Equating Translated Verbal Test Forms Using Multiple Channels

Avi Allalouf and Joel Rapp
National Institute for Testing and Evaluation (NITE),Jerusalem, Israel

Paper presented at the annual meeting of the
National Council on Measurement in Education
as part of an NCME symposium on test adaptations
New Orleans, LA, April 2002

Abstract

For a growing number of test translations, there is a need for equating which provides scores that can be used interchangeably for both source- and target- language forms. The basic equating requirements (common items and/or common) cannot usually be met in the cross-lingual case and especially in verbal tests. In this study, two equating channels, in addition to the popular common identical items method that uses only the non-DIF translated items in the anchor (***Channel 1***) were proposed: using non-verbal translated items as an (external) anchor (***Channel 2***), and using an internal within-language equating channel (***Channel 3***). All three equating channels were described in detail, and we elaborated on the experience acquired and the studies conducted at the Israeli National Institute for Testing and Evaluation (NITE). The three channels were evaluated by a number of criteria: theoretical background, stability of the equating process, accuracy in measuring the ability and drift over time. Two possibilities for implementation are suggested: (**a**) to continue to use the verbal anchor, as it has a stronger theoretical basis, and to use the two other channels for monitoring purposes, and (**b**) to weight the outcomes of all three methods according to an estimation of the appropriateness of each of the methods.

Equating Translated Verbal Test Forms
Using Multiple Channels

In the growing field of test translations, there is an increased need for equating which provides scores that can be used interchangeably for both source- and target- language forms. In any equating between two test forms at least one of two elements is required: (a) common identical items (an anchor) on the two forms that represent the forms, and (b) common examinees on the two forms. However, in the cross-lingual case, and especially in verbal tests, these requirements usually cannot be met, due to the following: (a) the translated items are not identical to the source language items because translation has an impact on item content, and, as a result, on the psychometric characteristics of the translated items, and (b) no examinees can be considered to have the same ability in both languages, which means that there are no "common examinees."

In some cases, often because equating is problematic, the linkage between the source- and translated- test forms is accomplished by relying on a specific assumption - which may be inaccurate - regarding the relation between population abilities or form difficulties (for example, relying on an assumption that two language groups have the same ability distribution). In many other cases the solution selected is to choose from *the common translated items* those items with no DIF and to use them as an anchor (after applying statistical methods for detecting those items which display DIF between the source- and target- language forms). While this procedure (labeled here as ***Channel 1***) does not make the anchor items identical in content, it maximizes the psychometric similarity of the common items. This solution is still far from perfect because there may be a systematic drift in the difficulty of the translated items following translation, which cannot be revealed by DIF methods. In addition, the anchor (what remains after the removal of the DIF items) may not represent the test well enough.

Results of studies on the accuracy of *Channel 1* procedure (which will be discussed further later in this paper) show that the accuracy is far less than that obtained in mono-lingual equating. This, in conjunction with the aspiration to achieve a more satisfactory equating method for the cross-lingual case, led to the development and study of two additional equating channels, labeled here as ***Channel 2*** and ***Channel 3***:

Channel 2: *Using common non-verbal translated items as an anchor (*instead of using the common verbal items).

Channel 3: *Using an internal, within-language equating channel*, in which every new translated form is equated to an already equated translated form.

This paper will elaborate on the experience acquired by and the studies performed at the Israeli National Institute for Testing and Evaluation (NITE) regarding all three channels for equating translated verbal tests to the source language tests. Also, the issue of assigning weights to each of the three channels will be discussed. Each equating channel will be described in detail and then evaluated by two means: (a) a general evaluation of the method, and (b) results of studies conducted at NITE regarding the equating method.

All the NITE studies were carried out on the Psychometric Entrance Test (PET) used in the admissions to universities in Israel. This is a high-stakes, multiple-choice test, composed of three subtests (or domains): verbal reasoning (V), quantitative reasoning (Q) and English as a foreign language (E). Each subtest has two parallel test sections. The verbal and quantitative subtests are similar in many aspects to the Scholastic Assessment Tests (SAT I) verbal and mathematical sections. PET is translated and adapted into five languages: Arabic, Russian, Spanish, French and Hebrew & English combined form. The test is administered to most of the candidates applying to institutions of higher education in Israel (see Beller, Gafni and Hanani, 2000, for further details).

Equating or Linking?

