Research Report

Reliability and the Nonequivalent Groups With Anchor Test Design

Tim Moses

Sooyeon Kim

Research & Development

April 2007
RR-07-16

Reliability and the Nonequivalent Groups With Anchor Test Design

Tim Moses and Sooyeon Kim
ETS, Princeton, NJ

April 2007

As part of its educational and social mission and in fulfilling the organization's nonprofit charter and bylaws, ETS has and continues to learn from and also to lead research that furthers educational and measurement research to advance quality and equity in education and assessment for all users of the organization's products and services.

ETS Research Reports provide preliminary and limited dissemination of ETS research prior to publication. To obtain a PDF or a print copy of a report, please visit:

http://www.ets.org/research/contact.html

Abstract

This study evaluated the impact of unequal reliability on test equating methods in the nonequivalent groups with anchor test (NEAT) design. Classical true score-based models were compared in terms of their assumptions about how reliability impacts test scores. These models were related to treatment of population ability differences by different NEAT equating methods. A score model was then developed based on the most important features of the reviewed score models and used to study reliability in a simulation study across a total of 45 measurement conditions (= 5 test and anchor reliability combinations × 3 population ability difference conditions × 3 sample sizes). Ten equating methods were considered: chained linear, chained equipercentile with raw and smoothed frequencies, Tucker, frequency estimation equipercentile with raw and smoothed frequencies, Levine observed using Angoff-estimated and the "correct" reliabilities based on the data generation model used in this study, and Levine true using Angoff-estimated and correct reliabilities. The results were consistent with what is known about equating functions and their variability. Unequal and/or low reliability inflates equating function variability and alters equating functions when population abilities differ.

Key words: Reliability, NEAT equating design, classical true-score model, classical congeneric model, generalizability theory model, chained methods, conditioning methods, Levine equating method

Acknowledgments

The authors thank Michael Walker, Paul Holland, Brad Moulder, Neil Dorans, Bob Smith, and Dan Eignor for helpful comments on earlier versions of this paper. The authors also gratefully acknowledge the editorial assistance of Kim Fryer.

Table of Contents

List of Tables

Page

List of Figures

Page

Introduction

Reliability is often regarded as an important aspect of acceptable test equating. One of the basic requirements of test equating is that the test forms to be equated be equally reliable (Allen & Yen, 1979; Angoff, 1971; Dorans & Holland, 2000; Kolen & Brennan, 2004; Lord, 1980; Petersen, Kolen, & Hoover, 1989). High reliability is also desirable, though not usually described as a specific requirement of equating.

The focus of this paper is the impact of reliability on equating for the nonequivalent groups with anchor test (NEAT) design. Reliability has particularly important implications for NEAT equating, where the objective is to separately identify the contributions of examinee ability and test form difficulty on test scores in order to adjust test scores for form difficulty differences. Reliability is first described in terms of its assumed role in different test score models. Next, the score models are described in terms of which equating model appropriately accounts for the score models' population ability differences. Finally, reliability's effects on equating functions are illustrated in a series of simulations.

The Impact of Reliability on Scores

There are many models of test scores, and reliability is given different roles in each model. This section compares two classical true score-based score models in terms of the roles they assign to reliability. The simplifying assumption that the tests' and anchors' true scores are perfectly related is made throughout this discussion and the paper.

Classical True Score Theory

In classical true score theory, observed scores are modeled as the sum of "truth" and "error,"

$$X = T + E. \quad (1)$$

In (1), the expected score equals the true score, $\varepsilon(X) = \mu_X = T$, and T and E are independent so that their covariance, $\sigma(T,E)$, is zero. The independence assumption allows observed score variance to be expressed as the sum of true score and error variance,

$$\sigma^2(X) = \sigma^2(T) + \sigma^2(E). \quad (2)$$

Reliability is defined as the ratio of true score variance to observed score variance,

$$rel_X = \frac{\sigma^2(T)}{\sigma^2(T) + \sigma^2(E)}. \quad (3)$$

The two models reviewed next retain the main characteristics of classical true score theory ($\varepsilon(X) = T$ and $\sigma(T,E) = 0$). The unique and reliability-relevant aspects of these models are in how they structure T and E to specify test form difficulty and examinee ability effects (Table 1).

Table 1

Summary of Score Models and the Roles These Models Give to Reliability and Examinee Ability

Model	Test score = [Test difficulty] + [Examinee ability] + [Error (unreliability)]
Classical congeneric model	$X = [\ \delta_X\ ] + [\ \sqrt{rel_X}\frac{\sigma(X)}{\sigma(T)}T\ ] + [\ E_X\ ]$
Generalizability theory	$X = [\ \sum_i \upsilon_i\ ] + [\ n_i\, t\ ] + [\ \sum_i \upsilon_{ti}\ ]$

Classical Congeneric Models

Classical congeneric models (Brennan, 1990; Feldt & Brennan, 1989) specify the contributions of difficulty, ability, and reliability on congeneric test and anchor scores. Tests (X) and anchors (A) are modeled as

$$X = (T_X) + (E_X) = (\lambda_X T + \delta_X) + (E_X),$$

$$A = (T_A) + (E_A) = (\lambda_A T + \delta_A) + (E_A). \quad (4)$$

The λ terms are defined as *effective test lengths* (Brennan, 1990) for which this paper's discussion makes use of a definition of effective test length as a reliability-dependent true score standard deviation (Angoff, 1953, 1971, pp. 114–115; Kolen & Brennan, 2004, pp. 112–113),

$$\begin{aligned}\sigma(T_X) &= \sqrt{rel_X}\,\sigma(X)\\ &= \sqrt{\frac{\lambda_X^2\sigma^2(T)}{\lambda_X^2\sigma^2(T) + \sigma^2(E_X)}}\sqrt{\lambda_X^2\sigma^2(T) + \sigma^2(E_X)}\,,\\ &= \lambda_X\sigma(T)\end{aligned}$$

so that $\lambda_X = \sqrt{rel_X}\,\frac{\sigma(X)}{\sigma(T)}$. T is the test taker true score and underlies the tests' and anchors' true scores. The test and anchor are congeneric, meaning that their true scores, $T_X = \sqrt{rel_X}\,\frac{\sigma(X)}{\sigma(T)}T + \delta_X$ and $T_A = \sqrt{rel_A}\,\frac{\sigma(A)}{\sigma(T)}T + \delta_A$, are perfectly related. The δ terms in (4) are constants that determine the difficulty or ease of the test and anchor. The E terms have expectations of zero and are independent of T. Test and anchor reliabilities in the classical congeneric model are

$$rel_X = \frac{rel_X \frac{\sigma^2(X)}{\sigma^2(T)}\sigma^2(T)}{rel_X \frac{\sigma^2(X)}{\sigma^2(T)}\sigma^2(T) + \sigma^2(E_X)} = \frac{rel_X \sigma^2(X)}{rel_X \sigma^2(X) + \sigma^2(E_X)} = \frac{\sigma^2(T_X)}{\sigma^2(T_X) + \sigma^2(E_X)},$$

$$rel_A = \frac{rel_A \frac{\sigma^2(A)}{\sigma^2(T)}\sigma^2(T)}{rel_A \frac{\sigma^2(A)}{\sigma^2(T)}\sigma^2(T) + \sigma^2(E_A)} = \frac{rel_A \sigma^2(A)}{rel_A \sigma^2(A) + \sigma^2(E_A)} = \frac{\sigma^2(T_A)}{\sigma^2(T_A) + \sigma^2(E_A)}. \quad (5)$$

