*.75*	*.94*		*17.3 (4.9)*	*10.8 (3.4)*	*.67*	*.93*
F5	V1	19	7265	*18.4 (5.7)*	*11.8 (3.9)*	*.77*	*.94*	3093	*18.4 (4.6)*	*10.2 (3.5)*	*.69*	*.95*
	V2	22		*19.2 (5.7)*	*14.2 (4.5)*	*.81*	*.97*		*19.0 (4.8)*	*12.9 (3.9)*	*.71*	*.97*
AVERAGE		17.8		*19.6*		*.76*	*.94*		*18.1*		*.68*	*.93*

[1] KR20 Reliability [2] Pearson correlation between test raw score and anchor raw score

score of 7 in a 10-item anchor is not equivalent to an anchor raw score of 7 in a 15-item anchor. Therefore, an average was not calculated for this column. As can be seen, the anchor sets used in practice usually consist of about thirteen items in TL-1 and eighteen items in TL-2 (out of 30 items in a section). As a result, the reliability of the anchor item set and the correlation between it and the whole section are low in TL-1 (on average, rel.= 0.49 and corr.= 0.82), higher in TL-2 (0.68 and 0.93) and highest in SL (0.71 to 0.76 and 0.89 to 0.94). Since the selection of the anchor set depends on the one hand on practical considerations (not all item types can be translated) and on the other hand on statistical considerations (the DIF procedure), the anchor set that is eventually used is far from being ideal according to the demands of proper equating. It is not necessarily content-representative nor item-type representative as required.

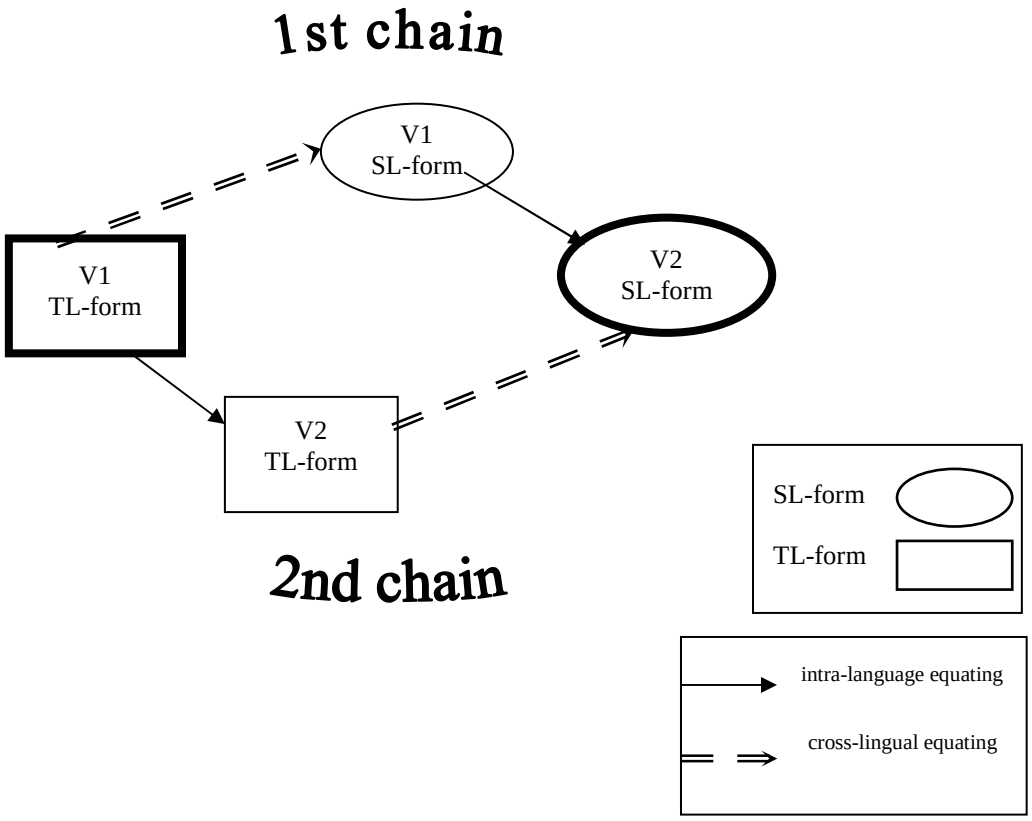
Another problem that was mentioned previously is the large difference in ability between the TL-1 and the SL examinee groups. Table 1 shows clearly that TL-1 examinees score considerably lower than SL examinees. The difference is approximately 4 raw score points (out of 30 items in a section). A similar disparity is found when comparing the group scores on the anchor sets in each individual PET version. The difference between the TL-2 and SL group scores is much smaller and is only about 1.5 points (Table 2). Possibly, the lower reliability of the TL-1 test version stems from the fact that the test is relatively more difficult for what seems to be the less able TL-1 examinee group.