The fact that general equating assumptions do not hold perfectly for cross-lingual equating has led some equating specialists to label the relationship between two language forms as *linking* rather than *equating*. A typical justification for using the term linking appears in Brennan (2001):

> I use the word "equating" to refer to a statistical relationship between scores on forms of a test constructed according to the same content and statistical specifications and administered under the same conditions. By contrast, when any of these conditions are not fulfilled, I use the term "linking" (p. 10).

However, this paper, which expresses the need for achieving a common scale for both language test forms, uses the term *equating* (even though this usage may be seen as expressing an aspiration rather than an empirical relationship).

Equating Translated Verbal Tests - ***Channel 1***

Equating through non-DIF translated verbal items

1.1 Description of the equating method

The popular method for cross-lingual equating applies the separate monolingual groups design (Hambleton, 1994; Sireci, 1997) in which source-language (SL) and target-language (TL) versions of a test are separately administered to source- and target- language examinee groups. Then, a set of translated items (an anchor), considered to be equivalent across the source- and target-language versions, is used to link the different languages tests onto a common score scale. Basically, this equating design is similar to the familiar "common-item non-equivalent groups" equating design (Angoff, 1971; Kolen & Brennan, 1995), except that in the present context, the anchor set consists of translated items.

Because the translated items usually are not completely identical to the source language items, a DIF analysis is usually performed in order to discard the most aberrant items from the common item set so as to reduce the item-by-language interaction. Because DIF in translated verbal items is usually large and easy to detect, the exact DIF detection method used is not critical.

After the anchor is set, one of the common-item non-equivalent groups equating methods can be applied. It can be an equipercentile method or a linear equating procedure such as the Tucker method (Tucker, 1951), Levine observed score or Levine true score, (Levine,1955), all based on classical test theory. Equating based on item response methods is also feasible.

1.2 General evaluation of the method

The separate monolingual groups procedure suffers from a theoretical flaw because it is practically impossible to ensure that the different language versions are fully equivalent and that they measure the same construct - a basic requirement for any equating system. Of course, this problem characterizes in particular tests in which the verbal aspect is critical. Specifically, it cannot be ascertained that the common (translated) items used for the equating continue to maintain the same content and psychometric characteristics following translation (see Angoff & Cook, 1988; Allalouf, Hambleton & Sireci, 1999). If there is a systematic drift in the anchor item difficulty following translation, the equating will result in over- or under-estimation of the target language examinees' ability due to the systematic inaccuracy in estimating the equating relationship that exists between the test forms in the two languages.

In addition, two additional factors may induce equating error in the *Channel 1* method: (**a**) the anchors usually do not represent the test forms in each language well enough; this is because there are items in the translated forms that are constructed especially for these forms and do not appear in the source language forms. In addition, the non-DIF items that serve as the anchor are not random samples of the whole test, and (**b**) in many cases ability differences exist between the groups of examinees taking the alternate test forms in the source and target languages. These ability differences induce more error in the equating results, and the set of common items, however appropriate, is not likely to overcome this problem (Angoff & Cook, 1988; Kolen & Brennan, 1995).

1.3 Results of studies conducted at NITE regarding the equating method

Several studies were conducted to evaluate the common translated verbal items equating method, which is currently used in equating the translated forms of PET. In a recent study, Rapp and Allalouf (2002) proposed a cross-lingual evaluative tool to estimate the degree of inconsistency in the procedure used to equate the verbal test. A research plan that applied a double- linking design modified to fit the special context of cross-lingual testing was developed. According to this plan, the error in cross-lingual equating of the verbal subtest of PET would be estimated by the difference between two equating conversion functions, each linking a source language verbal section and a parallel verbal section in the target language. The average difference found by this method over a number of test forms would reflect the degree of overall instability that existed in the cross-lingual equating process. The same study design was to be carried out on Hebrew (source) test sections and test sections in two target languages.

The findings indicated that the differences between the conversion functions in the two alternative links were not small. On average, the difference between the conversions was about 1.5 raw score points (for a test that consists of 30 items) in the equating process of one of the target languages to the source language and about 0.75 raw score points in the equating process of the other target language to the source language (these values represent about 0.3 standard deviations for the first language, and 0.15 standard deviations for the second language). These errors, although not large, are greater than the standard error of equating estimated for the Hebrew equating process which is no more than 0.06 standard deviations (Rapp & Allalouf, 1999).