The classical part of classical congeneric theory is the assumption that error variances are proportional to effective test length:

$$\sigma^2(E_X) = \lambda_X \sigma^2(E) = \sqrt{rel_X}\,\frac{\sigma(X)}{\sigma(T)}\sigma^2(E),$$

$$\sigma^2(E_A) = \lambda_A \sigma^2(E) = \sqrt{rel_A}\,\frac{\sigma(A)}{\sigma(T)}\sigma^2(E). \quad (6)$$

The classical congeneric model has two important implications for scores and equating. First, the proportionality of the error variances across the tests and anchors in (6) allows reliability and the test and anchor error variances to be estimated from the observed variances and correlations of the test and anchor scores (Angoff, 1953). Second, while the mean observed scores (μ_X) are equal to mean true scores (μ_{TX}), the population ability effect (μ_T) on mean observed scores is directly influenced by reliability,

$$\mu_X = \mu_{TX} = \sqrt{rel_X}\,\frac{\sigma(X)}{\sigma(T)}\mu_T + \delta_X\,,$$

$$\mu_A = \mu_{TA} = \sqrt{rel_A}\,\frac{\sigma(A)}{\sigma(T)}\mu_T + \delta_A\,. \qquad (7)$$

For classical congeneric models, the role of unreliability on test scores is not only to influence the proportion of true score to observed score variance. Unreliability also biases the extent to which overall test taker abilities are visible on observed scores.

Generalizability Theory

Generalizability theory extends classical true score theory by using analysis of variance models to separately identify different sources of the error left undifferentiated by classical true score theory (Brennan, 2001; Cronbach, Gleser, Nanda, & Rajaratnam, 1972). One of the simplest designs is sufficient for identifying the role of reliability on scores of test takers (*t*) sampled from some population of interest-taking items (*i*) sampled from a universe of admissible items. The score on item *i* for test taker *t* (X_{ti}) is modeled as

$$X_{ti} = t + \nu_i + \nu_{ti}. \qquad (8)$$

The *t* reflects test taker *t*'s ability and has expectation $\varepsilon(t) = \mu_t = \varepsilon_t\varepsilon_i\left(X_{ti}\right)$. The effect ν_i is interpretable as the influence of an easier or more difficult item, which introduces absolute error when decisions are made based on the absolute values of observed scores (e.g., classifications with respect to a cut-score). The effect ν_{ti} is the interaction of test takers with items that is confounded with all other sources of error, which introduces relative error for decisions based on the relative standing of test takers on their observed scores.

The error effects have zero expectations,

$$\varepsilon_i(\nu_i) = \varepsilon_t(\nu_{ti}) = \varepsilon_i(\nu_{ti}) = 0, \qquad (9)$$

and are assumed to be uncorrelated with the other terms in the model and the effects of other items ($i' \neq i$),

$$\sigma(t\nu_{ti}) = \sigma(\nu_i\nu_{ti}) = \sigma(t\nu_i) = \sigma(\nu_i\nu_{i'}) = \sigma(\nu_{ti}\nu_{ti'}) = 0. \qquad (10)$$

To emphasize how the generalizability theory model in (8) relates to the classical congeneric model in (4), (8) can be used to express test takers' scores (i.e., the sums of their n_i item scores) on tests ($X_{t.}$) and anchors ($A_{t.}$) that are based on items (not necessarily the same) sampled from a common universe:

$$X_{t.} = \sum_{i,X}^{n_{i,X}} X_{ti} = \sum_{i,X}^{n_{i,X}} (t + \nu_{i,X} + \nu_{ti,X}) = n_{i,X}(t) + \sum_{i,X}^{n_{i,X}} \nu_{i,X} + \sum_{i,X}^{n_{i,X}} \nu_{ti,X} \ ,$$

$$A_{t.} = \sum_{i,A}^{n_{i,A}} A_{ti} = \sum_{i,A}^{n_{i,A}} (t + \nu_{i,A} + \nu_{ti,A}) = n_{i,A}(t) + \sum_{i,A}^{n_{i,A}} \nu_{i,A} + \sum_{i,A}^{n_{i,A}} \nu_{ti,A} \ . \qquad (11)$$

From (11), examinees take both X and A, so that test taker variance (i.e., true score variance) contributes to the observed variances of both X and A. Because the items on X and A are parallel, item effects for items in X ($\nu_{i,X}$ and $\nu_{ti,X}$) and A ($\nu_{i,A}$ and $\nu_{ti,A}$) both have variances $\sigma^2(i)$ and $\sigma^2(ti)$. Observed score variances for the scores in (11) are therefore defined as

$$\sigma^2(X_{t.}) = n_{i,X}^2 \left(\sigma^2(t) + \frac{\sigma^2(i)}{n_{i,X}} + \frac{\sigma^2(ti)}{n_{i,X}} \right),$$

$$\sigma^2(A_{t.}) = n_{i,A}^2 \left(\sigma^2(t) + \frac{\sigma^2(i)}{n_{i,A}} + \frac{\sigma^2(ti)}{n_{i,A}} \right). \qquad (12)$$

Test and anchor reliabilities are defined as

$$rel_X = \frac{\sigma^2(t)}{\sigma^2(t) + \frac{\sigma^2(ti)}{n_{i,X}}},$$

$$rel_A = \frac{\sigma^2(t)}{\sigma^2(t) + \frac{\sigma^2(ti)}{n_{i,A}}}. \qquad (13)$$

The reliability coefficients in (13) are identical to coefficient alpha and, for dichotomously-scored items, the KR-20 reliability coefficient. Classical definitions of reliability focus

exclusively on relative error (ν_{ti}) rather than absolute error (ν_i) for defining error variance. Reliabilities across congeneric tests and anchors differ only with respect to test and anchor lengths, a fundamental assumption that follows from the test taker population and item universe that is used in the decision studies that typically accompany generalizability analyses.

Comparing the Classical Congeneric and Generalizability Theory Models

Important distinctions exist between the classical congeneric and generalizability theory models that are relevant for equating. First, the source of test form difficulty differences is different for each model. For generalizability theory models, test form difficulty differences are assumed to be due to random samples of items that do not always have mean difficulties that converge to their expected value of zero (i.e., while $\varepsilon_i(\nu_i) = 0$, $\sum_i^{n_i} \nu_i$ may not equal zero for every sample of items and therefore $\sigma^2(i)$ is not necessarily equal to zero). In classical congeneric models, test form difficulty differences are described as systematic difficulty effects rather than as sampling effects (i.e., the δ terms are defined as constants rather than as random variables). In addition, reliability across tests and anchors composed of parallel items in generalizability models does not directly affect the scores' extent of ability effects (changes in $\sigma^2(t)$ and $\sigma^2(ti)$ do not necessarily affect $\varepsilon_t(X_{t.}) = n_{i,X}\mu_t$), while in classical congeneric models, reliability has a direct effect on a score's extent of ability effects ($\mu_X = \mu_{TX} = \sqrt{rel_X}\frac{\sigma(X)}{\sigma(T)}\mu_T + \delta_X$).

The Impact of Reliability on Equating

The purpose of equating is to adjust the scores of test forms that are intended to be parallel for unintended differences in difficulty. NEAT equating matches nonequivalent administration groups on their ability, where ability differences are estimated from mean anchor score differences. When one of two test forms (*X* or *Y*) is given to an independent sample of one of two populations (*P* or *Q*) along with an anchor test (*A*), X_P is equated to test Y_Q, and the anchor scores (A_P and A_Q) are used to account for ability differences in the populations. The following presentation focuses on contrasting the major equating methods' treatment of population ability differences when scores are unreliable and follow either the classical congeneric model or the generalizability theory model. Some previous works have compared the equating methods that directly informed this section (Holland, 2004; Kolen & Brennan, 2004),

and other works have informed this section's relating of the classical congeneric model to Levine equating (Brennan, 1990; Hanson, 1991).