Method

The fact that each PET form, in both the SL and TL, contains two verbal sections built to similar specifications and measuring the same construct helped us in designing the research. According to the experimental design, in each particular PET form, the first-positioned verbal section (V1) in the TL version was equated via two equating chains to the second- positioned verbal section (V2) of the SL version (see

Figure 1). In the first chain, it was equated via section V1 in SL, while in the second chain it was equated via the V2 section in TL. Thus, both chains included one cross-lingual equating link and one same-language equating link, but in reverse order. The cross-lingual equating links in the plan are those conducted in practice and they are the topic of interest in this study. The "intra-language" links between two sections included in the same PET form were established especially for this study. These equating links were executed using the "single group equating design" and the mean-sigma linear method. They were based on large examinee groups, constituting most of the examinees to whom the forms were administered (about 2500 examinees in TL forms and about 7000 examinees in SL forms – see tables 1 and 2).

Figure 1 : Equating V1-TL to V2-SL: the "Double-Linking" Plan



Evaluative tools

The within-language equating link in each of the forms used in the study was assumed to be fairly stable, exhibiting only a relatively small standard error of equating. Thus, it was postulated that the difference found between the equating results in the two chains would mainly reflect the instability of the cross-lingual equating links. The averaged difference over several PET forms served to estimate the range of instability incorporated in the current cross-lingual equating procedure conducted between the TL- (1 or 2) and SL-versions of PET.

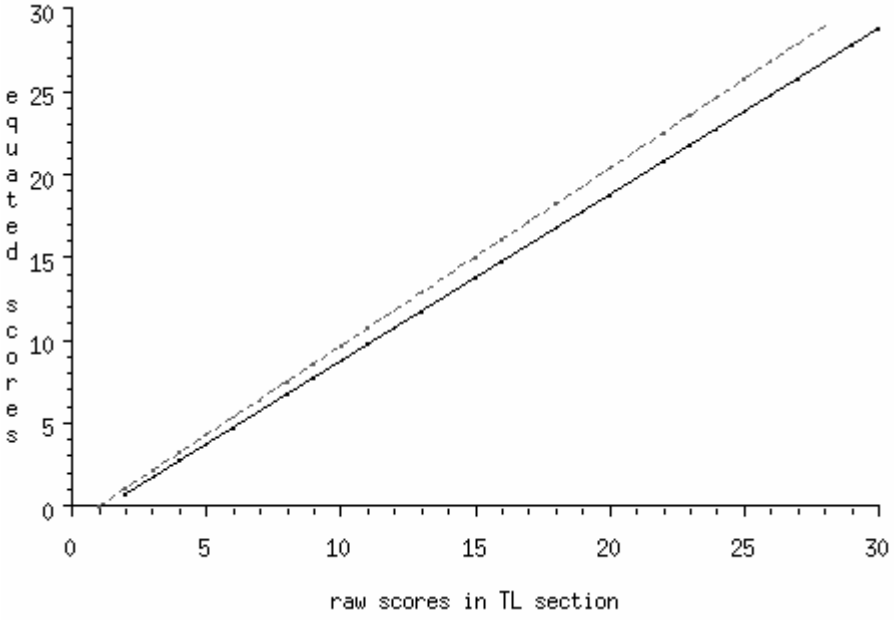
Results

Graph 1 demonstrates a typical result for one of the forms analyzed in the study. Graph 1a presents the equating functions corresponding to the two equating chains described above. Graph 1b presents the difference between these two functions (hereafter "the difference function" - DF). The main results of the study are presented in Graph 2 and Graph 3. Graph 2a presents 12 DF's, each corresponding to one of the TL-1 PET forms used in the study. Graph 2b summarizes these DF's and gives the mean (and sd) of the absolute DF functions presented in 2a. Graph 3a presents 9 DF's, each corresponding to one of the TL-2 PET forms used in the study. The mean and SD functions of the absolute DF's presented in 3a are given in 3b. The average difference between the equating functions in the two equating chains in each form clearly shows that the equating plan used in this study, in which the cross-lingual links are dominant, suffers from instability. As can be seen in graphs 2b and 3b, the average difference in equating two tests sections in two languages via two different links is about 1 to 2 raw score points in TL-1 and about 0.5 to 0.8 raw score points in TL-2.

