Linear and Equipercentile Methods for Equating PET

A comparative study of the linear and equipercentile methods for equating the Hebrew PET forms

Joel Rapp

National Institute for Testing and Evaluation

December 1999

Abstract

The present study was designed to determine whether there was any justification for replacing the linear equating method currently used for the Hebrew version of the Inter-University Psychometric Entrance Test (PET) with an equipercentile (curvilinear) equating method. Whereas a curvilinear equating function is more general and best represents form-to-form differences in difficulty, equipercentile equating is more complicated than is linear equating. In addition, the curvilinear method requires a larger sample in order to obtain the same range of random error obtained by linear equating.

This study is descriptive in nature and explores the equating relationships between different PET forms. Both the equating and the analysis are performed separately for each of PET's three test domains: Verbal Reasoning, Quantitative Reasoning and English as a foreign language. Data were collected from 19 Hebrew PET forms. For each form, both linear and equipercentile equating were performed on three pairs of sections within each test domain:
(1) the two operational sections, (2) the first operational section and the anchor section and (3) the second operational section and the anchor section. In the first pair, the equatings were based on all examinees who were administered the form, while in the other two pairs, equatings could be based only on the sample of examinees who were administered the anchor section.

In most cases, a convincing similarity was obtained between the linear and equipercentile equating functions. Differences between the two functions rarely exceeded one raw score point, and in most cases did not exceed 0.3 points. These results are quite close to the ranges of typical standard errors of equating. The average of the differences between the equipercentile and linear equating functions was fairly constant in the central range of the score scale and tended to increase towards the ends of the scale. Findings suggest that the differences between the two

functions were caused more by random factors than by typical non-linear relationships that presumably exist between sections.

A linear function thus appears to serve as an adequate estimate of the equating relationship between sections. Furthermore, as equipercentile method is also more complex and requires a larger equating sample, there is little justification, if any, for replacing the linear equating method with the equipercentile method in the Hebrew version of PET.

Introduction

Over the last 15 years, a linear method of equating has been used at the National Institute for Testing and Evaluation (NITE) to equate different forms of the Inter-University Psychometric Entrance Test (PET). To collect data for equating, a single group design is used with a sample of about 1,000 examinees. The linear method appears to be appropriate for equating PET forms because it requires a comparatively small sample of examinees and is relatively easy to perform. Moreover, studies conducted at NITE during this period, especially on the Hebrew PET form, have not indicated any fundamental problem with the equating method (see Allalouf, 1999 for a recent review). Nevertheless, it has been argued that the use of linear equating could produce erroneous results if the real form-to-form differences in difficulty are in fact curvilinear. For this reason, it was important to explore the nature of the equating relationship between various forms of PET, a topic that has never been investigated at NITE. The present study deals with this question and with the feasibility of replacing the linear method used at NITE for equating the Hebrew versions of PET with a curvilinear method. A separate study would be necessary to explore this same question with regard to the foreign language forms, since the process involved in equating the translated forms of PET to the Hebrew original is somewhat different (see Allalouf, 1999).

Equating is a process that converts scores on one form of a test to the score scale on another form. In equipercentile equating, a curve is used to describe form-to-form differences in difficulty, which makes equipercentile equating more general than linear equating. The equipercentile function allows for greater flexibility for differences in difficulty between forms than the linear function. For example, Form X could be more difficult than Form Y at high and low scores, but less difficult in the middle range.

The equating function is an equipercentile equating function if the distribution of scores on Form X converted to the Form Y scale is equal to the distribution of scores on Form Y. If a linear equating function is used when in fact, the relationship is non-linear, the linear function can be viewed as an estimate of the curvilinear relationship. This is because in linear equating, the first two moments (i.e. the mean and the standard deviation) of the score distributions in both forms are equated. By contrast, the equipercentile equating function is developed by identifying scores on one form that have the same percentile ranks as scores on another form. Hence, up to 100 distribution parameters (all percentile ranks) are involved in the estimation of the equating relationship, which results in a better estimation of it. Equipercentile equating is especially recommended over linear equating when the accuracy of the results is important along the entire score scale and not only around the mean of the scores. However, it should be noted that in order to maintain the same range of standard error of equating as in the linear equating process, a much larger equating sample is required. This is a major concern at NITE, as there is a limited pool of examinees available for equating.

Another disadvantage of the equipercentile method over the linear method is its complexity. It requires more complicated analyses, and presents its results in the form of a complex conversion table (or function). On the other hand, the method does not produce non-existing scores at the extremities of the score scale and obviates the need for performing a doglegging procedure.

Because the equipercentile equating method is complex and requires a large sample of examinees, its profitability is brought into question. While it is difficult to determine the fundamental equating relationships between PET forms, it can be determined whether or not equipercentile equating would result in outcomes that are radically different from those of the linear method. If there were to be significant differences in the outcomes of the two procedures, the option of replacing the linear equating method with a more complex one would have to be seriously considered.

In this exploratory study, we examine the differences between linear equating results and equipercentile results in pairs of PET sections. The comparison is performed separately in each of three test domains: Verbal Reasoning, Quantitative Reasoning and English as a foreign language. No statistical tests are used; instead, the main analytic tool consists of sets of graphs that describe the linear and equipercentile functions together, and corresponding graphs that describe the differences between the two functions along the score scale (hereafter: "Differences Function" -"DF"). Repetitive findings from a large sample of cases should create a sufficient empirical base to allow us to draw conclusions about the appropriate method for equating PET sections in each of the three test domains.

