Research Report

Population Invariance and Linear Equating for the Non-Equivalent Groups Design

Alina A. von Davier

Ning Han

Research & Development

October 2004
RR-04-47

Population Invariance and Linear Equating for the Non-Equivalent Groups Design

Alina A. von Davier and Ning Han

ETS, Princeton, NJ

October 2004

Abstract

This study investigates the population sensitivity of the commonly used linear equating methods in the Non-Equivalent-groups with an Anchor Test (NEAT) design: the Tucker, the Levine observed-score, and the chain linear methods. For a detailed analysis of the subject, we apply three distinctive approaches to a real data set from a NEAT design: a) the RMSD index for the NEAT design of von Davier, Holland, and Thayer (2004); b) the parallel-linking system of Dorans and Holland (2000), and c) the pseudo-NEAT design approach of von Davier, Holland, and Thayer (2003).

The data set is used to illustrate the derivations on male and female subpopulations. Comparisons of the results obtained using linear and equipercentile equating are also presented. For this particular data set the Levine function seems to vary less across subpopulations than the other methods; the Tucker method seems to vary the most.

Key words: Linear equating functions, Non-Equivalent-groups with an Anchor Test (NEAT) design, root mean squared difference (RMSD), population invariance, test equating

Acknowledgements

The authors thank Neil Dorans, Skip Livingston, and Dan Eignor for their suggestions and comments on the previous version of this paper and to Kim Fryer for editorial support. This research was supported in part by an ETS Summer Program in Research Graduate Internship awarded to Ning Han.

Introduction

Test equating methods are used to produce scores that are interchangeable across different test forms. One of the five requirements of equating functions mentioned in Dorans and Holland (2000) is that equating should be population invariant, that is, equating functions yielded from different subpopulations should be very close to the function from the whole population.

This present study adapts and compares previously introduced measures or approaches for studying population invariance of equating functions to linear equating functions in the Non-Equivalent-groups with the Anchor Test (NEAT) design.

Dorans and Holland (2000) introduced a measure of the extent to which an equating function is sensitive to the population on which it is computed. This measure, the root mean squared difference (RMSD), compares equating (or linking) functions computed on different subpopulations with the function computed for the whole population within single-group/equivalent-groups designs. Von Davier, Holland, and Thayer (2003, 2004) generalized the RMSD measure to the NEAT design.

Dorans and Holland (2000) also introduced a system of parallel-linear linking functions for investigating the degree of population sensitivity of an equating function within single-group/equivalent-groups designs. The idea is to preserve the same slope for the linear equating functions computed on subpopulations as that obtained from the linear equating function for the total population (under some assumptions that are discussed later). Hence the differences between the equating functions in the subpopulations and the function in the total population are captured in the differences between the intercepts of the parallel linear functions.

Von Davier, Holland, and Thayer (2003) described a new approach for diagnosing the observed-score equipercentile equating methods used in the NEAT design with respect to the population invariance requirement. They constructed a pseudo-NEAT design, where a test is reused on two subpopulations that are identifiable within the population(s) from which samples are drawn (a special case of the self-equating procedure). In this way, they provide a method to test the assumptions made by the equating functions in the NEAT design. Hence, one can check which equating method is least sensitive to violations of the population invariance requirement for a given set of data.

This paper investigates the population invariance requirement for the linear Tucker, Levine observed-score, and chain equating procedures. For a detailed analysis of the subject, we consider

the three distinctive approaches just described: the RMSD index, the parallel-linking system, and the pseudo-NEAT design approach. The goals of this paper are:

1. To adapt the RMSD formula to each of the linear equating functions used in the NEAT design.
2. To extend the system of parallel equating functions (Dorans & Holland, 2000) to the NEAT design and to apply it to the three linear equating methods (the Tucker, Levine, and chain equating).
3. To compare the outcomes of the combination of the pseudo-NEAT design method across the three linear equating functions being studied and then to compare them to the results for equipercentile equating functions reported in von Davier et al. (2003).

The same data as used in von Davier et al. (2003) are used in this study; this allows us to compare the pseudo-NEAT design approach for the linear equating functions (investigated here) with the results previously obtained by von Davier et al. (2003) for the equipercentile equating functions. One reason for doing this comparison is to check if the pseudo-NEAT design approach (that makes use of the self-equating method) is more stable when using linear equating than when using equipercentile functions (Kolen, 2003).

Method

In this section we introduce our notation, we summarize the methods available for assessing the population sensitivity of an equating function, and we describe each of the three linear equating methods that are studied in this paper.

Notation

The NEAT design is described in Holland (2002). Here, the basic notation is reviewed:

	X	V	Y	
P	✔	✔		X, V observed on P
Q		✔	✔	Y, V observed on Q

Usually, X and Y are the operational tests and V is the anchor test. X and V are given to a sample from the test administration P and, Y and V are given to a sample from the test administration Q. The anchor test score, V, can be either a part of both X and Y (the internal anchor case) or a separate score (the external anchor case). X, V, and Y are used to refer to both tests and test scores. The two tests, X and Y, are to be equated and this is to be done on a target population, T (see, for example, Braun & Holland, 1982, or Kolen & Brennan, 2004).

The target population T for the NEAT design is a mixture of P and Q denoted by

$$T = wP + (1-w)Q, \tag{1}$$

where w is a weight that lies between 0 and 1.

Many observed score equating methods are based on the *linear equating function*, defined on the target population, T, by setting standardized deviation scores (z-scores) on the two forms to be equal such that

$$\frac{x-\mu_{XT}}{\sigma_{XT}} = \frac{y-\mu_{YT}}{\sigma_{YT}},$$

where $\mu_{YT}, \sigma_{YT,} \mu_{XT},$ and σ_{XT} are the means and the variances of X and Y in T. Solving for y in the above equation results in the formula for the linear equating function,

$$\mathrm{Lin}_{XY;T}(x) = y = \mu_{YT} + \sigma_{YT}\left(\left(x-\mu_{XT}\right)/\sigma_{XT}\right), \tag{2}$$

which converts the observed scores on X to the scale of Y on T.