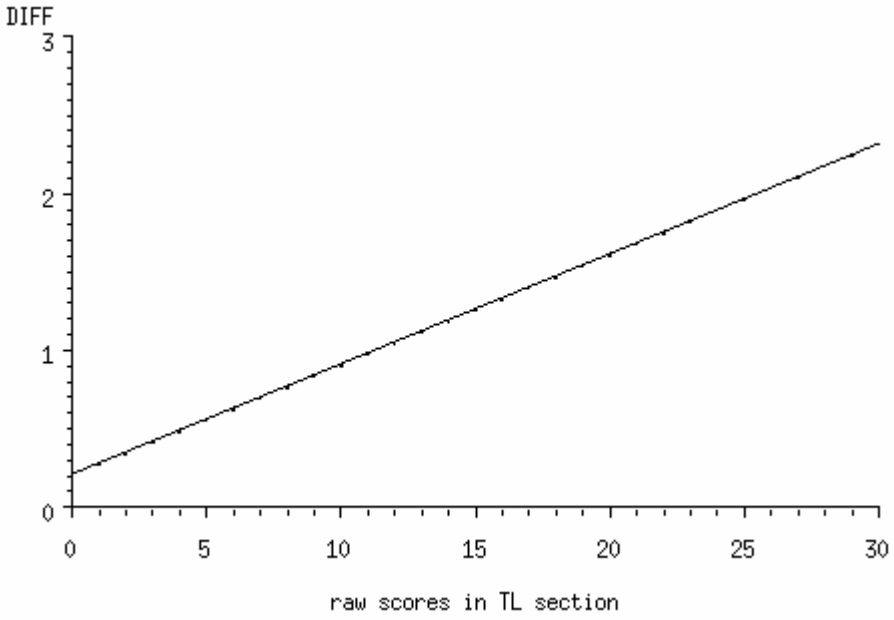
Graph 1 (1a and 1b):

Two equating functions for linking sections V1 - TL-1 and V2 - SL (graph 1a) and the difference between them (graph 1b) in a typical PET form.

1a:

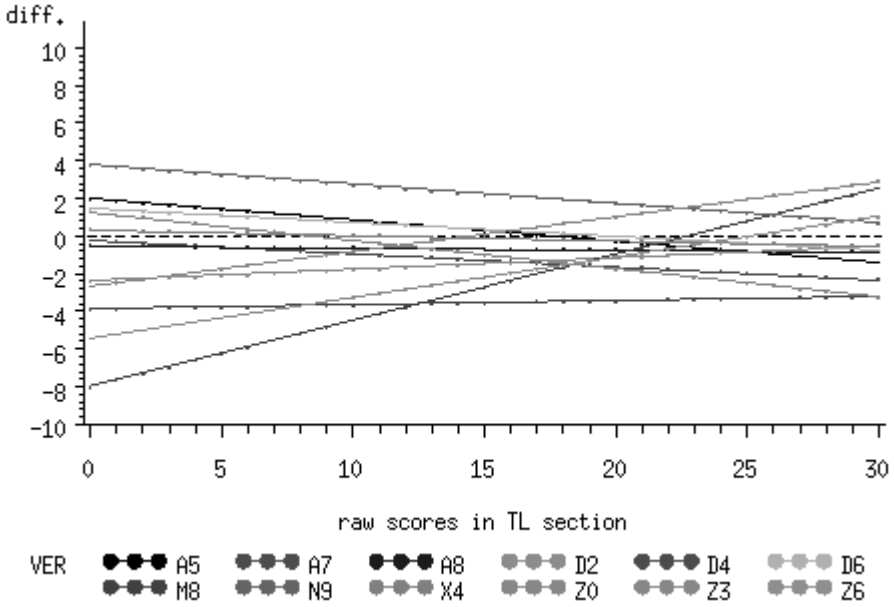


1b:

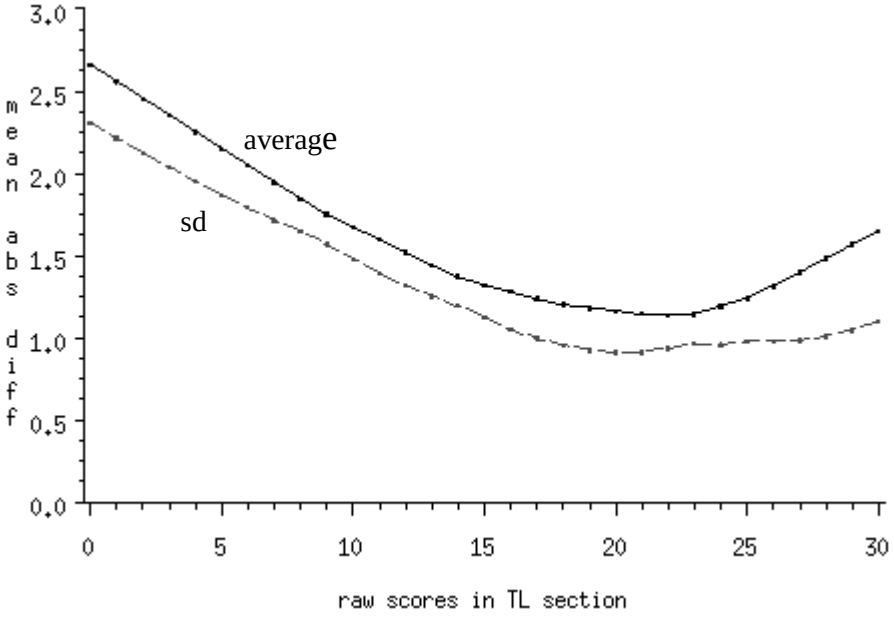


Graph 2 (2a and 2b):
Difference functions of twelve SL and TL-1 PET forms (graph 2a) and the mean and sd of the absolute values of these functions (graph 2b).

2a:

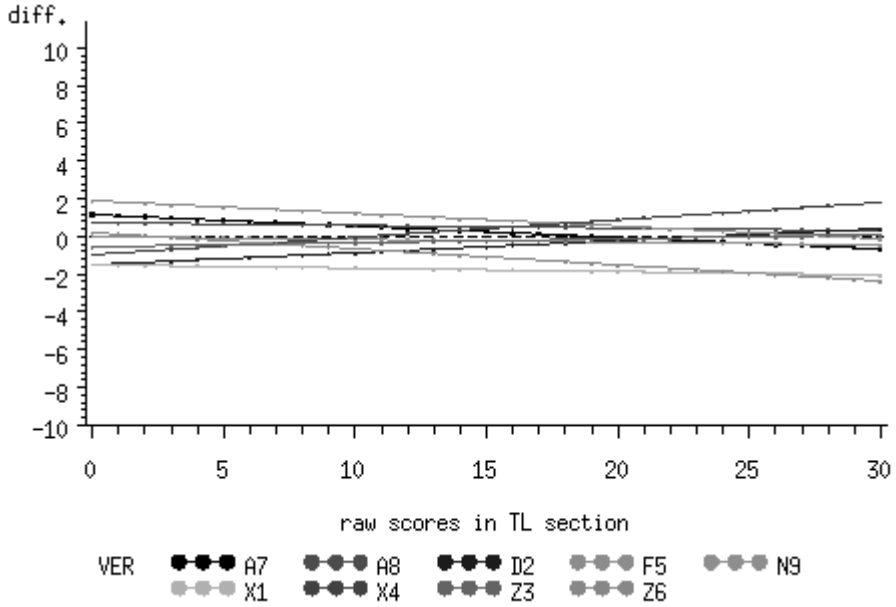


2b:



Graph 3 (3a and 3b):
Difference functions of nine SL and TL-2 PET forms (graph 3a) and the mean and sd of the absolute value of these functions (graph 3b).

3a:



3b:

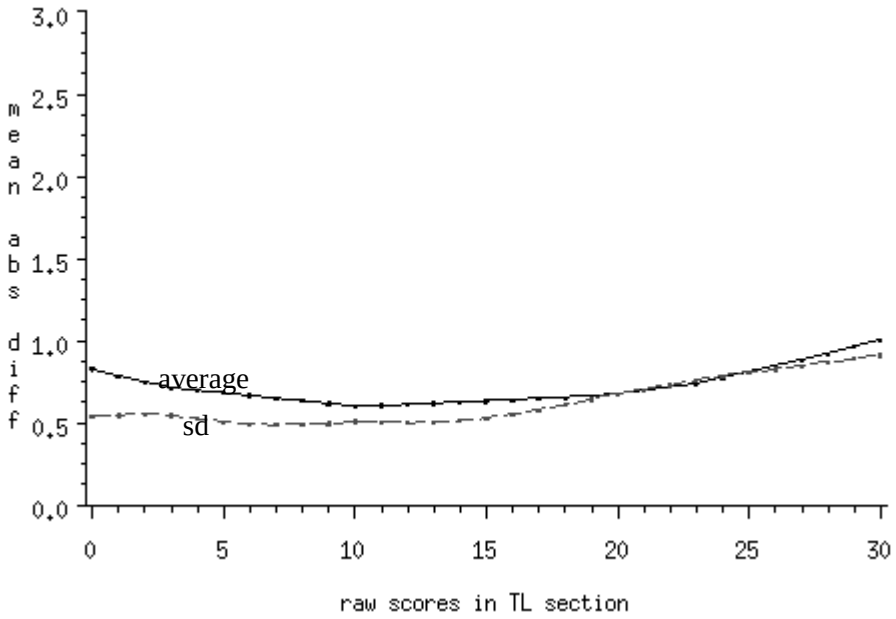


Table 3:
Absolute Differences Between Equating Functions in TL-1 at Specific Scores

PET form	Abs. Diff. at raw score 7 (Guessing score)	Abs. Diff. at raw score 18 (Middle range score)	Abs. Diff. at raw score 30 (Highest raw score)
X4	0.2	0.2	0.5
A5	1.2	0.0	1.4
A7	0.7	1.5	2.3
A8	0.6	0.7	0.8
D2	2.0	1.3	0.5
D4	3.7	3.5	3.2
Z6	3.9	1.6	1.0
D6	1.0	0.2	0.7
N9	3.0	1.9	0.7
Z3	1.4	0.7	2.2
Z0	0.2	1.4	3.2
M8	5.5	1.7	2.6
Av.	**1.9**	**1.2**	**1.7**

Table 4:
Absolute Differences Between Equating Functions in TL-2 at Specific Scores

PET form	Abs. Diff. at raw score 7 (Guessing score)	Abs. Diff. at raw score 18 (Middle range score)	Abs. Diff. at raw score 30 (Highest raw score)
X4	0.3	0.7	1.8
X1	1.6	1.8	2
A7	0.8	0.1	0.6
A8	0.6	0.5	0.3
D2	1.0	0.4	0.4
Z6	0.4	1.3	2.4
N9	0.1	0.2	0.4
Z3	0.4	0.2	0
F5	1.4	0.7	0.2
Av.	**0.7**	**0.7**	**1**

The absolute difference between the two equating conversion lines at specific raw score points is presented in Tables 3 and 4 (for Tl-1 and TL-2 respectively). The first score point represents the guessing level score, the third score represents the highest score and the second score is the middle range score in the two languages. Clearly, the difference in equating via the two linking chains is somewhat larger for TL-1 (about 1.9 raw score points at the lower end of the scale, 1.2 point at the medial score and 1.7 points at the highest end) than for TL-2, where differences are roughly half in size (about 0.7 raw scores points at the lower end of the scale and at the middle and about 1 point at the highest end). For comparison, the estimated standard error of cross-lingual equating of PET is about 0.1 raw scores at the medial score and about 0.2 at the ends.

Discussion

This study attempted to examine the instability inherent in the cross-lingual equating process adopted by a large-scale testing system, in which target languages forms are equated to source language forms using a set of translated items. In particular, we were interested in evaluating the degree of error inherent in the routine cross-lingual equating of the Verbal Reasoning sub-test of PET, a high stakes scholastic aptitude test that is adapted from Hebrew into 5 different target languages.

A research plan inspired by the "double-linking" method but modified to fit the special context of cross-lingual testing was used. According to this plan, the error in cross-lingual equating of the verbal subtest of PET was estimated by the difference between two equating conversion functions, each linking the SL verbal section and a parallel verbal section in TL. In principle, if the equating process in those two links were free of error, the equating relationship resulting from the two links could be expected to be similar. The average difference found by the same method over a number of test forms would reflect the degree of overall instability that exists in the cross-lingual equating process.

The findings indicated that the differences between the conversion functions in the two alternative links were high. On average, the difference between the conversions was between 1 to 2 raw score points in the equating process of one of the target languages to the source language and between 0.5 to 1 raw score points in the equating process of the other target language to the source language. Obviously, these differences were caused by a real and systematic problem that underlies the cross-lingual equating process. The problem is greater in the first target language than in the second.