Method

Data

Data were collected from 19 consecutive Hebrew PET forms administered between 1997 and 1999. A PET form consists of two parallel operational sections in each of three domains and two additional sections[1]. According to the equating design, external anchor sections, one in each of the three test domains, are administered to random "equating samples" of approximately 1,000 examinees. An anchor (equating) section, is an operational section from a previously-administered form.

In each form and in each domain, linear and equipercentile equating functions were computed between three pairs of sections: (1) the two operational sections, (2) the first operational section and the equating section, and (3) the second operational section and the equating section. In the first pair, data were used from all examinees who were administered the form. In the other two pairs, only data from the equating sample could be used.

---

[1] In addition to the six operational sections, in each form there are two sections used for research, pre-testing or equating purposes. While all examinees in a specific form receive the same operational sections, the supplementary sections might differ among examinees.

Calculation of linear equating

The linear equation, $l_y(X)$, that converts observed scores on Section X to the scale of Section Y was calculated by

$$l_y(x) = \frac{\sigma(Y)}{\sigma(X)}x + \left[\mu(Y) - \frac{\sigma(Y)}{\sigma(X)}\mu(X)\right] \quad . \tag{1}$$

This expression has a linear structure with

$$slope = \frac{\sigma(Y)}{\sigma(X)} \quad \text{and} \quad intercept = \mu(Y) - \frac{\sigma(Y)}{\sigma(X)}\mu(X) \, . \tag{2}$$

Basically, it was calculated using the routine procedure followed at NITE (see Allalouf, 1999 for details) but with certain differences:

(1) In the normal procedure, this linear function serves as the basis for modifications that follow. The ends of the score scale are modified to prevent the obtaining of too-low or too-high scores in the domain. Specifically, a doglegging procedure is performed to adjust the scores at the ends. In the present study, no doglegging was done and hence, the converted scores sometimes exceed the range of possible scores.

(2) In the normal procedure, minor adjustments might be added to the function, based on the results of quality control checks. For example, in rare instances, the evaluation of the function leads to minor changes in the intercept parameter.

(3) On the rare occasions that items are removed from the test (due to some major problem), the equating method described above is replaced with the Levine or Tucker method, depending on the degree of score differences found between the populations of the old and the new forms. This was not done in this study.

(4) Under normal conditions, equating is performed only between the operational sections and the anchor section of the same domain but not between the two operational sections. Here, for research purposes, equating was also performed between the two operational sections of each domain. The standard error incorporated in equating between operational sections is less than in the usual equating, since the number of examinees upon whom the equating is based is much larger than in the usual equating process.

Calculation of equipercentile equating

The definition of equipercentile equating developed by Braun & Holland (1982) and later adapted by Kolen & Brennan (1995, pp. 35-54) was used here: for a given raw score on Section X, find the percentage of examinees earning scores at or below that score. Next, find the Section Y raw score that has the same percentage of examinees at or below it. These Section X and Section Y raw scores are considered to be equivalent. For example, suppose that 20% of the examinees earned a Section X raw score of 12 or less, and 20% earned a Section Y raw score of 10 or less. The Section X raw score of 12 is then regarded as representing the same level of achievement as the Section Y raw score of 10. Hence, a Section X raw score of 12 would be equated to the raw score of 10 on Section Y. The same procedure is repeated for each of the X raw scores to obtain a conversion table from X raw scores to their percentile equivalents on the Y scale. Mathematically the process is expressed by

$$e_y(x) = y = Q^{-1}\left[P(x)\right], \tag{3}$$

where $e_y(x)$ is the equipercentile equivalents of observed scores on X on the scale of Section Y; $P(x)$ is the percentile rank function of X which gives the percentile of a given raw score x; and Q^{-1} is the inverse of the percentile rank function for Y and it gives the raw score on Section Y for a given percentile.

The definition of percentile rank may differ according to the definition of the percent at or below a given score and as a result, the equating function might change. The definition of percentile rank in this study is similar to the one used by the American College Testing Program (ACT) (see Pommerich, Hanson, Harris & Sconing, 1999).

Although X and Y have only discrete values, continuous approximations of X and Y are used. Accordingly, scores on X or Y are treated as if continuously and uniformly distributed over a range of $\pm\frac{1}{2}$ of the score (the rationale for this is presented in Holland & Thayer, 1989). Accordingly, the percentile of a given integer score of X, $P(x)$ is calculated by the percentage of examinees who earned a score below x, plus half the percentage of examinees who earned exactly x points. Thus:

$$P(x) = \Pr ob(X < x) + 0.5 \cdot \Pr ob(X = x). \tag{4}$$

When finding a score on Form Y that has the same percentile rank as a given score on Form X, the value obtained can be a non-integer. To calculate the percentile rank of a certain non-integer value, y, the integer closest to this score, defined as y*, is used (e.g., if y = 6.3, then y* = 6; if y = 6.9, then y*=7). Likewise, the percentile of a given score of Y, $Q(y)$, is calculated as the percentage of examinees who earned a score below y* plus the percentage of examinees who earned a score between y*-$\frac{1}{2}$ and y (computed as the appropriate proportion of the percentage of examinees who earned exactly y* points). Thus:

$$Q(y) = [y - (y^* - \tfrac{1}{2})] \cdot \Pr ob(Y = y^*) + \Pr ob(Y < y^*) . \qquad (5)$$

For example, the percentile rank of y = 6.3 is calculated according to the percentage of examinees who earned a score of between 5.5 and 6.3, added to the percentage of examinees who earned a score of less than 6.