Four common linear equating methods in the NEAT design (Tucker, chained linear, Levine observed, and Levine true) all incorporate ability differences between populations *P* and *Q* in the X_P–to-Y_Q equating function. Ability difference information is expressed as

$$\gamma_{ey} \frac{\sigma(Y_Q)}{\sigma(A_Q)}(\mu_{AP} - \mu_{AQ}) . \tag{14}$$

Here, (14) shows that the population ability difference observed in *P* and *Q*'s anchor means is standardized according to A_Q's variability and scaled to Y_Q's variability. The γ_{ey} is a term that is specific to each equating method and describes the unique way an equating method scales mean anchor score differences to Y_Q.

Here, (14) is part of the chained linear equating function,

$$cl_Y(X_P) = \frac{\sigma(Y_Q)}{\sigma(A_Q)}\frac{\sigma(A_P)}{\sigma(X_P)}(X_P - \mu_{XP}) + \mu_{YQ} + \frac{\boldsymbol{\sigma(Y_Q)}}{\boldsymbol{\sigma(A_Q)}}\boldsymbol{(\mu_{AP} - \mu_{AQ})} , \tag{15}$$

and the Levine true equating function,

$$lt_Y(X_P) = \frac{\sqrt{rel_{YQ}}}{\sqrt{rel_{AQ}}}\frac{\sigma(Y_Q)}{\sigma(A_Q)}\frac{\sqrt{rel_{AP}}}{\sqrt{rel_{XP}}}\frac{\sigma(A_P)}{\sigma(X_P)}(X_P - \mu_{XP}) + \mu_{YQ} + \frac{\sqrt{\boldsymbol{rel_{YQ}}}}{\sqrt{\boldsymbol{rel_{AQ}}}}\frac{\boldsymbol{\sigma(Y_Q)}}{\boldsymbol{\sigma(A_Q)}}\boldsymbol{(\mu_{AP} - \mu_{AQ})} . \tag{16}$$

Also, (14) is used to estimate the mean of the unobserved Y_P for the Levine observed method

$$\mu_{YP} = \mu_{YQ} + \boldsymbol{w}\frac{\sqrt{\boldsymbol{rel_{YQ}}}}{\sqrt{\boldsymbol{rel_{AQ}}}}\frac{\boldsymbol{\sigma(Y_Q)}}{\boldsymbol{\sigma(A_Q)}}\boldsymbol{(\mu_{AP} - \mu_{AQ})} . \tag{17}$$

Finally, (14) is used to estimate the mean of the unobserved $Y_{wP+(1-w)Q}$ for the Tucker method, expressed here with a $Y_Q A_Q$ correlation (ρ_{YQAQ}) based on the assumption of congeneric tests and anchors ($\rho_{YQAQ} = \sqrt{rel_{YQ} rel_{AQ}}$)

$$\mu_{YwP+(1-w)Q} = \mu_{YQ} + \boldsymbol{w}\sqrt{\boldsymbol{rel_{YQ}rel_{AQ}}}\,\frac{\boldsymbol{\sigma(Y_Q)}}{\boldsymbol{\sigma(A_Q)}}(\boldsymbol{\mu_{AP} - \mu_{AQ}}) . \qquad (18)$$

Scaling Population Ability Differences in Terms of Observed Scores

The chained linear method utilizes an observed score scaling of the population ability differences in (14) by setting γ_{ey} =1 in (15). The chained method's use of observed score variance rather than true score variance has the advantage of simplicity and of dealing with directly observable variances. When data follow a generalizability theory model, the observed score scaling is defensible because the observed score means are equal to average true ability ($\varepsilon_t\varepsilon_i\left(X_{ti}\right)=\mu_t$ in (8)). Observed score scaling of the difference in mean A_P ($\mu_{AP} = n_{iA}\mu_{tP}$) and mean A_Q ($\mu_{AQ} = n_{iA}\mu_{tQ}$) can be expressed as

$$\gamma_{ey}\frac{\sigma(Y_Q)}{\sigma(A_Q)}(\mu_{AP} - \mu_{AQ}) = (1)\frac{n_{iY}\sqrt{\left(\sigma^2(tQ) + \frac{\sigma^2(iY)}{n_{iY}} + \frac{\sigma^2(tQiY)}{n_{iY}}\right)}}{n_{iA}\sqrt{\left(\sigma^2(tQ) + \frac{\sigma^2(iA)}{n_{iA}} + \frac{\sigma^2(tQiA)}{n_{iA}}\right)}}(n_{iA}\mu_{tP} - n_{iA}\mu_{tQ}) . \qquad (19)$$

When data follow a classical congeneric model, the chained linear method's γ_{ey} =1 results in

$$\gamma_{ey}\frac{\sigma(Y_Q)}{\sigma(A_Q)}(\mu_{AP} - \mu_{AQ}) = (1)\left(\frac{\sigma(Y_Q)}{\sigma(A_Q)}\right)(\sqrt{rel_{AP}}\,\frac{\sigma(A_P)}{\sigma(T_P)}\mu_{TP} + \delta_A - \sqrt{rel_{AQ}}\,\frac{\sigma(A_Q)}{\sigma(T_Q)}\mu_{TQ} - \delta_A),$$

$$= \left(\frac{\sigma(Y_Q)}{\sigma(A_Q)}\right)(\sqrt{rel_{AP}}\,\frac{\sigma(A_P)}{\sigma(T_P)}\mu_{TP} - \sqrt{rel_{AQ}}\,\frac{\sigma(A_Q)}{\sigma(T_Q)}\mu_{TQ}) . \qquad (20)$$

If the anchor scores are very unreliable in a classical congeneric model, chained linear method's observed score scaling of ability differences will reflect a biased estimate of true population ability differences, $\left[\sqrt{rel_{AP}}\,\frac{\sigma(A_P)}{\sigma(T_P)}\mu_{TP} - \sqrt{rel_{AQ}}\,\frac{\sigma(A_Q)}{\sigma(T_Q)}\mu_{TQ}\right] \le \left[\frac{\sigma(A_P)}{\sigma(T_P)}\mu_{TP} - \frac{\sigma(A_Q)}{\sigma(T_Q)}\mu_{TQ}\right]$.

Scaling Population Ability Differences in Terms of True Scores

The Levine true and observed methods utilize a true score scaling of the population ability differences in (14), meaning that they set γ_{ey} equal to the ratio of Y_Q and A_Q's root reliabilities, $\frac{\sqrt{rel_{YQ}}}{\sqrt{rel_{AQ}}}$ in (16) and (17). The true score scaling of ability differences is especially defensible when the data follow a classical congeneric model. True score scaling of ability differences can be justified when the data follow a generalizability theory model under limited conditions.