In summary, equating through common translated items is not sufficiently stable. Alternative equating channels are needed either for monitoring the equating results or for using in addition to the common translated items method.

Equating Translated Verbal Tests - ***Channel 2***

Equating through non-verbal translated items

2.1 Description of the equating method

The second channel applies the same equating method as *Channel 1* except that here the anchor is not a set of common verbal translated items. The second equating channel signifies an attempt to avoid the impact of translation difficulties on equating by using as anchors only items in which the linguistic aspect is less important or not important at all, such as mathematical items or visual items. These items serve to equate between the source- and the target- verbal tests. The items are used as an external anchor, in contrast to *Channel 1* where the verbal anchor is an internal anchor (i.e., the anchor items are part of the items used in scoring).

Angoff and Cook (1988), who developed and examined the equating process employed between the SAT and the Spanish PAA (Prueba de Aptitud Academica), found that mathematics items tend to preserve their psychometric characteristics across languages more than verbal items do. Similar findings were presented by Gafni and Melamed (1991) and Allalouf (1999). These studies suggest that mathematics is a more universal language. Hence, it seems that equating verbal tests across language versions using non-verbal, rather than verbal, items as an anchor, would achieve more stability than using the verbal translated items as an anchor. In addition, in the case of PET, there would be more common items available for the equating procedure and it is likely that the reliability of the quantitative common item set would be higher than that of the verbal common item set.

2.2 General evaluation of the method

The theoretical justification for using quantitative items as anchor items to equate the verbal sub-test is questionable, especially because quantitative items are not representative of the verbal domain. Such usage violates basic assumptions needed for a proper implementation of any equating method. For example, in order to equate using the Levine observed score method it must be assumed that (a) the two tests equated and the common item set measure the same property, (b) the scores of the common item set correlate highly with the scores in the two tests equated, and (c) the linear regression of the scores in the anchor item set (the quantitative sections) on the test scores (the verbal section scores) is the same for both populations to whom the tests were administered. Obviously, the first assumption does not hold because the anchor is not measuring the same ability as the test. The second assumption may hold, as the correlation between verbal and quantitative scores is about 0.7 (Donlon, 1984, for the SAT; Beller, 1994 for

the PET) and that is a reasonable correlation between an external anchor and a test. As for the third assumption, in the present case, one has to assume that the relationship between verbal and quantitative abilities is independent of language or cultural background (this seems to be impossible to prove). For the purpose of the study, we will assume that using the quantitative sections as anchors is appropriate and will result in an accurate equating relationship.

2.3 Results of studies conducted at NITE regarding this equating method

A study by Rapp and Allalouf (2001) explored the equating design by applying it to data collected from various PET forms. The study compared the use of a verbal anchor to the use of non-verbal anchor and will be described here in detail. The PET forms chosen were all administered in the source language and in one of two target languages (TL-1, 10 forms, and TL-2, 8 forms) in the years 1997-2000. In the study, verbal sections in a given target language were equated to their respective source-language in sections using a 25-item translated quantitative PET section as anchor. The Levine linear observed score equating method with an external anchor was used for the equating procedure. The first verbal section (V1) in a given PET form in a target language was equated to the second verbal section (V2) in the source language version of the same form via two different equating chains, and the results from the two equating chains were compared. The research plan is displayed in Figure 1.

Figure 1 - Equating between V1-TL and V2 -SL via two equating chains using the quantitative sections as anchors.