In summary, the equipercentile function applied here is given by

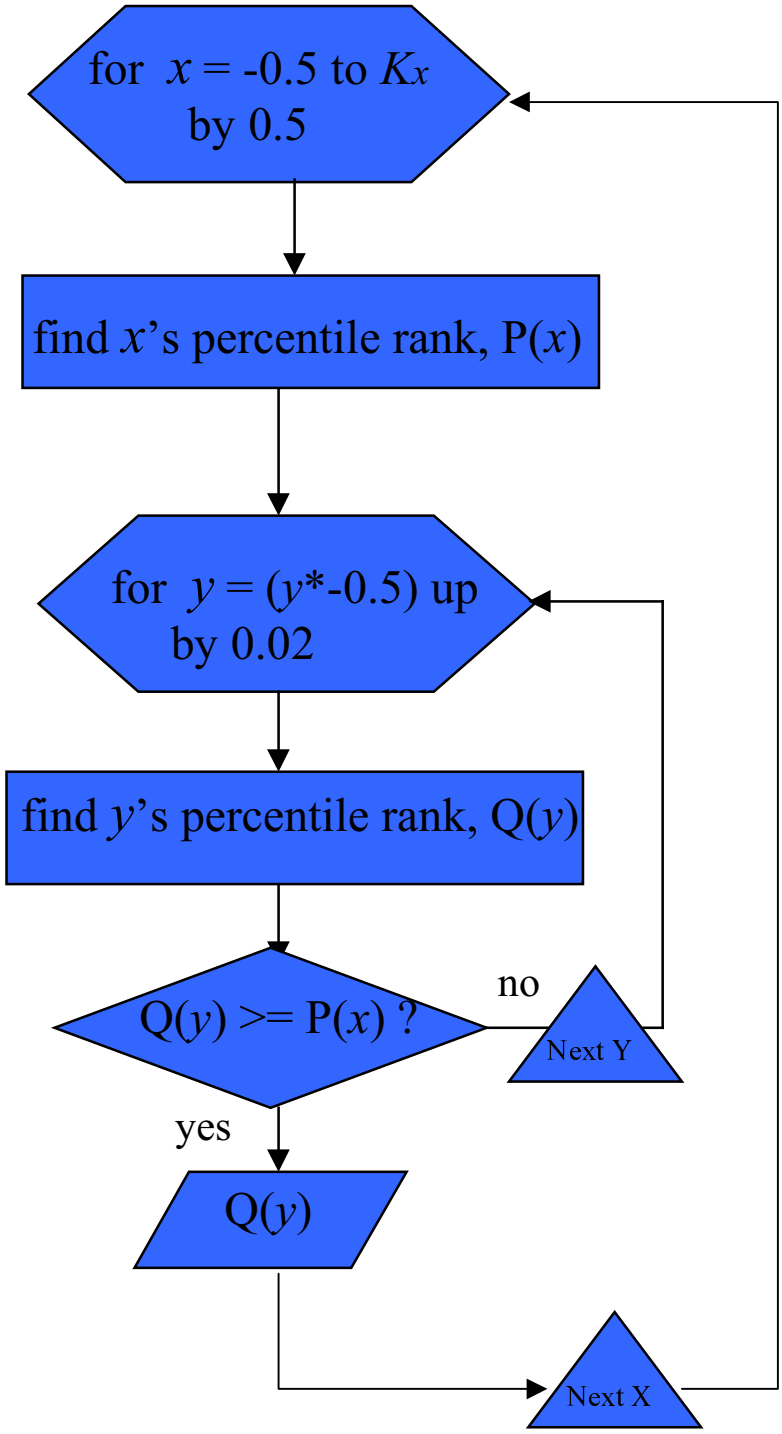
$$e_y(x) = y = \frac{\left[\Pr ob(X < x) + 0.5 \cdot \Pr ob(X = x)\right] - \Pr ob(Y < y^*)}{\Pr ob(Y = y^*)} + y^* - 0.5 , \qquad (6)$$

where y* is defined as the smallest integer value of Y such that

$$\Pr ob(X < x) + 0.5 \cdot \Pr ob(X = x) < \Pr ob(Y \le y^*) . \qquad (7)$$

This formulation was applied in the present study using a computer program in SAS. The algorithm of the program is presented below. In principle, the program first computed percentile ranks for all x values from 0 to (k_x + 0.5) in increments of 0.5 (where k_x is the number of test items in Section X) and then, for each x value, found the y value that has the same rank, as follows: After identifying the appropriate y* value, it screened all y values from y*-$\frac{1}{2}$ and up,in increments of 0.02, and calculated the percentile rank for each. Whenever there was a match between the percentile rank (of *y)* and the target percentile rank (that of *x)*, it gave the y value as the equipercentile equivalent.

Algorithm of the Equipercentile Equating Computer Program



There were several specific consequences to using this particular program:

(1) It led to an accuracy of 0.02 in the inverse function Q^{-1}.

(2) When there was no unique y score for a given percentile rank (for example, when no examinee earned a specific score on Section Y), the smallest suitable y score was chosen. Although Kolen and Brennan (1995) used the middle score and not the smallest score in a similar situation, they claim that the choice is an arbitrary one.

(3) The program ignored cases at the lower extremity of the scale where the cumulative frequency of the X scores was zero. Calculating these missing cases according to the same method would have resulted in y = -0.5, since it is the smallest y value that has a percentile rank of zero.