Measures of Population Invariance

In this subsection, we summarize the methods and formulas available for investigating the population invariance of an equating function. First, we will present the methods available for single-group/equivalent-groups designs. Then, we will describe the methods available for the NEAT design. At the end of the subsection, we indicate which are the formulas we use in this study for investigating the population invariance of the linear equating methods within the NEAT design.

1. *Measures of population invariance of an observed-score equating function when there is only one population underlying the data collection design (i.e., measures of invariance based on single-group/equivalent-groups design).*

    1. Dorans and Holland (2000) introduce a measure of the degree to which an equating function is sensitive to the population on which it is computed. This measure, the root mean squared difference (RMSD), compares equating (or linking) functions computed on different subpopulations with the function computed for the whole population. The formula introduced by Dorans and Holland (2000) is applicable to both the equivalent-groups and single-group designs, where the sample(s) are drawn from one population, *P*. The formula for the RMSD difference is given in (3).

$$\mathrm{RMSD}(x) = \frac{\sqrt{\sum_j w_j \left[e_{P_j}(x) - e_P(x) \right]^2}}{\sigma_{YP}}, \tag{3}$$

    where *P* is the target population in the equivalent-groups and single-group designs, which is the population from which the samples were drawn. In (3), e_P denotes the equating function that equates *X* to *Y* on the whole population *P;* e_{Pj} denotes the equating function that equates *X* to *Y* on the subpopulation P_j of *P*; and w_j denotes the relative proportion of P_j in *P*. It is standardized by dividing by the standard deviation of *Y* on *P* so that it is a type of effect size and it may be interpreted as a percentage of the *y*-standard deviation on *P*. Von Davier et al. (2004) present arguments for giving equal weight, w_j, to each subpopulation link for computing the RMSD values.

    2. Dorans and Holland (2000) also introduce the system of parallel-linear linking functions for multiple subpopulations in the single-group/equivalent-groups designs framework (where there is only one population, *P*):

$$\mathrm{Lin}_{Y;P_j}(x) = \mu_{YP_j} + \frac{\sigma_{YP}}{\sigma_{XP}}\left(x - \mu_{XP_j}\right), \tag{4}$$

The differences between the linking functions in the subpopulations and the function in the total population are reflected in the differences between the intercepts of the parallel linear functions. An advantage of using a system of parallel-linear linking functions is that the RMSD measure has a constant value at each point x.

Obviously, the system of parallel linking functions is useful if a linear conversion is appropriate for the data and if the main differences among the distributions of the two tests are in their means, or at least that the ratio of the standard deviations of the two tests is similar across subpopulations.

2. *Measures of population invariance of an observed-score equating function when there are two populations underlying the data collection design.*

   a. Von Davier et al. (2003) generalized RMSD for observed-score equating to the NEAT design. As mentioned before, in the NEAT design there are two populations from which the samples of examinees are drawn. T denotes the target population in the NEAT design defined in (1).

   Let the examinees from each population P and Q be divided into subgroups, $\{P_j\}$ and $\{Q_j\}$, that are defined the same way for P and Q. Let w_{Pj} define the relative proportion of each subgroup P_j, as shown in (5); similarly, let w_{Qj} define the relative proportion of each subgroup Q_j in Q, again, as shown in (5) (note that w_{Pj} can also be some other set of weights that sum to unity; similarly for w_{Qj} —see Yang, 2004).

   This is denoted by

$$P = \sum_{j} w_{Pj} P_j \quad \text{and} \quad Q = \sum_{j} w_{Qj} Q_j. \tag{5}$$

   The weights, $\{w_{Pj}\}$ and $\{w_{Qj}\}$, are allowed to be different in P and Q, if necessary. The target population for each subgroup represents populations P and Q in the same proportion as does the target population for the total group: T_j, the target population for the subgroup j is defined, following (1), as

$$T_j = wP_j + (1 - w)Q_j \,, \tag{6}$$

where the same weight, w, as in (1) was used.

The weights for the RMSD formula are computed as

$$w_j = w(w_{Pj}) + (1 - w)w_{Qj}. \quad (7)$$

Von Davier et al. (2004) present arguments for giving equal weight, w_j, to each subpopulation link for computing the RMSD values.

Let $e_{Tj}(x)$ be a function that equates X to Y on T_j, and $e_T(x)$ be a function that equates X to Y on T. Both equating functions $e_{Tj}(x)$ and $e_T(x)$ are assumed to be "computed in the same way" (i.e., both are derived using the same equating function). RMSD(x) is defined by von Davier et al. (2003) as

$$\mathrm{RMSD}(x) = \frac{\sqrt{\sum_j w_j \left[e_{T_j}(x) - e_T(x) \right]^2}}{\sigma_{YT}}, \quad (8)$$

where the choice of the denominator, σ_{YT}, depends on the equating method and on the assumptions the methods make. Since Y is not observed in T, the standard deviation of Y in a synthetic population, T, is computed following the assumptions of the equating methods: for example, the σ_{YT} can be computed following the Tucker method (see Kolen & Brennan, 2004) or the σ_{YT} can be computed following the chain linear equating method (see von Davier et al., 2004). One could eventually use the standard deviation of the equated scores [i.e., $\sigma(\mathrm{e}_T(x))$] since the equated scores are on the Y-scale; however, in this paper the interest is to standardize the measure by using the standard deviation of the target distribution, which is Y on T.