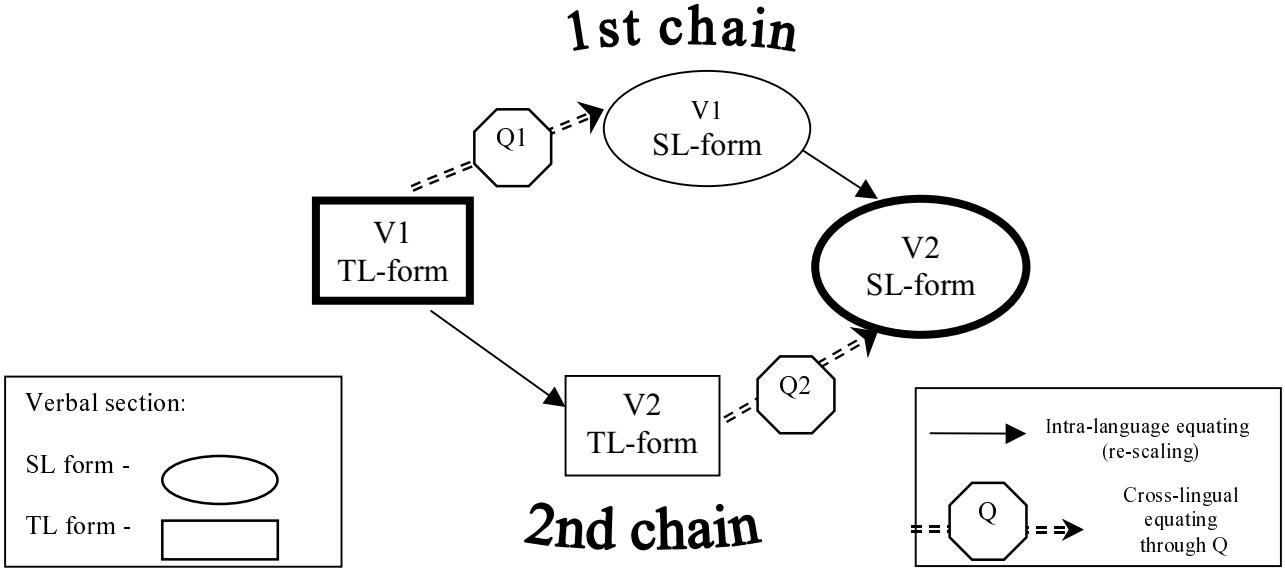


Figure 2 displays the results obtained for one of the PET forms explored. The two higher functions in the figure are the equating conversion lines between V1 in TL and V2 in SL using the quantitative sections in that form (Q1 or Q2) as anchor sets. The two lower functions plotted on the figures are the conversion lines obtained using the common translated verbal

items internal anchor (about 15 non-DIF items) as calculated in Rapp and Allalouf (2001). This typical case indicates that the equating conversion functions via the two quantitative links fall very close to each other, while the two equating functions via the verbal anchors do not. The other central finding is that the two quantitative conversion functions clearly result in higher scores than the scores obtained using the verbal item anchors.

Figure 2: Functions of equating sections V1 in TL and V2 in SL via quantitative sections as anchors and via verbal anchor sets in a typical PET form.

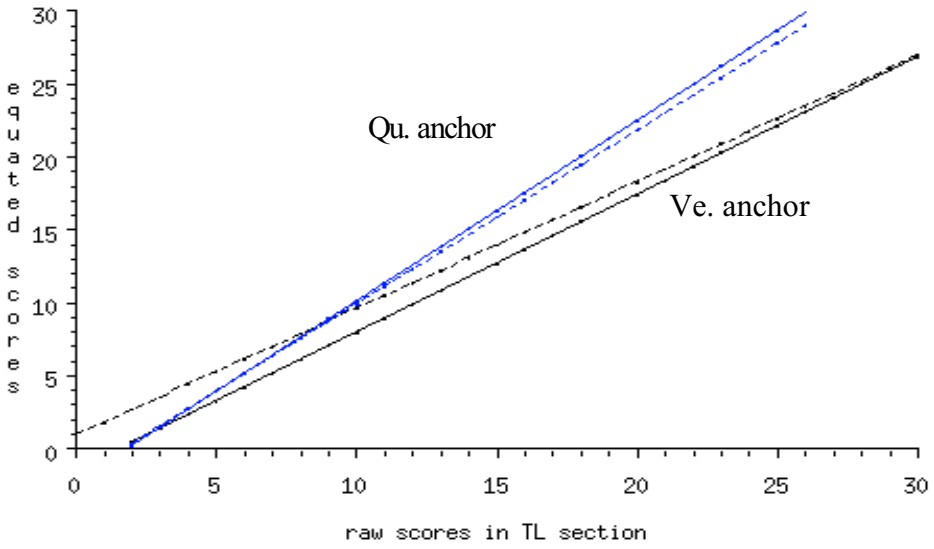
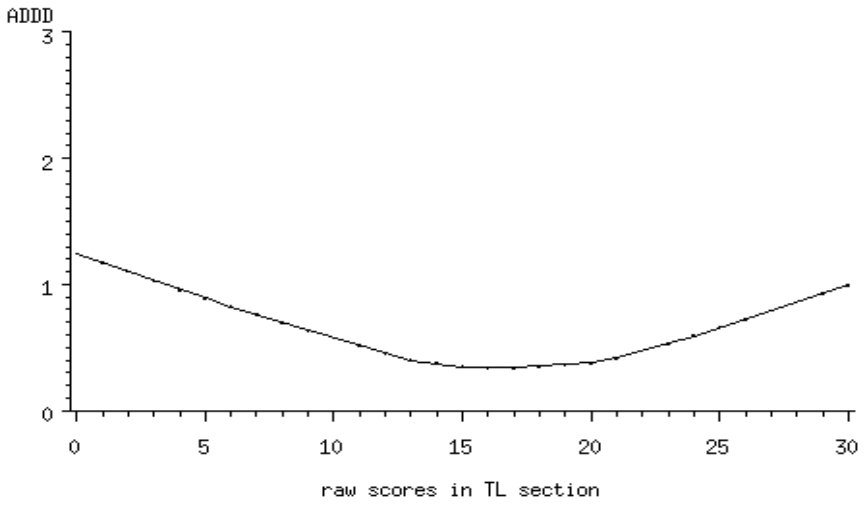