Once the equipercentile equivalents for all X values were found, they were plotted against the X scores and connected by lines. Whereas in a real application of the method, we would also smooth the equipercentile equivalent function, here, no smoothing was applied and the function kept its crude form. Therefore, the function exhibits a larger random error than it would have had one of the smoothing methods been applied. However, it is important to note that smoothing could in itself introduce a systematic error.

Results

The linear equating function vs. the corresponding equipercentile equating function and the differences between the two functions were ploted for all 19 PET forms included in the study. Inspection of the graphs reveals that the linear and equipercentile equating functions are fairly similar. Differences between them rarely exceed one raw score point, and in most cases do not exceed ± 0.3 points.

To summarize the differences obtained between the two equating functions, an average of the DF (difference functions) was calculated in each domain. The average

DF within a test domain was calculated four times, each time with slight modifications, as follows:
(a) The average DF across all 57 equatings performed within a test domain (see Figures 1a, 2a and 3a for the Verbal, Quantitative and English domains, respectively).
(b) The average DF calculated for the 19 equatings performed between the two operational sections (hereafter: the "op.-op." pairs) and for the 38 equatings performed between one of the operational sections and the equating section (hereafter: the "op.-eq." pairs). (Figures 1b, 2b and 3b for the Verbal, Quantitative and English domains, respectively.) This was done to control for random error that might have occurred in the DF's of the op.-eq. pairs. As explained previously, the equating samples for the op.-eq. equatings were considerably smaller than for the op.-op. equatings, and thus, the resulting equating functions and the corresponding difference contained greater random error. (c) Averaging as in (a) and (b) but using absolute differences instead of directional differences (Figures 1c and 1d, 2c and 2d, 3c and 3d for the Verbal, Quantitative and English domains, respectively.) This was done to prevent counterbalancing of opposite differences that would yield a smaller and non-representative overall result.

Two vertical broken lines at the ends of the score scale in all of the graphs mark the range of scores for which the linear equating function would have changed had the doglegging procedure been applied. The lines are positioned at modal values, derived from real doglegging procedures performed over the past years. This range represents about 90% of the score distribution. In addition, the equating results tend to be more accurate in this range than at the ends of the scale where only a small number of examinees is available. Thus, evaluations and comparisons between functions or graphs are confined to the central range of the score scale.

Verbal Domain

Figure 1a shows the average DF across all 57 equatings performed within the Verbal domain. As can be seen, the average DF ranges from -0.10 to +0.20 raw scores (increasing towards the ends). Similar results occur in Figure 1b, where we calculated the average DF by types of section pairs (i.e. op.-op pairs vs. op.-eq. pairs).

The average of the absolute DF (Figure 1c) is similar to the original results presented in Figure 1a. In general, the average absolute DF does not exceed 0.25 points. Figure 1d shows that the average absolute DF tends to be smaller for the op.-op. pairs than for the op.-eq. pairs. As explained previously, this could be due to the larger standard error that results from equating between an operational and an anchor section as compared with equating between two operational sections.

Quantitative Domain

Figure 2a shows the average DF across all 57 equatings performed within the Quantitative domain. As can be seen, the average DF ranges from -0.10 to +0.10 points. A comparison of the average DF for op.-op. pairs and op.-eq. pairs reveals only small differences, which are not in any specific direction (figure 2b).

Figure 2c presents the results of the average absolute DF. It can be seen that the average absolute DF does not exceed 0.2 points. This finding is somewhat smaller than in the parallel result for the Verbal domain. However, it should be kept in mind that the range of the raw score scale of the Quantitative sections is narrower than that of the Verbal sections (there are 25 items in a Quantitative section and 30 items in a Verbal section). Hence, the potential for differences between the two equating functions in the Verbal sections is greater than in the Quantitative sections, and the differences should be viewed relative to the number of items in a section. In fact, the maximal difference relative to the range of the raw score scale is almost identical in the two test domains: 0.2/25 = 0.008 in the Quantitative domain and 0.25/30=0.00833 in the Verbal domain.

Another difference between the two test domains is that when averaging the DF's in the Quantitative domain by types of section pairs, no dissimilarities were found (figure 2d), as opposed to the Verbal domain (figure 1d). This may indicate that there is a smaller random equating error in the Quantitative domain.

Verbal Domain

Figure 1a: Average DF across all equatings in the test domain.

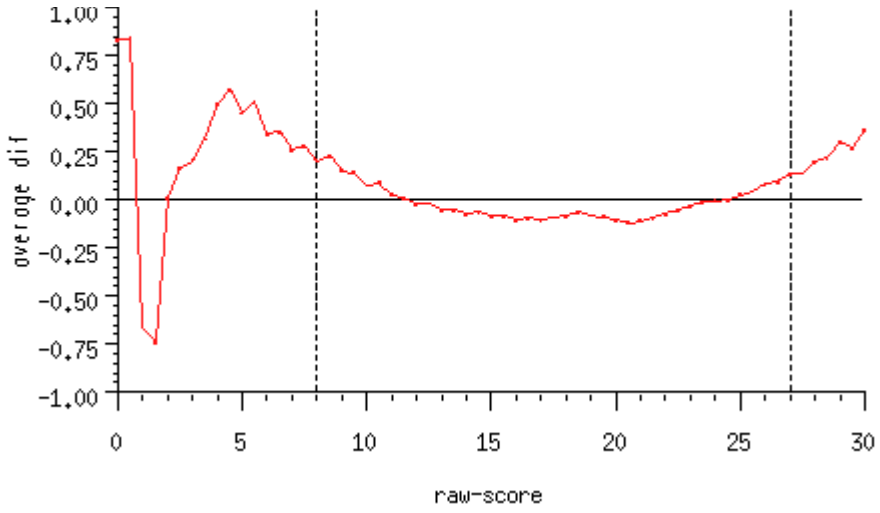


Figure 1b: Average DF by type of section pairs
(operational-operational pairs vs. operational-equating pairs).