2. The system of parallel-linear linking functions from (4) can be easily extended to the NEAT design, as the following:

$$\mathrm{Lin}_{Y;T_j}(x) = \mu_{YT_j} + \frac{\sigma_{YT}}{\sigma_{XT}}(x - \mu_{XT_j}), \quad (9)$$

where the means and the standard deviations of X and Y on the target population T and on the target subpopulations T_j, are obtained through different formulas depending on the equating procedure being used.

3. Von Davier et al. (2003) investigated a combination of pseudo-NEAT design and RMSD for comparing equating methods with respect to subpopulation differences.

In the NEAT design, X is observed on P and Y is observed on Q, but neither X or Y are observed on P and on Q. Thus, X and Y are not both observed on T, regardless of the choice of w. For this reason, all equating methods must make assumptions to overcome this lack of complete information in the NEAT design. These assumptions usually claim population invariance of different statistical relationships about the quantities that are not observed (the regressions of the tests on the anchor are assumed to be population invariant, in the case of the Tucker method; and the linking function between the tests and the anchor are assumed to be population invariant, in case of the chain method; the regressions of the true scores of the tests on the true score of the anchor are population invariant, in the case of the Levine function, etc.)—see von Davier et al. (2003).

The idea developed in von Davier et al. (2003) is to find a way to test these nontestable assumptions with data at hand. Von Davier et al. (2003) create two pseudo-NEAT designs, each one inside one of the two populations, P and Q. The focus is on one total population at a time, for example, on P, and to partition it into two subpopulations of interest, like males and females. Then, since all the examinees from P took test X and the anchor V, we can observe X and V in each of the two subpopulations of interest, males and females. Now, this pseudo-NEAT design looks like a NEAT design in Table 1, where P and Q are replaced by male and female subpopulations, and Y is replaced by X, and where we observe data everywhere. By having data for X and V everywhere we can test the assumptions made by different equating methods and see which equating method has the assumptions fulfilled to a greater degree. In order to achieve this goal, von Davier et al. (2003) make use of the fact that equating X to X through V in this pseudo-NEAT design should result in the identity function (up to sampling variation).

Note that this is a different usage of the self-equating procedure than previously discussed in the literature: here the self-equating is used to check the assumptions behind each of the equating methods. Moreover, the investigation of the assumptions done inside each population cannot be generalized to other populations or other assessments; at least not as is. An attempt to generalize these findings, although very interesting, is beyond the scope of this study.

Von Davier et al. (2003) applied the RMSD to the pseudo-NEAT design as follows:

$$RMSD(x) = \frac{\sqrt{w_M\,(e_M(x) - x)^2 - w_F\,(e_F(x) - x)^2}}{\sigma_{XP}}, \qquad (10)$$

where the expected equating function on the total population P is the identity function (after equating X to X through V on P, where $P = w_M M + w_F F$ and w_M and w_F are the weights described in (5) for j=2). In (10), $e_M(x)$ and $e_F(x)$ are the equating functions from X to X through V, if one considers $P = M$ and $P = F$, respectively, in the formulas for equating functions (see von Davier, et. al, 2003, for computational details). The denominator is the standard deviation of X in the total population, now P.

In this study, the RMSD formula from (8) will be applied to the Tucker, the Levine, and the chain linear equating methods, respectively.

Then, the parallel-linking system described in (9) will be applied to the three linear equating methods in our study.

Finally, we create two pseudo-NEAT designs inside each population, P and Q, respectively, and we apply the RMSD from (10) to each of the three linear functions.

Linear Equating Methods in a NEAT Design

This subsection recalls the formulas used for computing the linear equating functions in a NEAT design. Von Davier et al. (2003) and von Davier and Kong (in press) showed that all linear observed-score equating functions (including chain linear) in a NEAT design can be viewed as (2), where the parameters $\mu_{YT}, \sigma_{YT}, \mu_{XT}$, and σ_{XT} are obtained through different *method functions*.

Based on these results, a theoretical framework can be developed using the general form in (1), without referring explicitly to the formulas and assumptions of each of the three equating functions. The assumptions for the Tucker and Levine observed-score equating functions can be found in Kolen and Brennan (2004); the assumptions for the chain equating can be found in von Davier et al. (2004).

Equation 2 can be developed for each of the three equating methods using the operational formulas that follow (see Kolen & Brennan, 2004, pp. 103-115, for the derivations):

For the Tucker and Levine observed-score functions, the operational formulas are

$$\mu_{XT} = \mu_{XP} - (1-w)\Delta_P(\mu_{VP} - \mu_{VQ}), \tag{11}$$

$$\mu_{YT} = \mu_{YQ} + w\Delta_Q(\mu_{VP} - \mu_{VQ}), \tag{12}$$

$$\sigma^2_{XT} = \sigma^2_{XP} - (1-w)\Delta^2_P(\sigma^2_{VP} - \sigma^2_{VQ}) + w(1-w)\Delta^2_P(\mu_{VP} - \mu_{VQ})^2, \tag{13}$$

$$\sigma^2_{YT} = \sigma^2_{YQ} + w\Delta^2_Q(\sigma^2_{VP} - \sigma^2_{VQ}) + w(1-w)\Delta^2_Q(\mu_{VP} - \mu_{VQ})^2. \tag{14}$$

The four Δ-parameters that distinguish the two equating methods—Tucker and Levine—have the following formulas.

For the Tucker method,

$$\Delta_P = \alpha_P = \frac{\sigma_{XVP}}{\sigma^2_{VP}}, \tag{15}$$

$$\Delta_Q = \alpha_Q = \frac{\sigma_{YVQ}}{\sigma^2_{VQ}}, \tag{16}$$

where σ_{XVP} denotes the covariance of X and V in P and σ_{YVQ} denotes the covariance of Y and V in Q.