When the data follow a classical congeneric model, true score scaling in (14) results in

$$= \left(\frac{\sqrt{rel_{YQ}}}{\sqrt{rel_{AQ}}}\right)\left(\frac{\sigma(Y_Q)}{\sigma(A_Q)}\right)\left(\sqrt{rel_{AP}}\,\frac{\sigma(A_P)}{\sigma(T_P)}\mu_{TP} - \sqrt{rel_{AQ}}\,\frac{\sigma(A_Q)}{\sigma(T_Q)}\mu_{TQ}\right). \qquad (21)$$

When $rel_{AP} = rel_{AQ}$ and $\sigma(A_P) = \sigma(A_Q)$, (21) can be written directly in terms of Y_Q's true score variance as

$$\sqrt{rel_{YQ}}\,\sigma(Y_Q)\left(\frac{\mu_{TP}}{\sigma(T_P)} - \frac{\mu_{TQ}}{\sigma(T_Q)}\right). \qquad (22)$$

When the data follow a generalizability theory model, true score standard deviations are the product of test length and the test taker variance component ($n_{iY}\sigma(tQ)$ and $n_{iA}\sigma(tQ)$, as in (12), so that true score scaling essentially involves the lengths of A_Q (n_{iA}) and Y_Q (n_{iY}),

$$\left(\frac{\sqrt{rel_{YQ}}}{\sqrt{rel_{AQ}}}\right)\left(\frac{\sigma(Y_Q)}{\sigma(A_Q)}\right)(\mu_{AP} - \mu_{AQ}) = \frac{\sqrt{\dfrac{\sigma^2(tQ)}{\sigma^2(tQ) + \dfrac{\sigma^2(tQi)}{n_{iY}}}}\; n_{iY}\sqrt{\sigma^2(tQ) + \dfrac{\sigma^2(i)}{n_{iY}} + \dfrac{\sigma^2(tQi)}{n_{iY}}}}{\sqrt{\dfrac{\sigma^2(tQ)}{\sigma^2(tQ) + \dfrac{\sigma^2(tQi)}{n_{iA}}}}\; n_{iA}\sqrt{\sigma^2(tQ) + \dfrac{\sigma^2(i)}{n_{iA}} + \dfrac{\sigma^2(tQi)}{n_{iA}}}}(n_{iA}\mu_{tP} - n_{iA}\mu_{tQ}). \qquad (23)$$

When n_{iA} and n_{iY} are large and/or $\sigma^2(i)$ is relatively small (i.e., items and forms are long and/or do not differ widely in difficulty), $\sqrt{\sigma^2(t) + \dfrac{\sigma^2(i)}{n_i} + \dfrac{\sigma^2(ti)}{n_i}} \approx \sqrt{\sigma^2(t) + \dfrac{\sigma^2(ti)}{n_i}}$, so that (23) becomes

$$= \frac{n_{iY}\sigma(tQ)}{n_{iA}\sigma(tQ)}(n_{iA}\mu_{tP} - n_{iA}\mu_{tQ}) = n_{iY}(\mu_{tP} - \mu_{tQ}) \; .$$

When scaling according to true score variance in generalizability theory, the ability difference is scaled using the ratios of the *actual* lengths of Y_Q and A_Q ($\frac{n_{iY}}{n_{iA}}$).

When scaling according to true score variance in classical congeneric theory, the ability difference is scaled using the ratios of the *effective* lengths of Y_Q and A_Q ($\frac{\sqrt{rel_{YQ}}\sigma(Y_Q)}{\sqrt{rel_{AQ}}\sigma(A_Q)}$). For both score models, the average true score differences between the populations are potentially scalable according to how each theory defines the true score scale of Y_Q.

Using Population Ability Differences in an Observed-Score Regression

Tucker equating assumes that the linear regressions of observed test scores on observed anchor scores are test- and anchor-specific rather than population-dependent. Assumptions about true scores and errors do not directly inform the linear regression used by the Tucker method, though some correspondence to true score theory can be observed by noting that the correlation between congeneric tests and anchors is expressible in terms of test and anchor reliabilities. An estimate of the mean of $Y_{wP+(1-w)Q}$, $\mu_{YwP+(1-w)Q}$, can be obtained, as in (18), by applying the synthetic population-weighted $Y_Q|A_Q$ regression at score μ_{AP}. When unreliability weakens the $Y_Q|A_Q$ regression, it essentially discounts the extent to which anchor score mean differences are incorporated in the Tucker equating function. This discounting is very different from the way the chained linear and Levine methods utilize the mean ability difference information in (14), and it is inconsistent with how reliability is assumed to affect scores generated from classical congeneric and generalizability theory models.

Unreliability and Equating Bias

From the previous section's discussion, the impact of unreliability on equating bias can be understood as a misinterpretation of the anchor score information (i.e., using an equating method that incorrectly scales μ_{AP}-μ_{AQ}), as in (14). For example, if reliability's effect was only on true score and error variances and not on ability effects, the Levine method would assume that true ability differences were bigger than the observed μ_{AP}-μ_{AQ} and overmatch on the ability

difference by setting $\gamma_{ey} = \frac{\sqrt{rel_{YQ}}}{\sqrt{rel_{AQ}}}$. In contrast to the Levine method, the Tucker method would incorrectly discount the observed μ_{AP}-μ_{AQ} and undermatch for μ_{AP}-μ_{AQ} by applying a regression and setting $\gamma_{ey} = w\sqrt{rel_{YQ}rel_{AQ}}$.

If reliability's effect was to bias the true ability difference information observed in μ_{AP}-μ_{AQ} (as in classical congeneric models), then the chained linear and Tucker methods would undermatch on the true ability in μ_{AP}-μ_{AQ}. The Tucker method would undermatch more so than the chained linear method because the former's setting $\gamma_{ey} = w\sqrt{rel_{YQ}rel_{AQ}}$ would incorrectly reduce the observed ability difference more than the chained linear method's setting $\gamma_{ey} = 1$. If the observed test and anchor scores followed population-invariant linear regression models, Tucker would correctly utilize the regressions, and the chained linear and Levine methods would incorrectly overmatch on μ_{AP}-μ_{AQ}, the Levine method more so than the chained linear one because the Levine method's incorrect setting $\gamma_{ey} = \frac{\sqrt{rel_{YQ}}}{\sqrt{rel_{AQ}}}$ will likely magnify μ_{AP}-μ_{AQ} while the chained linear method's setting $\gamma_{ey} = 1$ will be closer to the Tucker method's $\gamma_{ey} = w\sqrt{rel_{YQ}rel_{AQ}}$.

Unreliability and Equating Standard Errors

The impact of reliability on equating variability can also be understood in terms of the different equating methods' versions of (14). Specifically, μ_{AP}-μ_{AQ} has a sampling variance that will impact equating standard errors. The Tucker method's tendency to downweight μ_{AP}-μ_{AQ} based on the test-anchor correlation would also reduce equating standard errors. The Levine method's tendencies to magnify μ_{AP}-μ_{AQ} based on the ratio of test and anchor root reliabilities would magnify equating standard errors. These statements correspond to previous findings of the relative ordering of equating function standard errors, where approaches that use the anchor as a conditioning variable (Tucker and frequency estimation equipercentile) are less variable than approaches that use the anchor to form a chain of links between two test forms (chained linear and chained equipercentile), which are in turn less variable than the Levine approaches (von Davier, Holland, & Thayer, 2004; von Davier & Kong, 2005; Kolen & Brennan, 2004; Wang, Lee, Brennan, & Kolen, 2006). To the extent that unreliability makes equating functions more

variable, it should do so more for the Levine and chained approaches than for the conditioning approaches.

Method

The issues of reliability on test equating results were explored in a simulation study. A data generation model was developed to reflect the following main features of classical true score theory, classical congeneric models, and generalizability theory applied to congeneric tests and anchors:

- Observed scores are the sum of test-taker truth plus error.
- The expected value of the observed scores is the true score.
- The true scores are independent of error.
- The correlation of the test and anchor true scores is 1.

The data generation model was also developed in order to manipulate reliability independently of other test score features, including observed score variances, true population ability differences, observed lengths, and no-test-form difficulty differences. This conception of reliability was the basis for studying how equating method averages and standard errors were impacted across combinations of sample size, reliability, equating method, and population ability difference.