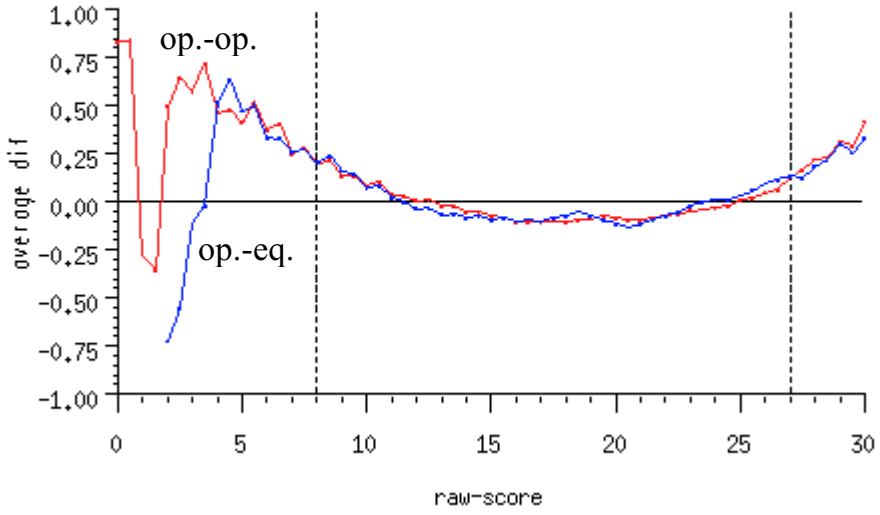


Figure 1c: Average Absolute DF across all equatings in the test domain.

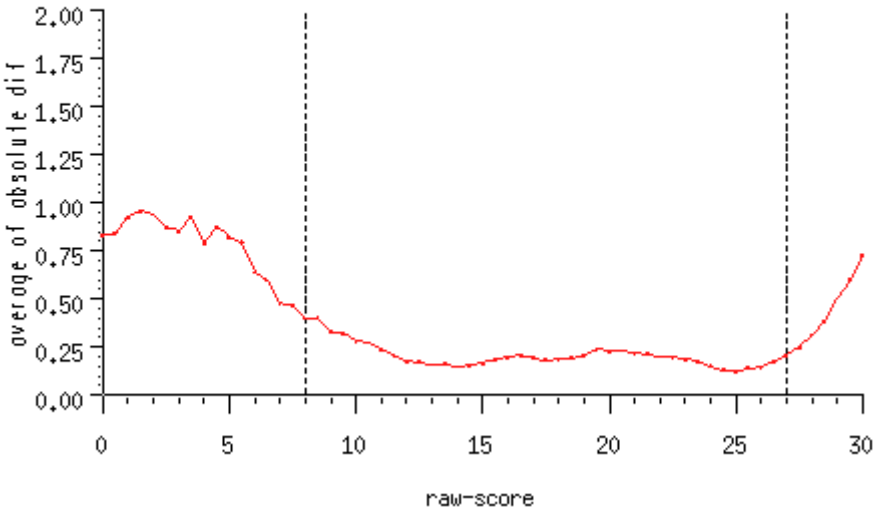
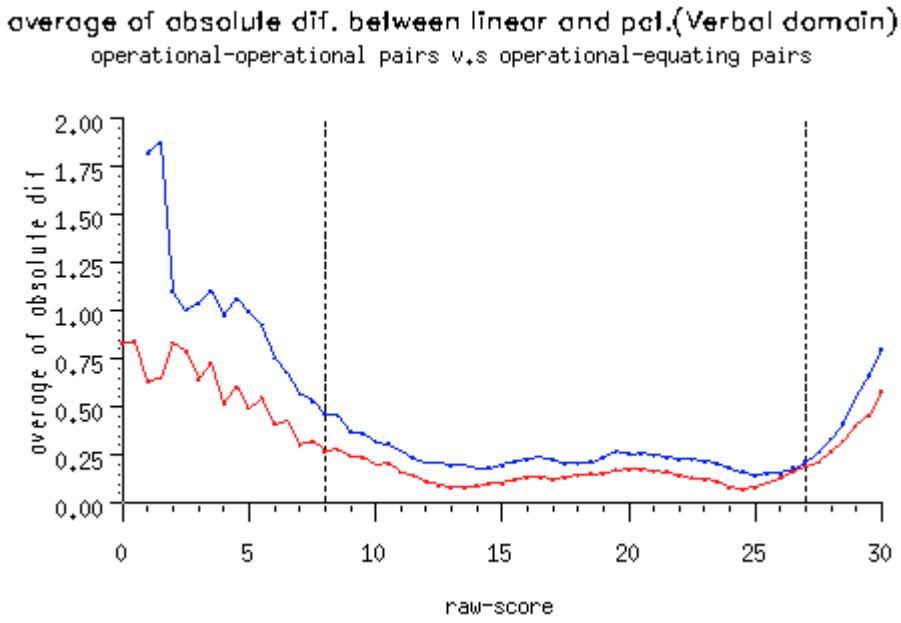


Figure 1d: Average Absolute DF by type of section pairs
(operational-operational pairs vs. operational-equating pairs).