Figure 3 summarizes the differences (which are estimation of the equating error) between the two chains of equating using quantitative sections as anchors, expressing the difference between the two upper functions in Figure 2. The figure shows the mean absolute difference in 10 PET forms in TL-1in raw score points. Smaller differences were found in the second language. The differences represent 0.10 to 0.15 standard deviations of the scores.

Figure 3: Mean of absolute difference between two equating chains that use a quantitative anchor, for 10 PET forms in TL-1 and SL.

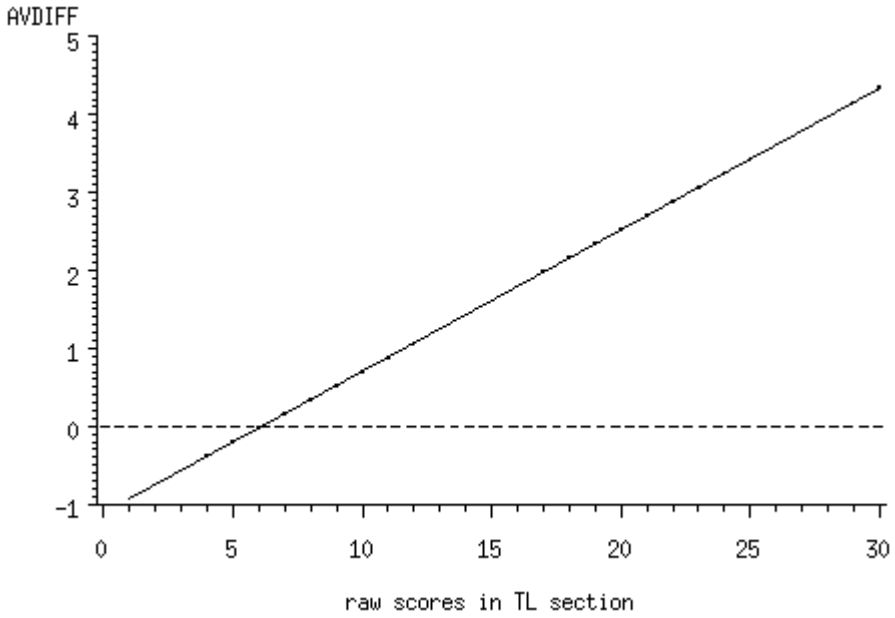


In both TL-1 and TL-2, the mean difference function between the two equating chains is quite small, although in TL-2 the difference tends to be smaller. The range of difference between the equating results of the two quantitative anchor chains is about half of that found between the same two equating chains when verbal instead of quantitative anchors are used (see Rapp & Allalouf, 2002). The small differences can also be attributed to a longer quantitative anchor (25 items) in comparison to the verbal anchor (about 15 items).

Figure 4 summarizes the differences found between using a quantitative section as an anchor and using a set of translated verbal items as an anchor for several PET forms. The figure shows the mean difference in 10 PET forms in TL-1. The average difference is about 2 raw score points around the middle raw score and about 4.5 raw scores points at the higher end. The differences were smaller in TL-2.

In summary, it was found that the equating results using the non-verbal anchor were more stable than the equating results using the verbal anchor. However, using the non-verbal anchor produced higher scores.