What sort of problem could create such error in equating across two parallel language forms of the same test? We can name a few important factors that could interfere in the cross-lingual equating comparisons and introduce error: (1) the distortion of item content due to the translation process, which may also change the statistical and psychometrical properties of items and of a whole test section; (2) the composition of the anchor item set, which is not properly representative of the various item types; (3) the difference in the ability level and other relevant characteristics between the examinee groups participating in the process; and (4) differential performance in the two Verbal sections depending on the section position along the test. This "location effect" can be of different nature in the language groups due to cultural factors, differences in motivation, level of training and so on.

Aside from the inevitable and uncontrollable problem of content distortion that stem from the translation process, the other, more methodological problems, can be controlled to a certain extent, especially the use of a non-representative anchor and the attempt to equate groups that are too dissimilar in ability. As to the last mentioned problem of differential location effect on performance in the two language groups, this problem is more related to the within-language equating process but it might have an indirect influence on the findings of the present study because the equating links that were compared each included a within-language equating link as well as a cross-lingual equating link. Rapp, Pinku and Allalouf (in press) reported that there is a high probability that a differential location effect is occuring in the TL groups used in the current study. If an examinee group performed differentially on the first and second sections of the Verbal Reasoning subtest in a given PET form, this may have contributed to the difference in equating found in this study. Consequently, this

difference may have wrongly been attributed to problems in the cross-lingual equating processes rather than in the intra-language equating. It is likely that part of the differences found in the current study were caused by problems of instability in the group performance on the test and not by problem in the cross-lingual linking. In fact, the overall instability of the cross-lingual process may actually be smaller than that found in the present study, and in reality, may not be so bad after all. In order to answer these questions properly and to be more confident about the present suggestion, more controlled research is needed to analyze the performance of the different language groups and the relationship between it and factors such as the location of a section, cultural factors, etc. It would be interesting to try to estimate the degree of influence of such effects on the cross-lingual equating difference.

It is important, however, to note that some of the problems mentioned are typical of the PET testing conditions but do not necessarily exist in other testing systems, especially the problem regarding the differences in examinee characteristics between the language groups (difference in level of performance, in location effect on performance etc…). The study should be replicated using a wider empirical base, other cognitive tests, studying other languages in other cultures and countries, etc., especially in conditions where smaller differences exist between examinees in the different language groups.

For the moment, the importance of this research lies in presenting a new evaluative tool for testing a cross-lingual equating process. It can be used, as in the present study, to evaluate a long-term equating process but also for an online routine evaluation of cross-lingual equating. Once a standard of error in cross-lingual equating is established, one can measure the difference between two cross-lingual links and compare it to the expected error. This process can be integrated into the cross-lingual equating procedure.

Another extension of the basic idea used in the study is to improve the cross-lingual equating plan by simultaneously double-equating each new TL form. More specifically, the aim is to create a linking plan in which forms are equated both to other forms in the same language and to forms in the source language. In this “double linking” method, every new TL form would be linked both across-languages using a

method similar to the method used in the present study, and within-language to an older form in the same language (using a single-group method or using common items). These within-language links between forms can be considered relatively error-free, as well as free of translation problems. The within-language equating outcome can be used to evaluate and reaffirm the cross-lingual equating outcome or be used in conjunction with it. It is suggested that such a plan could provide greater equating stability relative to the regular cross-lingual plan, in which every new TL form is linked only through a cross-lingual link to the source version of the form.

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.

Berk, R.A. (Ed.) (1982). Handbook of methods for detecting item bias. The John Hopkins University Press.

Budesco, 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 non-representative 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.

Hambleton, R. K. (1994). Guidelines for adapting educational and psychological tests: A progress report. European Journal of Psychological Assessment, 10, 229-244.

Hambleton, R. K., & Patsula, L. (1998). Adapting tests for use in multiple languages and cultures. Social Indicators Research, 45, 153-171.

Holland , P.W., & Wainer, H. (Eds.) (1993). Differential Item Functioning . Hillsdale, NJ: Erlbaum.

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.

Rapp, J., & Allalouf, A. (2001), Cross-lingual equating of a Verbal Subtest Using a Quantitative Subtest. NITE's report (in press).

Rapp, J., Pinku, G. & Allalouf, A. (2002), Testing the Population Equating Independence Property in the Cross-lingual Context. NITE's report (in press).

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.

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.

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.

Woodcok, R.W., & Munoz-Sandoval, A.F. (1993). An IRT approach to cross-language test equating and interpretation. European Journal of Psychological Assessment, 3, 1-16.