2. Quantitative Domain

Figure 2a: Average DF across all equatings in the test domain.

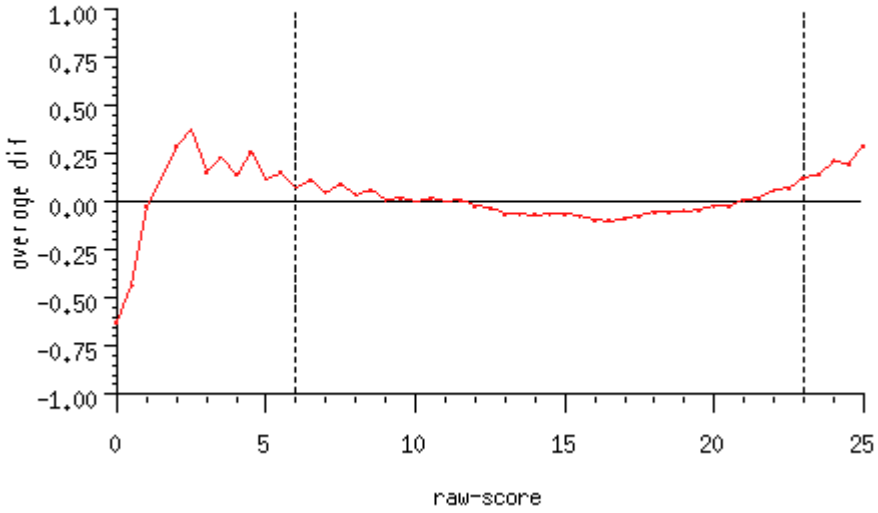


Figure 2b: Average DF by type of section pairs
(operational-operational pairs vs. operational-equating pairs).

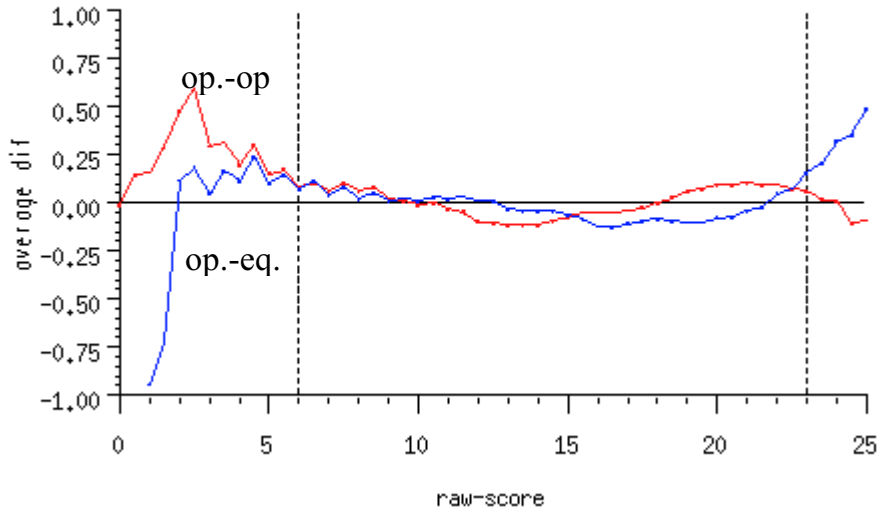


Figure 2c: Average Absolute DF across all equatings in the test domain.

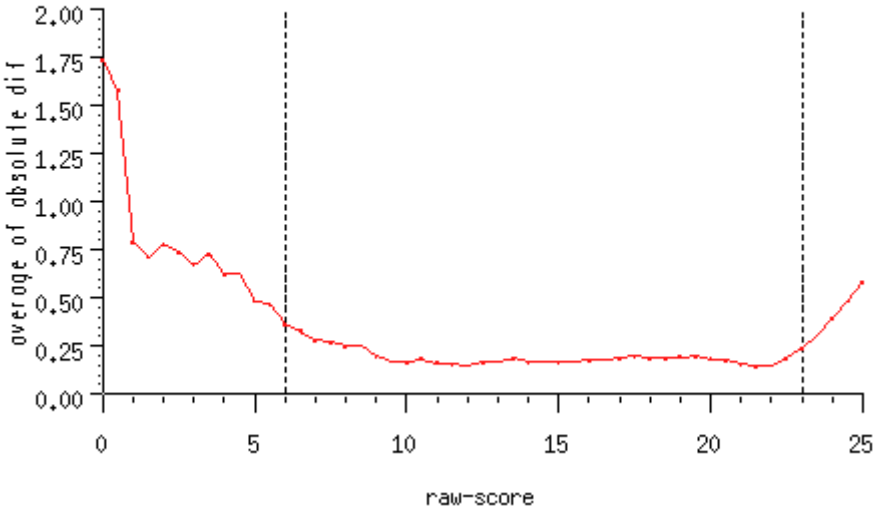
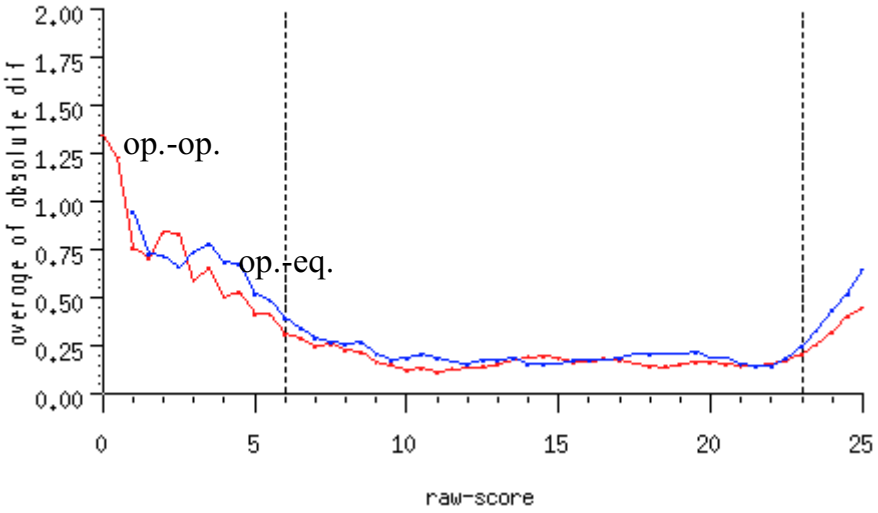