For the Levine method with an external anchor:

$$\Delta_P = \gamma_P = \frac{\sigma^2_{XP} + \sigma_{XVP}}{\sigma^2_{VP} + \sigma_{XVP}} \tag{17}$$

and

$$\Delta_Q = \gamma_Q = \frac{\sigma^2_{YQ} + \sigma_{YVQ}}{\sigma^2_{VQ} + \sigma_{YVQ}}. \quad (18)$$

For the Levine method with an internal anchor:

$$\Delta_P = \gamma_P = \frac{\sigma^2_{XP}}{\sigma_{XVP}} \quad (19)$$

and

$$\Delta_Q = \gamma_Q = \frac{\sigma^2_{YQ}}{\sigma_{YVQ}}. \quad (20)$$

For the chain equating the means and the standard deviations of X and Y on T are given in von Davier et al. (2004) and are repeated here

$$\mu_{XT} = \mu_{XP} + \frac{\sigma_{XP}}{\sigma_{VP}}(\mu_{VT} - \mu_{VP}), \quad (21)$$

$$\mu_{YT} = \mu_{YQ} + \frac{\sigma_{YQ}}{\sigma_{VQ}}(\mu_{VT} - \mu_{VQ}), \quad (22)$$

$$\sigma_{XT} = \frac{\sigma_{VT}}{\sigma_{VP}}\sigma_{XP}, \quad (23)$$

$$\sigma_{YT} = \frac{\sigma_{VT}}{\sigma_{VQ}}\sigma_{YQ}. \quad (24)$$

Data

The same data as used in von Davier et al. (2003) are used in this study; this allows us to compare the pseudo-NEAT design approach for the linear equating functions (investigated here) with the results previously obtained by von Davier et al. (2003) for the equipercentile equating functions. One reason for doing this comparison is to check if the pseudo-NEAT design approach (that makes use of the self-equating method) is more stable when using linear equating than when using equipercentile functions (Kolen, 2003).

The data are from the 1998 and 2000 administrations of the AP® English Language & Composition Examination. These data sets have 79,434 examinees in the 1998 administration and 112,868 examinees in the 2000 administration (see Table 1 for the summary statistics).

Table 1

Summary Statistics From the 1998 and 2000 Administrations of the AP English Language & Composition Exam

	X_P	V_P	V_Q	Y_Q
Mean	32.8	9.258	8.996	35.51
SD	10.83	3.16	3.22	11.39
Number of items	56	14	14	55

Note. P denotes the data from the 1998 administration. *Q* denotes the data from the 2000 administration.

Each particular AP exam has a composite score, which is a weighted sum of scores from the multiple-choice and free-response parts. Here we use the multiple-choice (MC) data only. This AP exam uses a NEAT design with the year 2000 test being equated back to the one given in 1998. The anchor test (which has 14 items) is an embedded anchor within the MC component of the whole test (the test given in 1998 has 55 items, the test given in 2000 has 56 items). The reliability for the two MC tests was the same, 0.82 (given that the two tests have the same reliability, we decided that we can use the Levine observed equating, or also called the "Levine equally reliable equating method" as we initially planed). The two subpopulations we examined were males and females. In 1998 there were 30,217 male and 49, 217 female test takers, and in 2000 there were 42,317 male and 70,551 female test takers.

Table 2

Effect Size Calculations for Males/Females Differences on the Anchor Test

Year	Males	Females	All	Effect size
1998	9.408 (3.16)	9.166 (3.15)	9.258 (3.16)	7.7%
2000	9.152 (3.21)	8.903 (3.23)	8.996 (3.22)	7.7%
Effect size	8.1%	8.2%		

Note. MC anchor test data from the 1998 and 2000 administrations of the AP English Language & Composition Exam.

The effect size for the difference between 1998 and 2000 for all examinees is (9.258 – 8.996)/3.19 = 8.2%. (3.19 is the average of 3.22 and 3.16). Thus, the 7.7% effect size for the males/females differences each year are similar to the 8.2% effect size for the difference between the 2 years.

The observed correlations between the Y and anchor test V_Q (in the sample from 1998) and between X and V_P (in the sample from 2000) are both about 0.8.

Results

Figure 1 shows the RMSD (from Equation 8) for the Tucker, the Levine, and the chain linear methods. The RMSD values range from 0.043 (or 4.3% of a standard deviation) to 0.005 (or 0.5% of a standard deviation), which are considered to be small for many types of decisions regarding the equating process. The Levine function seems to show greater population dependence at the lowest scores than the other methods, but given the fact that there is not much data at the lowest scores, the accuracy of these equated results is low.

There are no clear differences between the three linear equating methods with respect to their deviations from the population invariance assumption. It is clearly the case that the RMSD are larger at the lower end of the score range, for all three methods.

In Figure 2 we plot the system of the parallel-linking functions for the Tucker equating. Actually, in Figure 2 we plot the differences between the Tucker linear function for the total group and the parallel Tucker linear functions for the subgroups (hence, the zero line would be the ideal differences plot, where there would be no difference between the equating function for the total

and the functions on the subpopulations). This plot shows that the three equating functions that correspond to the total population, the male population, and the female population are very close to each other, with the equating results for the male population deviating slightly more from the equating results for the total. This information is consistent with Figure 1, in the sense that the equating functions are population sensitive in a small degree.

The three RMSD values for each function are very small for the system of parallel linear functions: RMSD (Tucker) = 0.0185; RMSD (chain) = 0.0183, and RMSD (Levine) = 0.0185. The plots for the chain and the Levine functions for the system of parallel linking look almost identical with those for the Tucker and are omitted.