Figure 4: Mean difference in 10 PET forms between equating V1- TL1 to V2-SL by a quantitative and by a verbal anchor set.



Equating Translated Verbal Tests - ***Channel 3***

Internal- within-language equating

3.1 Description of the equating method

The internal within-language equating channel is, for practical purposes, the single group linear equating method where the same examinees are tested on both test form X and test form Y. In our terminology, the same examinees (in the target language) take both the operational new test, which needs to be equated, and an anchor test, which has already been equated. Here, in contrast to the two previous channels, there is no routine equating between the source- and the target- language forms. The method is applicable and relevant when there are several or many translated forms that can be equated to another. It is not relevant when there is only one translated form.

A linear equating method is used. It assumes that the standard-score deviate for any given scaled score in the operational test equals the standard-score deviate for any given scaled score in the anchor test (Angoff, 1984).

3.2. General evaluation of the method

The internal within-language equating method used for the translated forms has the advantage of simplicity. No complicated theoretical assumptions are needed, and the relation between the two forms is straightforward (mean and standard deviation transformation). The key here is the first linkage of the translated form to the "already equated" translated form. The already equated form must be equated via *channel 1* and/or *channel 2.* This equating should be done carefully and accurately because it will serve for several translated forms. Evaluating this method should take into account that using the internal within-language method continuously (without additional equating to the source language) may induce a systematic drift in the translated forms scale from the standard scale over time.

3.3. Results of studies conducted at NITE regarding the equating method

Within-language equating is the equating design currently used for the Hebrew forms of PET, so quite a few studies have accumulated on various aspects of the method. However, these studies were not conducted on the translated forms, so generalization to our case is restricted. Stoller and Allalouf (1996) estimated the size of the equating error in each domain

of PET. They made use of a fixed test section for each domain, which was administered to samples of about 300 examinees who took PET during the years 1993-1995. An analysis showed that these fixed sections were stable over time; they could, therefore, serve as a tool for detecting equating errors and possible changes in equating errors over time. Two types of estimated scores were obtained for the two test sections in each domain using two designs: (a) the design used in the regular equating of PET, and (b) a design using the fixed section as the anchor section. The differences between the two types of estimated scores were analyzed. The results of the analysis showed that the mean sizes of the equating errors ranged between 1 and 1.4 points (on a standard scale ranging between 50 to 150 with a standard deviation of about 20). This is equivalent to about 0.06 of the scores' standard deviation. In a recent study, Rapp and Allalouf (1999) found similar results.

In summary, the anticipated equating error in the internal within-language equating is very small. The crucial point here is the necessity for a high level of confidence regarding the "already equated" translated form, to which the other translated forms are equated.

Discussion

For a growing number of test translations, there is a need for equating which provides scores that can be used interchangeably for both source- and target- language forms. The basic equating requirements (common items that represent the forms to be used in the anchor, and/or common examinees having the same ability in the two languages) cannot usually be met in the cross-lingual case. The situation is more problematic in verbal tests, where translation has more impact on item content and, as a result, on the psychometric characteristics of the translated items.

However, despite these problems, the need for accurate cross-lingual equating still exists. In this study, two equating channels, in addition to the popular common identical items method that uses only the non-DIF translated items in the anchor (***Channel 1***) were proposed: using non-verbal translated items as an (external) anchor (***Channel 2***), and using an internal within-language equating channel (***Channel 3***). All three equating channels were described in detail, and we elaborated on the experience acquired and the studies conducted at the Israeli National Institute for Testing and Evaluation (NITE) regarding all three equating channels for equating translated verbal tests. Each of the three channels that were examined has advantages and disadvantages. The two main criteria for comparing the three channels can be called the reliability and validity of the cross-lingual equating.

Reliability of equating means that the equating is consistent and stable; that if we repeat the equating process, the results will be very similar. Here, in the cross-lingual equating of verbal tests, the results of *Channel 1* equating process were found to be not sufficiently stable: the estimated equating error ranged between 0.15 to 0.3 standard deviations of the scores. Equating via a non-verbal anchor (in this case, a quantitative anchor) resulted in much more stable results (the estimated equating error was 0.10 – 0.15 SD). *Channel 3*, inter-language equating, showed the smallest equating error (0.06 SD). It should be noted that the latter value was estimated for inter-language equating when the language was Hebrew, so it may be somewhat different for other languages.