Data Generation Model

The X_P, A_P, A_Q, and Y_Q scores were generated as sums of independently and normally distributed truth and error variables

$$\begin{aligned} X_P &= T_{XP} + E_{XP}, \\ A_P &= T_{AP} + E_{AP}, \\ A_Q &= T_{AQ} + E_{AQ}, \\ Y_Q &= T_{YQ} + E_{YQ}. \end{aligned} \tag{24}$$

In (24), the true scores in population P (T_{XP} and T_{AP}) and in population Q (T_{AQ} and T_{YQ}) were perfectly correlated, differing only in their variances ($\sigma^2_{TXP} \neq \sigma^2_{TAP}$ and $\sigma^2_{TYQ} \neq \sigma^2_{TAQ}$). The mean test and anchor scores were functions of actual lengths,

($\varepsilon(X_P) = \varepsilon(T_{XP}) = n_{XP}\mu_P$, $\varepsilon(A_P) = \varepsilon(T_{AP}) = n_{AP}\mu_P$, $\varepsilon(Y_Q) = \varepsilon(T_{YQ}) = n_{YQ}\mu_Q$, and $\varepsilon(A_Q) = \varepsilon(T_{AQ}) = n_{AQ}\mu_Q$). The test lengths (n_{XP} and n_{YQ}) were set equal to 100 and the anchor lengths (n_{AP} and n_{AQ}) were set equal to 30. The anchors were external to the tests. The variances of the true scores and error scores in (24) were manipulated to produce test and anchor scores of desired reliability levels (e.g., $rel_{XP} = \frac{\sigma^2(T_{XP})}{\sigma^2(T_{XP}) + \sigma^2(E_{XP})}$, Equation 3), while achieving desired observed score variances (e.g., $\sigma^2(X_P) = \sigma^2(T_{XP}) + \sigma^2(E_{XP})$, Equation 2). The observed score variances were kept equal for the tests ($\sigma^2_{XP} = \sigma^2_{YQ}$) and the anchors ($\sigma^2_{AP} = \sigma^2_{AQ}$). To also consider equipercentile methods, the final scores were rounded to integer units and truncated into desired score ranges. The means and variances of the X_P, A_P, A_Q, and Y_Q scores in (24) are summarized in Table 2.

There are several implications of the data generation model in (24):

- The standardized difference in anchor score means was equal to the standardized difference in the test means throughout the study. The equality of test and anchor standardized mean differences was an operationalized definition of no difficulty differences across test forms *X* and *Y*. In other words, no systematic difficulty effects were built into (24) and no standardized mean differences on test forms *X* and *Y* were present that could not also be observed in the standardized mean differences on the anchor.

- The data generation model corresponded to the chained linear method's observed score focus. Note that the expression of the chained linear equating function (15) based on this study's data generation model's constraints resulted in the identity equating function, the equating function that would be appropriate when test forms have no true difficulty differences:

$$cl_Y(X_P) = \frac{\sigma(Y_Q)}{\sigma(A_Q)}\frac{\sigma(A_P)}{\sigma(X_P)}(X_P - \mu_{XP}) + \mu_{YQ} + \frac{\sigma(Y_Q)}{\sigma(A_Q)}(\mu_{AP} - \mu_{AQ}) ,$$

$$= (X_P - \mu_{XP}) + \mu_{YQ} + \frac{\sigma(Y_Q)}{\sigma(A_Q)}(\mu_{AP} - \mu_{AQ}) ,$$

because the test variances were kept equal and the anchor variances were also kept equal, and

$$= X_P + \mu_{YQ} - \mu_{XP} + \frac{\sigma(Y_Q)}{\sigma(A_Q)}(\mu_{AP} - \mu_{AQ}),$$

$$= X_P, \text{ because } \frac{\mu_{YQ} - \mu_{XP}}{\sigma(Y_Q)} = \frac{\mu_{AQ} - \mu_{AP}}{\sigma(A_Q)}.$$

Table 2

Data Collection Design (Nonequivalent Groups With an Anchor Test Design), Equation 24, Where Score = T + E

Population	Score	*N*	μ (= ε(*T*)/*n*)	Score mean (= *n*μ)	Score SD[a]
Population standardized ability difference = 0					
P	*X*	100	.5000	50.000	18.0
P	*A*	30	.5000	15.000	5.4
Q	*A*	30	.5000	15.000	5.4
Q	*Y*	100	.5000	50.000	18.0
Population standardized ability difference = .15					
P	*X*	100	.5135	51.350	18.0
P	*A*	30	.5135	15.405	5.4
Q	*A*	30	.4865	14.595	5.4
Q	*Y*	100	.4865	48.650	18.0
Population standardized ability difference = .30					
P	*X*	100	.5270	52.700	18.0
P	*A*	30	.5270	15.810	5.4
Q	*A*	30	.4730	14.190	5.4
Q	*Y*	100	.4730	47.300	18.0

[a] The score standard deviation is determined as $\sqrt{\sigma^2(T) + \sigma^2(E)}$, where $\sigma^2(T)$ and $\sigma^2(E)$ are determined to obtain a desired reliability and observed score standard deviation.

The manipulation of test and anchor reliabilities, which are independent of test and anchor lengths, and of observed score variances was inconsistent with the assumptions of generalizability theory and the classical congeneric models. This inconsistency was deliberately created to set up situations where the Levine equating method using the classical congeneric-based Angoff (1953) reliability estimates could be studied when the Angoff reliability estimates were incorrect.

Population Standardized Ability Differences

Three population standardized ability difference conditions were defined as standardized mean differences on the anchor scores for the *P* and *Q*. This study considered population standardized mean differences of 0, .15, and .30. For nonzero population standardized mean differences, *P* was more able than *Q*.

Reliability

Five combinations of test and anchor reliabilities are presented in Table 3, where reliabilities ranged from high (.9 and .8), medium (.7 and .6), to very low (.5 and .4). The reliabilities of anchors A_P and A_Q were always equal and always less than the reliabilities of tests X_P and Y_Q. Two reliability combinations were such that the reliabilities of X_P and Y_Q were unequal (reliability combinations of $rel_{XP}\_rel_{AP}\_rel_{AQ}\_rel_{YQ}$ =.9_.5_.5_.7 and .7_.5_.5_.9). Unequal reliabilities among total tests mean that, technically, equating cannot be done. This study's references to equating method results when the tests have unequal reliabilities are intended as descriptions for the performance of the equating methods under conditions where adequate equating is impossible. Holding the observed standard deviations for the test and anchors constant but varying reliability produced a situation where the Angoff (1953) reliability estimates were not accurate. Their extent of inaccuracy for the reliability conditions and observed score standard deviations in this simulation is shown in Table 4.

Sample Size

Three sample size conditions were considered for *P* and *Q*: $N_P = N_Q$ = 500, 1,000, and 5,000.

Equating Methods

Ten NEAT equating methods were considered for equating X_P to Y_Q through anchors A_P to A_Q. These are the linear methods described in the introduction (chained linear, Tucker, Levine

observed and Levine true) and the equipercentile counterparts of the linear methods (chained equipercentile with raw and smoothed frequencies and frequency estimation equipercentile with raw and smoothed frequencies). Loglinear smoothing (Holland & Thayer, 1987, 2000) was used with the equipercentile methods to preserve four moments on the test and anchor distributions and one cross-product moment between the tests and anchors. The Levine observed and true equating methods were considered using Angoff (1953) reliability estimates for a test and an external anchor and also using the correct reliabilities (i.e., the reliabilities by which the data were actually generated).