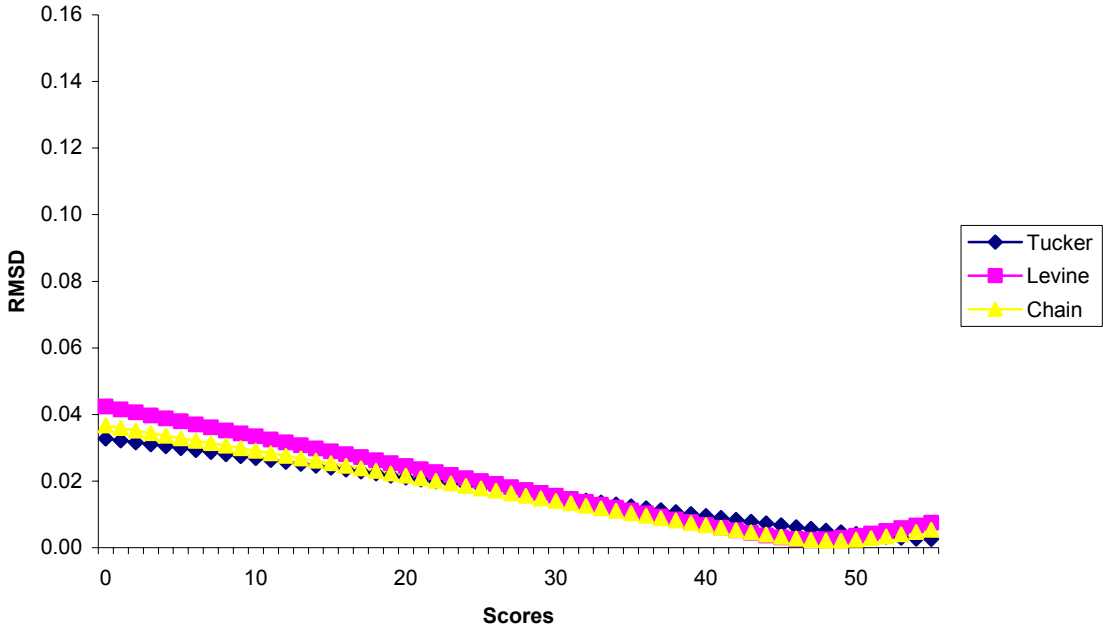

Figure 1. **The RMSD values for three linear equating functions on the data from the 1998 and 2000 administrations.**

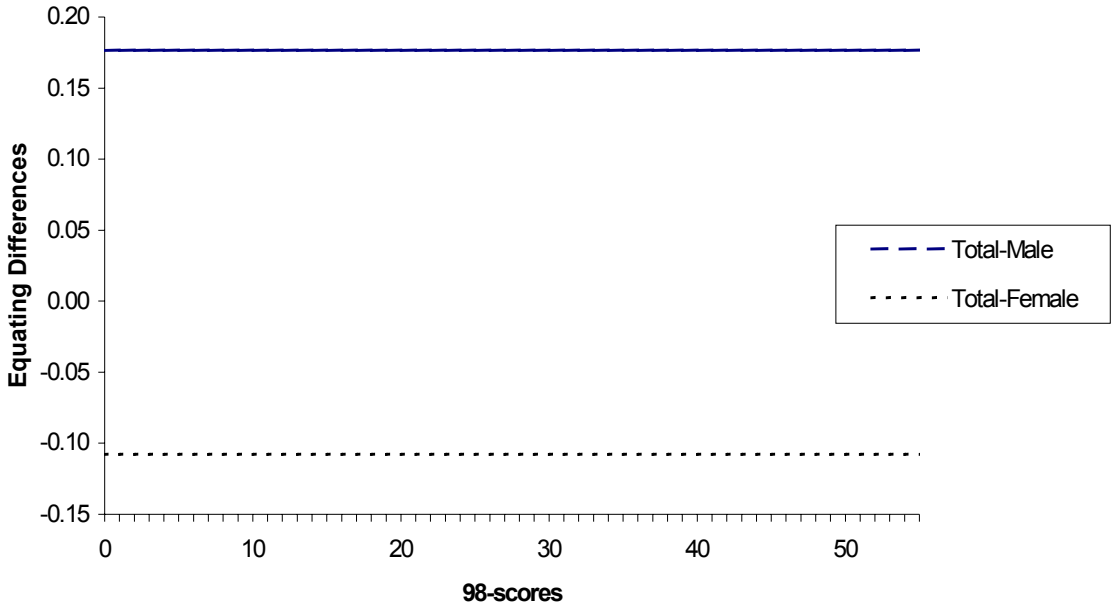

Figure 2. **The differences plots for the parallel linking functions (computed on the total group, and on the males and females groups) for the Tucker equating function on the data from the 1998 and 2000 administrations.**

Figures 3 and 4 illustrate the pseudo-NEAT design approach for both populations *P* and *Q*. Figure 3 shows the results for the 1998 administration and Figure 4 shows the results for the 2000 administration. We notice a significant difference between the RMSD values (from Equation 10) for the pseudo-NEAT design across years, the values for 1998 being larger than the values for 2000 for almost the whole score range. Although the RMSD values for the pseudo-NEAT designs are larger than the one for the original NEAT design (Figure 1), the values are still relatively small.

In order to facilitate the comparison between the linear and equipercentile procedures we reproduce the RMSD plots obtained from the pseudo-NEAT design approach in von Davier et al. (2003) in Figures 5 and 6.The same data were used in both studies. Von Davier et al. (2003) investigated two equipercentile methods within the NEAT design, the frequency estimation (the curvilinear equivalent of the Tucker method) and the chain equipercentile equating (the curvilinear equivalent of the chain linear method).

It is interesting to note the differences in shapes between the plots of the RMSD values for the equipercentile functions versus those of the RMSD values for the linear functions (or even among the linear functions). Our conjecture is that this difference reflects the shape of the equating

functions as well as the variation of the degree of population dependence across scores. More detailed studies on the RMSD index and its distribution are interesting topics for future research.