Figure 2d: Average Absolute DF by type of section pairs
(operational-operational pairs vs. operational-equating pairs).



3. English Domain

Figure 3a: Average DF across all equatings in the test domain.

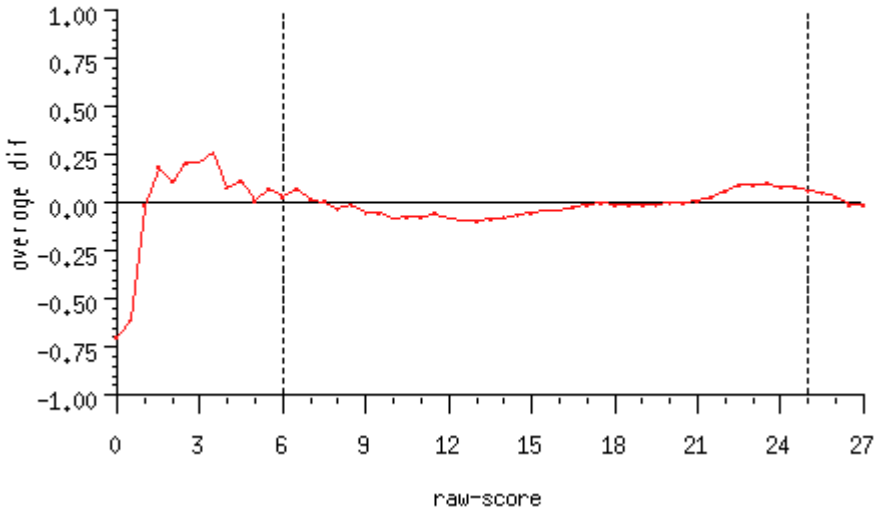


Figure 3b: Average DF by type of section pairs
(operational-operational pairs vs. operational-equating pairs).

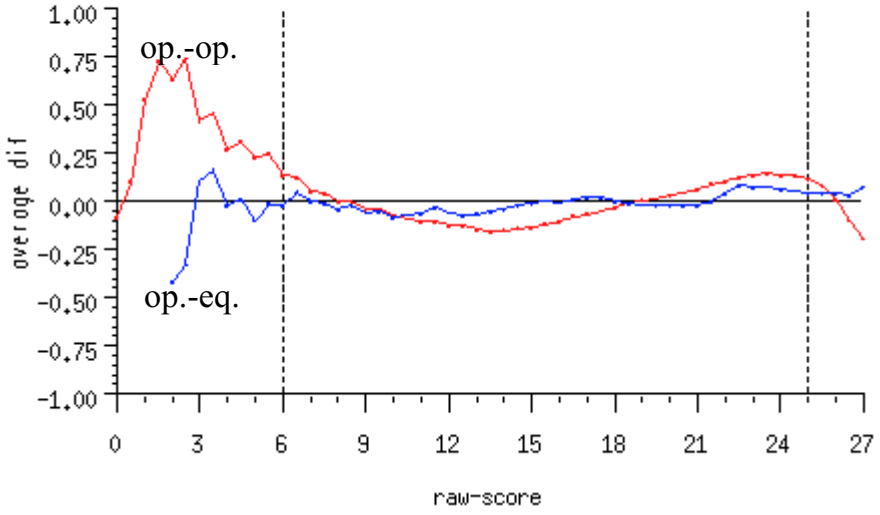


Figure 3c: Average Absolute DF across all equatings in the test domain.