Table 3

Reliability Levels in Two Test Scores and Two Anchor Scores

Combination	X_P	A_P	A_Q	Y_Q
1	.9	.8	.8	.9
2	.9	.6	.6	.9
3	.7	.4	.4	.7
4	.9	.5	.5	.7
5	.7	.5	.5	.9

Table 4

Correct and Angoff-Estimated Reliabilities for the Conditions of This Study

Combination	X_P	A_P	A_Q	Y_Q
1	.9 (.93)	.8 (.78)	.8 (.78)	.9 (.93)
2	.9 (.87)	.6 (.62)	.6 (.62)	.9 (.87)
3	.7 (.74)	.4 (.38)	.4 (.38)	.7 (.74)
4	.9 (.83)	.5 (.54)	.5 (.45)	.7 (78)
5	.7 (.78)	.5 (.45)	.5 (.54)	.9 (.83)

Simulation

For the simulation, 200 random datasets of X_P, A_P, A_Q, and Y_Q scores were generated for particular combinations of sample size, reliability, and population standardized ability differences. The 10 equating methods were used to equate X_P to Y_Q in each of these 200 datasets.

For each possible score on X_P (0–100), averages and standard deviations of the 200 equated scores were computed for each equating method. These equating method averages and standard deviations (i.e., empirical standard errors) were then analyzed across equating method, sample size, reliability combination, and population standardized ability difference.

Evaluation of Results

Analysis of variances (ANOVAs) and source mean squares were used to identify the strongest influences on equating method averages and standard errors. Specifically, the equating method averages and standard errors of converted scores at an X_P score of 50 were analyzed in 15-factor ANOVAs composed of the 4 main effects (i.e., the 10 equating methods, 3 sample sizes, 5 reliability combinations, and 3 population standardized ability differences); 6 two-way interactions; 4 three-way interactions; and 1 four-way interaction. The percentages of total variance in these 15 effects gave a general indication of how each manipulated variable contributed to the variation in equating method averages and standard errors. The ANOVA results, like ANOVA results from any controlled study, directly reflect the levels of the variables considered in the study (which were selected because they spanned a range of situations encountered by this study's authors in their equating work). Additional follow-up analyses for the equating method averages and standard errors were also conducted to describe the results not adequately captured by the ANOVA analyses.

Results

Equating Method Averages

Table 5 presents the mean squares from the ANOVA of the 15 effects for the equating method averages for score $X_P = 50$. These mean squares are ranked in terms of their proportion of variance explained on equating method averages. Over 99% of the variation in equating method averages is attributable to the main and interaction effects of equating method and population standardized ability differences and the interaction of these two effects with reliability combinations (equating, equating × ability, equating ×reliability, and equating × ability × reliability). The sample size effect on equating method averages was negligible.

Table 5

Ranked Mean Squares and Their Percentage of Total Variation in Equating Method Averages

Source	DF	Mean square	% of total variance	Cumulative % variance
Equating	9	18.92	72	72
Equating × ability	18	6.36	24	96
Equating × reliability	36	0.62	2	98
Equating × ability × reliability	72	0.21	1	99
Ability	2	0.15	1	100
Reliability	4	0.03	0	100
Ability × reliability	8	0.03	0	100
Ability × sample size	4	0.02	0	100
Ability × reliability × sample size	16	0.01	0	100
Reliability × sample size	8	0.01	0	100
Sample size	2	0.00	0	100
Equating × sample size	18	0.00	0	100
Equating × reliability × sample size	72	0.00	0	100
Equating × ability × reliability × sample size	144	0.00	0	100
Equating × ability × sample size	36	0.00	0	100
Total	449	26.35	100	

Note. $X_P = 50$.

The influences of population standardized ability differences and reliability combination are illustrated in Table 6 (population standardized ability difference = 0), Table 7 (population standardized ability difference = .15), and Table 8 (population standardized ability difference = .30), which give the equating method averages for each equating method across the five reliability combinations averaged across all of the sample sizes. The equating method averages are essentially equal to the criterion equated score of 50 across reliability conditions when

population abilities do not differ (Table 6). When abilities differ (Tables 7 and 8), the equating method averages become more dependent on equating method and also on reliability levels, so that the conditioning methods (Tucker, raw and smoothed frequency estimation equipercentile) give progressively lower equating method averages as reliability declines, the chained methods (chained linear, raw and smoothed chained equipercentile) change only slightly, and the Levine methods give progressively higher equating method averages as reliability declines.

Table 6

Equating Method Averages Across Reliability Combinations When P and Q Abilities Were Equal and Population Standardized Ability Difference = 0

Equating method	Reliability combination				
	.9_.8_.8_.9	.9_.6_.6_.9	.7_.4_.4_.7	.9_.5_.5_.7	.7_.5_.5_.9
Tucker	50.02	50.02	49.99	50.04	50.05
Raw frequency estimation equipercentile	50.02	50.03	49.96	50.06	50.06
Smoothed frequency estimation equipercentile	50.01	50.02	49.99	50.04	50.05
Chained linear	50.01	50.02	50.00	50.05	50.05
Raw chained equipercentile	50.03	50.00	49.94	50.12	50.09
Smoothed chained equipercentile	50.00	50.00	49.99	50.08	50.04
Levine observed-correct reliabilities	50.01	50.02	50.00	50.06	50.05
Levine observed-Angoff reliabilities	50.01	50.02	50.00	50.06	50.05
Levine true-correct reliabilities	50.01	50.01	50.00	50.06	50.04
Levine true-Angoff reliabilities	50.01	50.02	50.00	50.06	50.05

Note. $X_P = 50$.

Table 7

Equating Method Averages Across Reliability Combinations When P and Q Abilities Were Unequal and Population Standardized Ability Difference = .15

Equating method	Reliability combination				
	.9_.8_.8_.9	.9_.6_.6_.9	.7_.4_.4_.7	.9_.5_.5_.7	.7_.5_.5_.9
Tucker	49.60	49.27	48.75	48.98	49.00
Raw frequency estimation equipercentile	49.62	49.29	48.75	48.96	48.99
Smoothed frequency estimation equipercentile	49.63	49.28	48.80	48.99	49.04
Chained linear	50.00	49.99	49.96	49.95	49.99
Raw chained equipercentile	50.01	50.00	49.98	49.97	49.97
Smoothed chained equipercentile	50.02	49.96	49.99	49.93	49.99
Levine observed-correct reliabilities	50.16	50.59	50.80	50.64	50.69
Levine observed-Angoff reliabilities	50.25	50.47	50.99	50.68	50.74
Levine true-correct reliabilities	50.16	50.59	50.78	50.58	50.71
Levine true-Angoff reliabilities	50.25	50.47	50.99	50.70	50.71

Note. $X_P = 50$.

Table 8

Equating Method Averages Across Reliability Combinations When P and Q Abilities Were Unequal and Population Standardized Ability Difference = .30

Equating method	Reliability combination				
	.9_.8_.8_.9	.9_.6_.6_.9	.7_.4_.4_.7	.9_.5_.5_.7	.7_.5_.5_.9
Tucker	49.17	48.58	47.44	48.00	48.02
Raw frequency estimation equipercentile	49.19	48.58	47.49	48.00	48.04
Smoothed frequency estimation equipercentile	49.20	48.61	47.51	48.06	48.08
Chained linear	49.99	50.02	49.96	49.97	50.00
Raw chained equipercentile	50.00	49.97	49.99	49.95	50.00
Smoothed chained equipercentile	49.98	49.99	49.96	49.95	49.99
Levine observed-correct reliabilities	50.32	51.24	51.68	51.37	51.41
Levine observed-Angoff reliabilities	50.49	51.00	52.08	51.47	51.50
Levine true-correct reliabilities	50.30	51.21	51.67	51.25	51.44
Levine true-Angoff reliabilities	50.49	51.00	52.09	51.51	51.45

Note. $X_P = 50$.