Validity of equating means that the outcome of equating, the scores computed after the conversion of raw scores to standard scores, are an accurate estimation of examinee ability or achievement. Here, in the cross-lingual case, it means that scores of the source language form and those of the translated language form are on the same scale, that is, ideally, no examinee is under- or over- scored because of the language he was tested in. Regarding the three equating channels studied, we found a large discrepancy between *Channel 1* and *Channel 2*. The scores that were computed based on the non-verbal anchor were higher. Applying a

non-verbal anchor in cross-lingual equating must be based on a strong assumption that the relationship between verbal and non-verbal abilities is independent of language examinee groups. The validity of equating through *Channel 3* depends on the accuracy of the first link between the translated form and the source language form. If this link was accurate, then the intra-language equating will be valid (except for the concern regarding possible drift over time).

Table 1 evaluates the three channels by a number of criteria, including reliability and validity as discussed above.

Table 1 – Comparing the three equating channels

	Channel 1 *Common translated verbal items as an anchor*	***Channel 2*** *Non-verbal translated items as an anchor*	***Channel 3*** *Internal within language equating*
Criteria			
1. Theoretical basis	Intermediate +	Intermediate	Strong
2. Stability (Reliability)	Relatively low Equating error: 0.15 – 0.30 SD	Medium Equating error: 0.10 – 0.15 SD	Relatively high Equating error: 0.06 SD
3. Accuracy in measuring the ability (Validity)	High	Questionable	Depends on the first link
4. Drift over time	Not expected	Not expected	May happen

At the moment, it seems that there are two possibilities for implementation which would be appropriate in practice: (**a**) to continue to use the verbal anchor, as it has a stronger theoretical basis, and to use the two other channels for monitoring purposes, and (**b**) to weight the outcomes of all three methods according to an estimation of the appropriateness of each of the methods. Choosing the second alternative requires further studies which would be dedicated to estimating these weights and would require an equating design that allows for application of all the three channels. This challenge should serve as a basis for additional study in the near future.

This paper did not deal with issues of translations and adaptations of tests. We should always keep in mind that some improvement in measuring and equating can be attained by improving the translation techniques and by carefully controlling the translation process (see Hambleton & Patsula, 1998).

In summary, all equating specialists (see, for example, Kolen & Brennan, 1995) agree that multiple equating channels are always welcome. In some cases, some of the channels can serve as quality control for the main channels; in other cases, the channels should be used together according to various weights. In any equating situation, and especially in the cross-lingual equating of translated verbal tests, relying on more than one channel, so that one can enjoy the advantages of all channels while trying to avoid their shortcomings, is highly recommended.

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. (1971). *The College Board admissions testing program*. New York: College Entrance Examination Program.

Angoff, W. (1984). *Scales, norms and equivalent scores*. Princeton, NJ: Educational Testing Service.

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.

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. & 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.

Brennan, R. L., (2001). Some problems, pitfalls, and paradoxes in educational measurement. *Educational Measurement: Issues and Practice, 20,* (4), 6-18.

Donlon, T. F. (Ed.) (1984*). The College Board technical handbook for the Scholastic Aptitude Test and Achievement Tests.* New York: College Entrance Examination Board.

Gafni, N. & Melamed, E. (1991). *Equating different language versions of a psychometric test*. Research Report No.148, Jerusalem: NITE.

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

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

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

Levine, R. S. (1955). *Equating the score scales of alternate forms administered to samples of different ability*. (RB-55-23). Princeton NJ: ETS.

Rapp, J., & Allalouf, A. (1999). *Using a double linking plan for equating the PET*. Research Report No. 262. Jerusalem: NITE.

Rapp, J., & Allalouf, A. (2001, July). *Cross-lingual equating of a verbal sub-test using a quantitative sub-test.* Paper presented at the 3rd meeting of NITE's scientific council. Tel-Aviv.

Rapp, J., & Allalouf, A. (April ,2002). *Evaluating Cross-Lingual Equating.* Paper presented at the Annual Meeting of the American Educational Research Association, New-Orleans, LA

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

Stoller, R., Allalouf, A. (1996). *Evaluation of the PET equating by fix sections*. Research Report No. 221. Jerusalem: NITE.

Tucker, L. R. (1953). Scales minimizing the importance of reference groups. In: *Proceedings of the 1952 Invitational Conference of Testing Problems.* Princeton, NJ: ETS, 22-28.

* * * * *

* * *

*