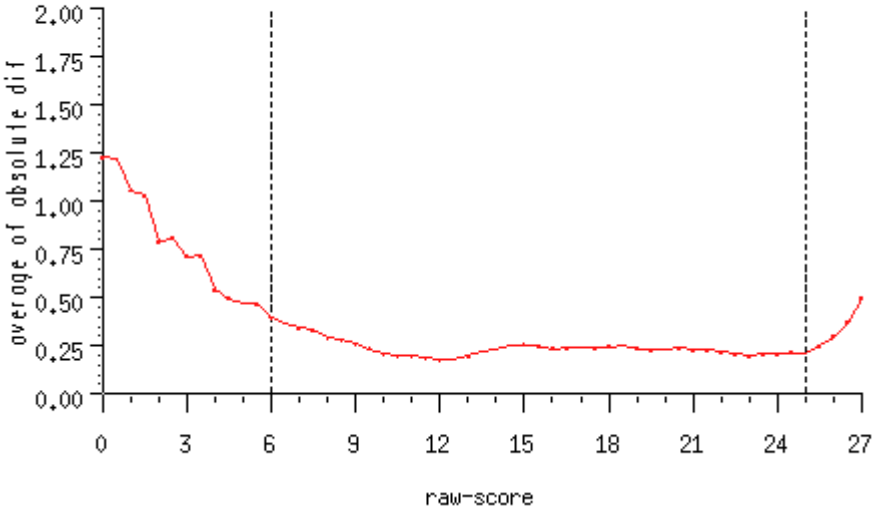
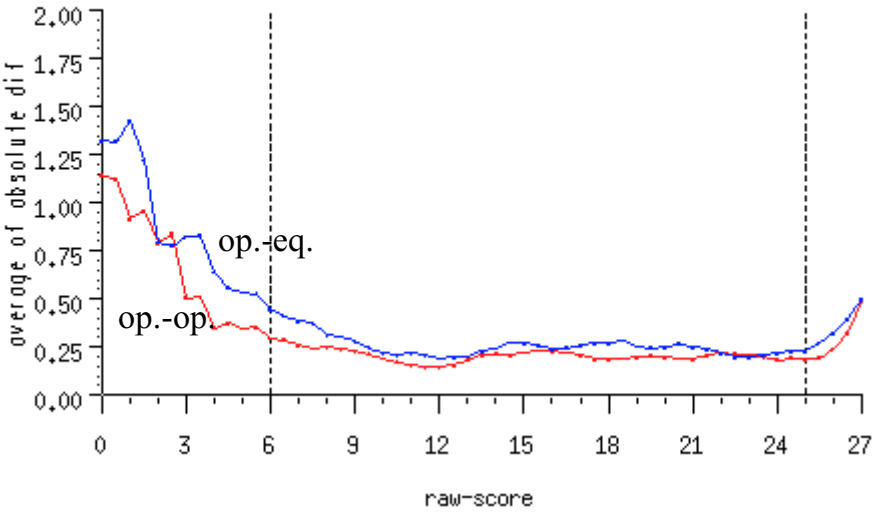


Figure 3d: Average Absolute DF by type of section pairs
(operational-operational pairs vs. operational-equating pairs).



English Domain

Figure 3a shows the average DF across all 57 equatings performed within the English domain. As can be seen, the average DF ranges from -0.20 to +0.20 points. There were small differences between the functions of the average DF by types of section pairs (Figure 3b). Unexpectedly, the average DF tended to be higher for the op.-op. pairs than for the op-eq pairs. But since this was not the case when we calculated the average of the absolute DF, (see Figure 3d) it was assumed that there was greater counterbalancing between differences in opposite directions in the op.-eq. pairs than in the op.-op. pairs, which lead to a smaller average DF.

Figure 3c shows that the average of the absolute DF generally did not exceed 0.25 points. The maximum value relative to the range of the raw score scale (0-27) in an English section was 0.25/ 27=0.0093, which is slightly higher than in the other two domains.

Discussion

Linear equating, as currently performed at NITE, and equipercentile equating yielded very similar results. This was true for PET's three test domains - Verbal Reasoning, Quantitative Reasoning and English as a foreign language. For the most part, differences between the equipercentile and linear functions tended to be fairly close to the ranges of the random equating error that NITE is familiar with . For example, the random equating error between Verbal sections was estimated at approximately 0.1-0.2 raw scores in the normal linear equating performed at NITE with a sample of 1,000 examinees, and at approximately 0.2-0.4 raw scores in the equipercentile equating performed with the same equating sample size (Rapp, 1999). Furthermore, with the exception of the ends of the scale where the differences tended to increase, the average DF's were fairly stable along the scale and did not produce a specific pattern. Differences between linear and equipercentile equating functions

were not found to be larger in certain ranges of the score scale and they were neither consistent nor systematic, either in direction or in size.

Thus, differences in most cases between the two methods appear to be due more to random factors than to typical non-linear relationships that presumably exist between sections. The results reinforce the working assumption that a linear function can accurately describe typical section-to-section differences in difficulty in PET; i.e., the fundamental equating relationship between scores on two PET sections (for a given domain) is nearly linear. This may be attributable to the fact that the score distributions on most sections in a given domain are fairly similar, indicating careful preparation of the forms (Brennan, in Cohen, 1999).

In conclusion, since the equipercentile method of equating appears to produce results similar to those of the linear equating method, there seems to be little, if any, justification for replacing the linear method of equating with one that is more complex and which requires a larger equating sample.

References

Allalouf, A. (1999). Scoring and Equating at the National Institute for Testing and Evaluation, NITE report, Jerusalem. (In preparation).

Braun, H. I., & Holland, P. W. (1982). Observed-score test equating: A mathematical analysis of some ETS equating procedures. In P.W. Holland & D. B. Rubin (Eds.) Test equating (pp.9-49). New York: Academic Press.

Cohen, Y. (1999). Proceedings of the 2nd bi-annual meeting of NITE Scientific Council, NITE report, Jerusalem (Internal report).

Holland, P.W., & Thayer, D.T. (1989). The Kernel method of equating score distributions . research report No. 89-7). Princeton, NJ: Educational Testing Service.

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

Pommerich, M., Hanson, B. A., Harris, D. J., & Sconing, J. A. (1999). Issues in creating and reporting concordance results based on equipercentile methods. A paper presented at the annual meeting of the National Council on Measurement in Education, Montreal, Quebec, Canada, April 1999.

Rapp, J. (1999). Research on equating PET forms. Paper presented at the 2nd bi-annual meeting of NITE scientific council, Jerusalem, July 1999.