Two additionally important findings could not be observed in the variability of the equating method averages at X_P scores of 50. The averages for the Levine true equating method were substantially influenced by an interaction between the reliability estimation method and reliability levels. Figure 1 plots the difference in the Levine true correct equating method averages from the identity function for the population standardized ability difference of 0 and the two reliability combinations that featured different test reliabilities. The slope for the Levine true

with correct reliabilities equating method changed considerably, decreasing when Y_Q's reliability was smaller than X_P*'s* reliability (e.g., reliability combination .9_.5_.5_.7) but increasing when Y_Q's reliability was larger than X_P*'s* reliability (e.g., reliability combination .7_.5_.5_.9). Figure 2 plots the difference in the averages of the Levine true with Angoff reliabilities equating method from the identity function. The slopes of the equating functions shown in Figure 2 were opposite and of relatively smaller magnitude than those of Figure 1.

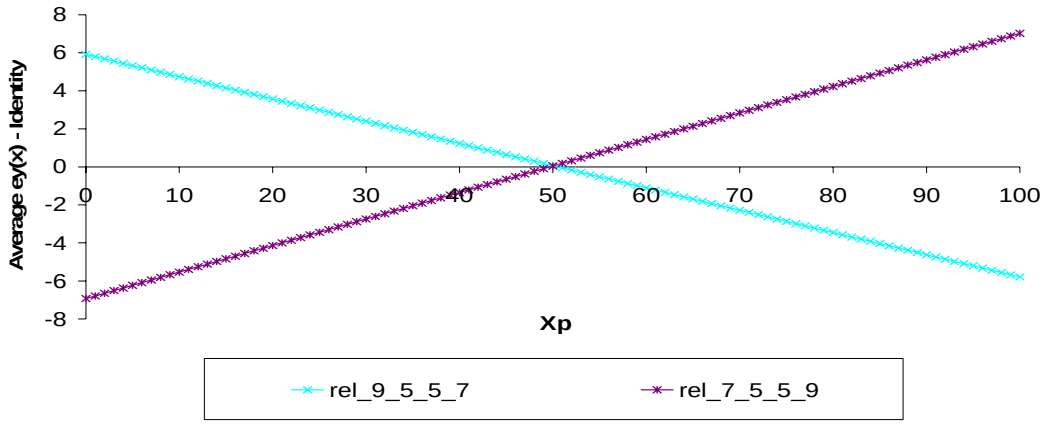

Figure 1. **Levine true with correct reliabilities equating method averages, identity population standardized ability difference = 0.**

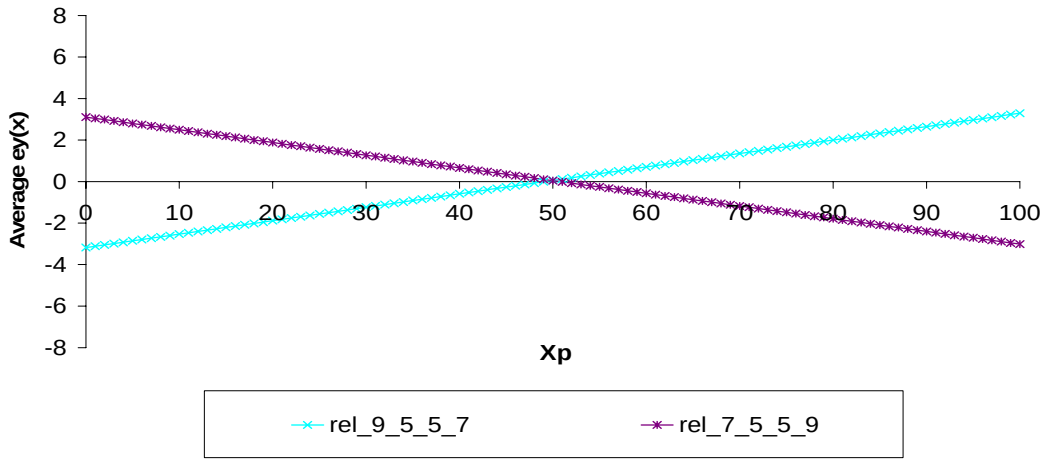

Figure 2. **Levine true with Angoff reliabilities equating method averages, identity population standardized ability difference = 0.**

Equating Method Standard Errors

Table 9 presents the mean squares from the ANOVA of the 15 effects for the equating method standard errors for score $X_P = 50$. These mean squares are ranked in terms of their proportion of variance explained by equating method standard errors. The strongest effects are sample size (90%), followed by reliability (7%), and equating method (2%). These three effects accounted for more than 99% of the total variance in equating method standard errors. Figures 3-32 plot the standard errors conditional on scores across the five reliability combinations for each equating method and sample size condition when population standardized ability differences = 0. From these plots, the increase in equating method standard errors as sample sizes decrease is shown, as is the relatively smaller increase in equating method standard errors as reliabilities decrease. The conditioning methods (Tucker, raw, and smoothed frequency estimation equipercentile) in Figures 3–11 had equating method standard errors that were smaller and less responsive to reliability changes than the chained methods (Figures 12–20). The Levine methods in Figures 21–32 had relatively large equating method standard errors that were strongly influenced by reliability changes. The series of equating method standard errors are U-shaped for the linear methods and dog bone shaped for the equipercentile methods. Finally, the figures also show an interactive effect of reliability and sample size, which accounted for about 1% of the variation in equating method standard errors. The reliability × sample size interaction is that reliability had a much more visible effect on equating method standard errors for smaller sample sizes than for larger sample sizes.

Discussion

The purpose of this study was to evaluate the impact of reliability on test equating methods used in the NEAT design. An essential part of this evaluation was a description of reliability's interaction with the influence of population ability differences on anchor means. Two test score models were summarized and compared in terms of their assumptions about the contribution of reliability and examinee ability on observed scores. The implicit assumptions of different equating methods for addressing reliability and ability differences were related to the assumptions made by different test score models, so any equating method might be inaccurate when test scores are not perfectly reliable, populations differ in ability, and the equating method incorrectly specifies the reliability-ability difference interaction. A simulation was conducted to

illustrate the influence of reliability on several equating methods across levels of population ability difference, anchor and test reliability levels, and sample size.

Table 9

Ranked Mean Squares and Their Percentage of Total Variation in Equating Method Standard Errors

Source	DF	Mean square	% of total variance	Cumulative % variance
Sample size	2	21.17	90	90
Reliability	4	1.62	7	96
Equating	9	0.55	2	99
Reliability × sample size	8	0.17	1	99
Equating × sample size	18	0.06	0	100
Ability × reliability × sample size	16	0.02	0	100
Ability × sample size	4	0.02	0	100
Ability × reliability	8	0.02	0	100
Equating × reliability	36	0.01	0	100
Ability	2	0.01	0	100
Equating × reliability × sample size	72	0.00	0	100
Equating × ability × reliability	72	0.00	0	100
Equating × ability × reliability × sample size	144	0.00	0	100
Equating × ability	18	0.00	0	100
Equating × ability × sample size	36	0.00	0	100
Total	449	23.64	100	

Note. $X_P = 50$.