The range of the RMSD values (from Equation 10) for the linear and equipercentile methods is similar. For the 1998 administration, the range of RMSD for linear methods is slightly higher than for the equipercentile ones. As in von Davier et al. (2003), the regression-based methods (the Tucker and the frequency estimation methods) seem to be more sensitive to deviations from the population invariance assumption than the chain equating method.

The third linear function investigated in this study and for which there is no equipercentile counterpart, the Levine function, seems to vary less across subpopulations than the other methods; the Tucker method seems to vary the most.

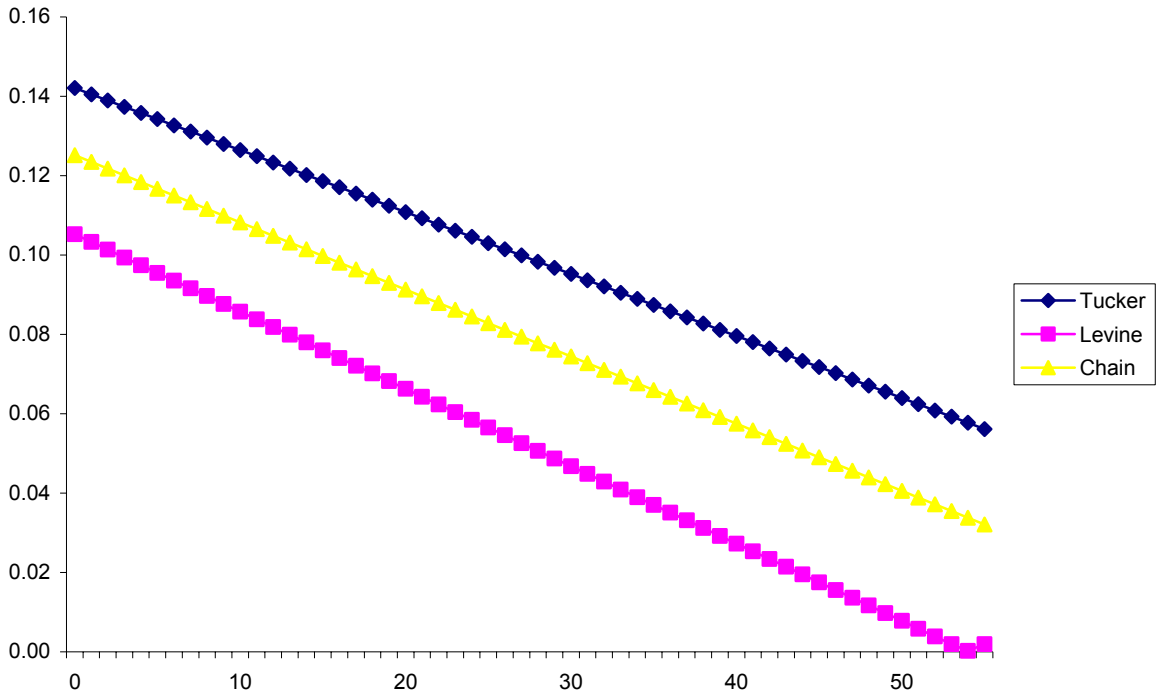

Figure 3. **The RMSD values for three linear equating functions on the data from the the pseudo-NEAT design from the 1998 administration.**

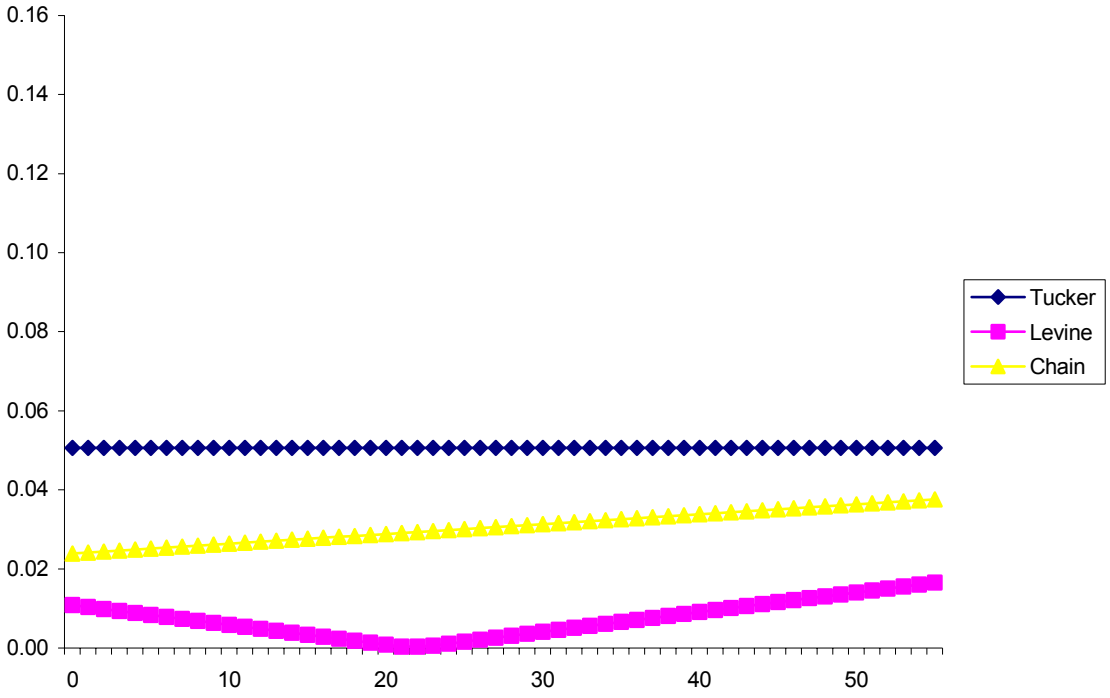

Figure 4. **The RMSD values for three linear equating functions on the data from the the pseudo-NEAT design from the 2000 administration.**

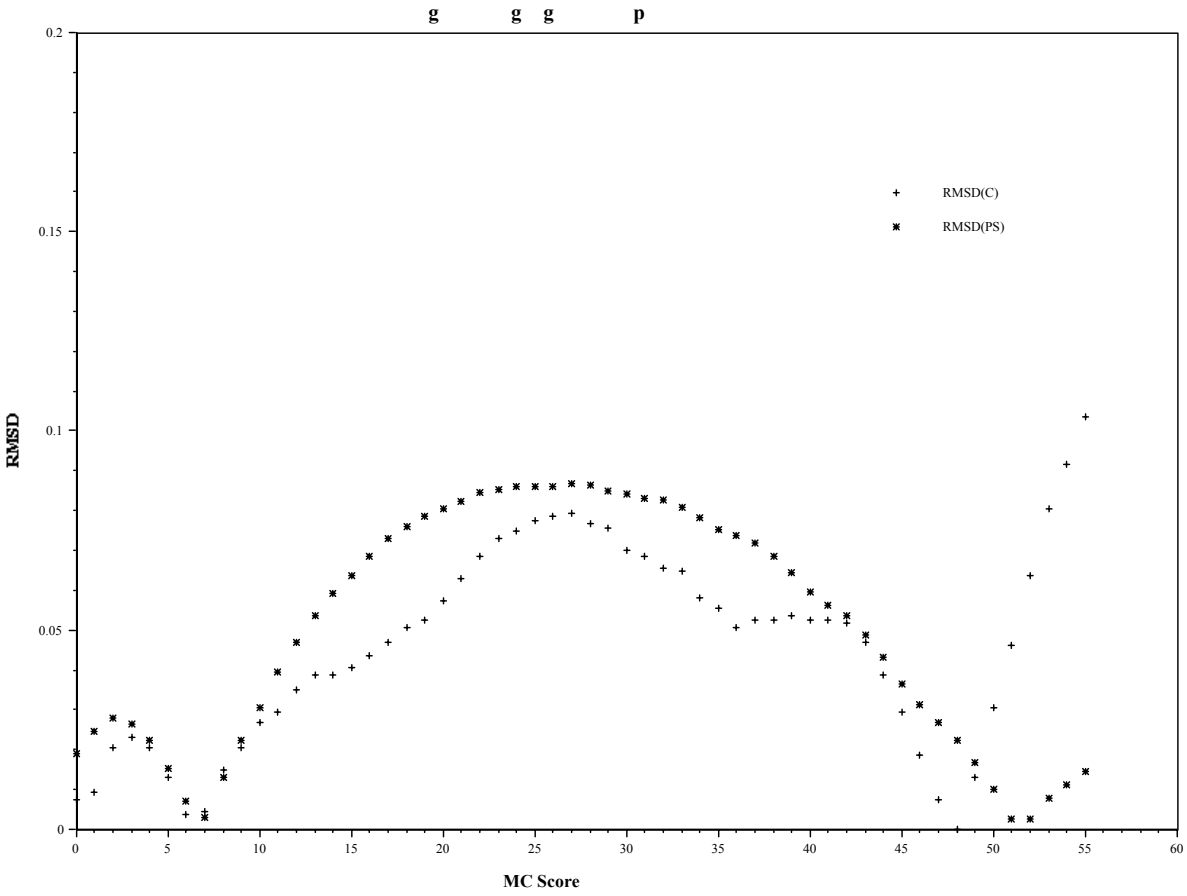

Figure 5. **The RMSD values for two equipercentile equating functions on the data from the pseudo- NEAT design from the 1998 administration (Figure 1 from von Davier et al., 2003).**

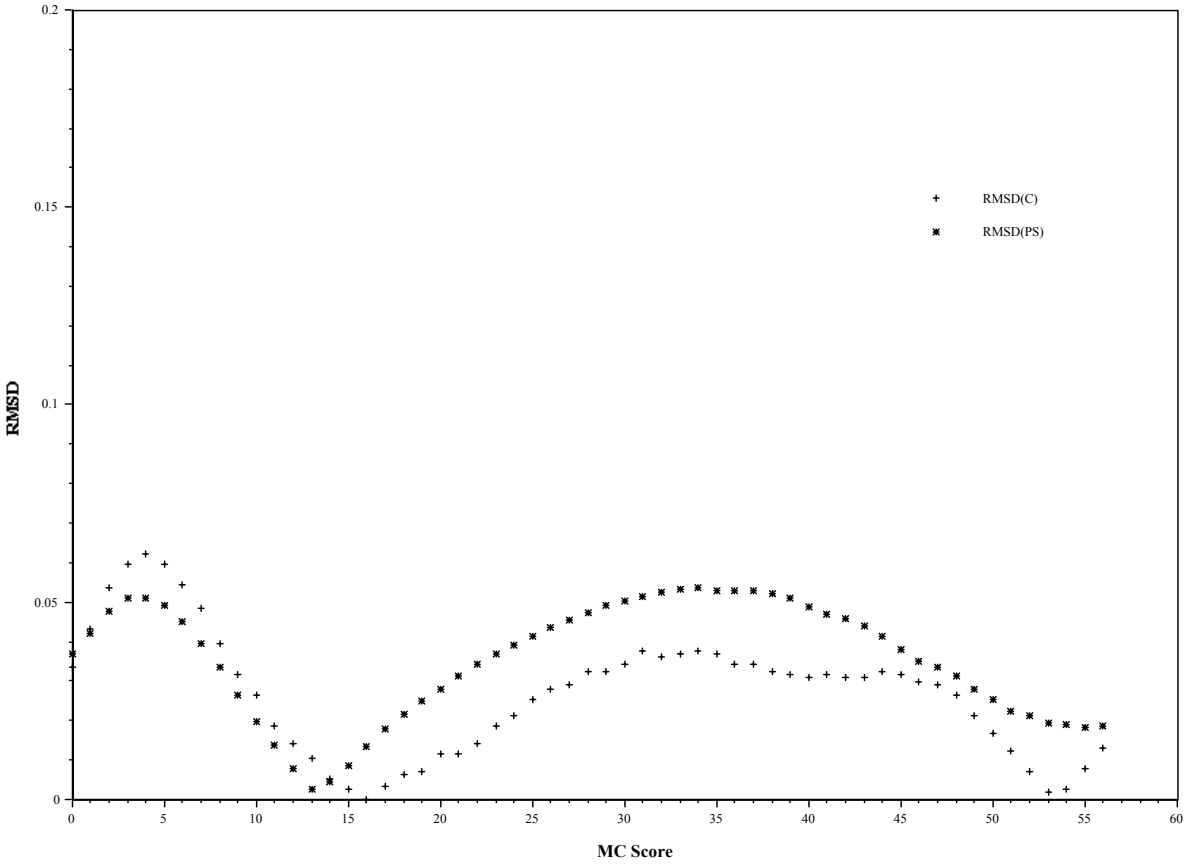

Figure 6. **The RMSD values for two equipercentile equating functions on the data from the pseudo- NEATdesign from the 2000 administration (Figure 2 from von Davier et al., 2003).**

Conclusions and Discussion

This study describes three different approaches for investigating the population sensitivity of the three commonly used linear equating methods under the NEAT design. The equating methods studied here are: the Tucker, the Levine, and the chain linear methods.

We applied the RMSD formula from (8) to each of the three equating functions. The RMSD values (for comparing the equating function computed on the total group with the functions computed on the subgroups) are considered to be small for many types of decisions regarding the equating process.

We also extended the system of parallel linear functions (Dorans and Holland, 2000) to the NEAT design. The three equating functions that correspond to the total population, the male population, and the female population are very close to each other for each of the investigated methods.

Finally, we applied the pseudo-NEAT design method developed in von Davier et al. (2003) to the linear equating functions. Kolen (2003) suggests that the pseudo-NEAT design approach might show less sensitivity for linear equating functions than for the equipercentile functions. This suggestion is based on the fact that the equipercentile equating functions require the estimation of the whole distribution of X and Y on T, while the linear equating function from (2) only requires the estimation of the means and standard deviations X and Y on T. However, when we compared our findings for the linear functions with those reported by von Davier et al. (2003) for nonlinear conversions, we found a similar degree of population sensitivity in linear equating as that reported in von Davier et al. (2003) for nonlinear conversions.

One of the implications of these findings is that the nontestable assumptions that underlie the equating methods in the NEAT design are met (or not) in a similar degree, for the linear and their counterpart equipercentile methods. However, the investigation of the assumptions done inside each population through the pseudo-NEAT design approach cannot be generalized to other populations or other assessments.

In conclusion, for this particular data set, all approaches that we investigated indicate that the three linear equating functions seem to be population invariant to an acceptable degree (and about as much as their nonlinear correspondent methods). For this example, the Levine function varied least across subpopulations; the Tucker method varied the most.

References

von Davier, A. A., Holland, P. W., & Thayer, D. T. (2003). Population invariance and chain versus post-stratification methods for equating and test linking. In N. Dorans (Ed.), *Population invariance of score linking: Theory and applications to Advanced Placement Program® examinations* (ETS RR-03-27, pp. 19–36). Princeton, NJ: ETS.

von Davier, A. A., Holland, P. W., & Thayer, D. T. (2004). The chain and post-stratification methods for observed-score equating: Their relationship to population invariance. In N. Dorans (Ed.), Population invariance of score linking: Theory and applications to Advanced Placement Program® examinations [Special issue]. *Journal of Educational Measurement, 41,* 15-32.

von Davier, A. A., & Kong, N. (in press). A unified approach to linear equating for the Non-Equivalent-groups design. *Journal of Educational and Behavioral Statistics*.

Dorans, N. J., & Holland, P. W. (2000). Population invariance and the equatability of tests: Basic theory and the linear case. *Journal of Educational Measurement, 37,* 281–306.

Dorans, N. J., Holland, P. W., Thayer, D. T., & Tateneni, K. (2003). Invariance of score linking across gender groups for three Advanced Placement Program® exams. In N. J. Dorans (Ed.), *Population invariance of score linking: Theory and applications to Advanced Placement Program® examinations* (ETS RR-03-27, pp. 19–36). Princeton, NJ: ETS.

Holland, P. W. (2002, April). *Overview of population invariance.* Paper presented at the meeting of the National Council on Measurement in Education, New Orleans, LA.

Kolen, M. J. (2003). Evaluating population invariance: A discussion of "Population invariance of score linking: Theory and applications to Advanced Placement Program® Examinations." In N. J. Dorans (Ed.), *Population invariance of score linking: Theory and applications to Advanced Placement Program® examinations* (ETS RR-03-27, pp. 19–36). Princeton, NJ: ETS.

Kolen, M. J., & Brennan, R. J. (2004). *Test equating, scaling, and linking: Methods and practices* (2nd ed.). New York: Springer.

Livingston, S. A., Dorans, N. J., & Wright, N. K. (1995). What combination of sampling and equating methods works best? *Applied Measurement in Education, 3*(1), 73–95.

Yang, W-L. (2004). Using subpopulation invariance to assess test score equity. In N. Dorans (Ed.), Population invariance of score linking: Theory and applications to Advanced Placement Program® examinations [Special issue]. *Journal of Educational Measurement, 41,* 15–32.