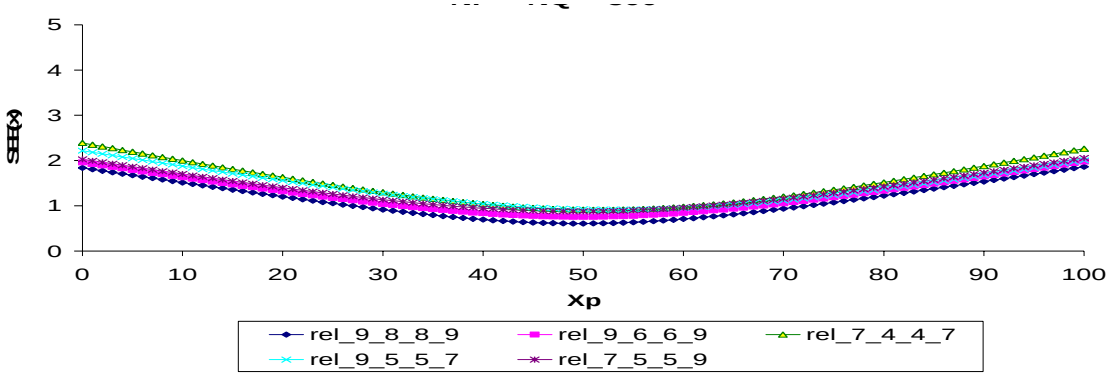

Figure 3. **Tucker equating method standard errors, population standardized ability difference = 0, $N_P = N_Q = 500$.**

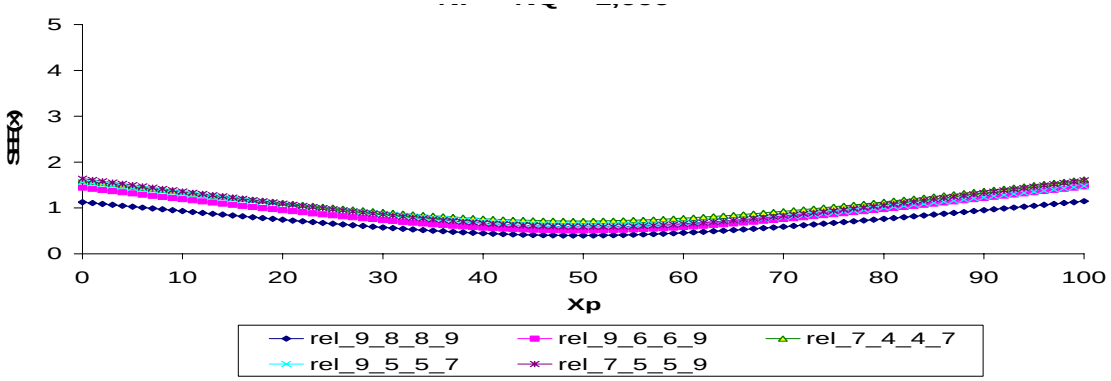

Figure 4. **Tucker equating method standard errors, population standardized ability difference = 0, $N_P = N_Q = 1{,}000$.**

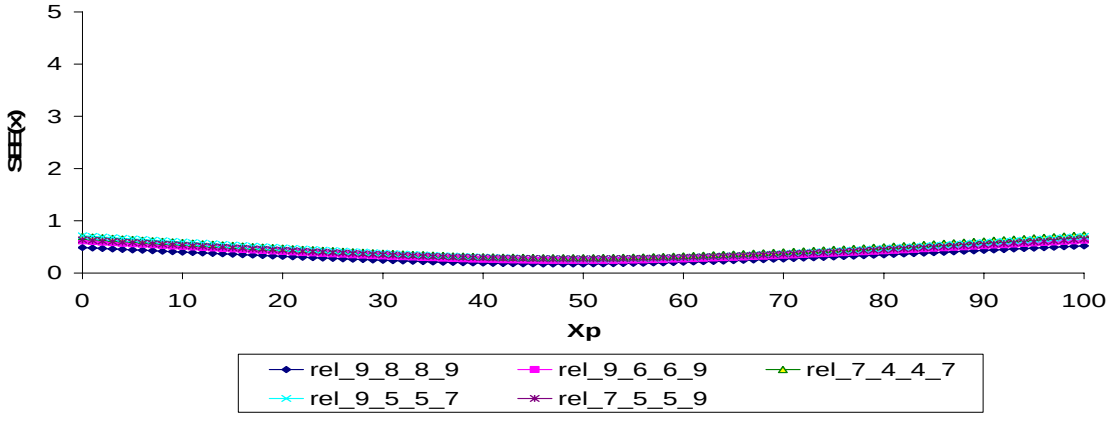

Figure 5. **Tucker equating method standard errors, population standardized ability difference = 0, $N_P = N_Q = 5{,}000$.**

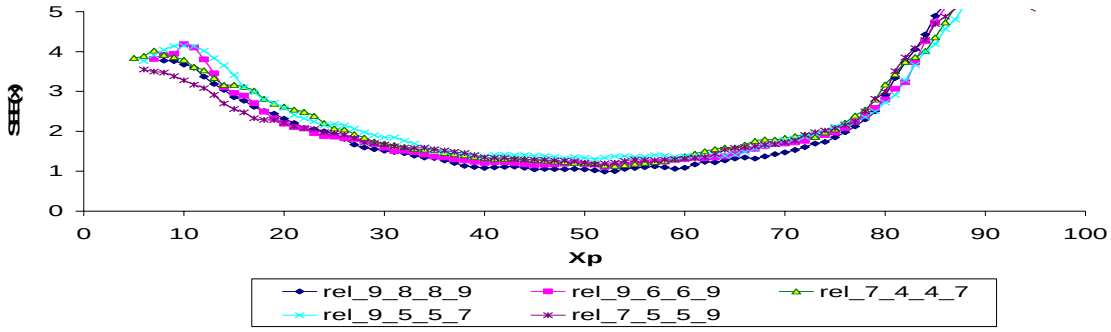

Figure 6. **Raw frequency estimation equipercentile equating method standard errors, population standardized ability difference = 0, $N_P = N_Q = 500$.**

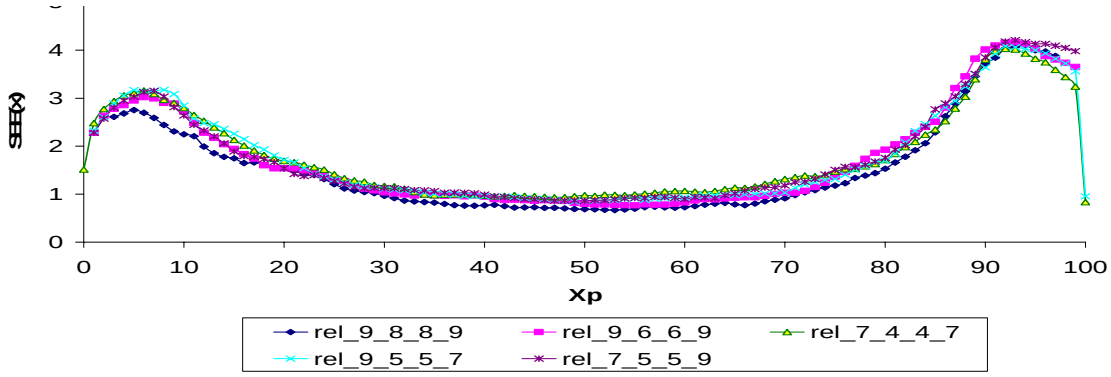

Figure 7. **Raw frequency estimation equipercentile equating method standard errors, population standardized ability difference = 0, $N_P = N_Q = 1{,}000$.**

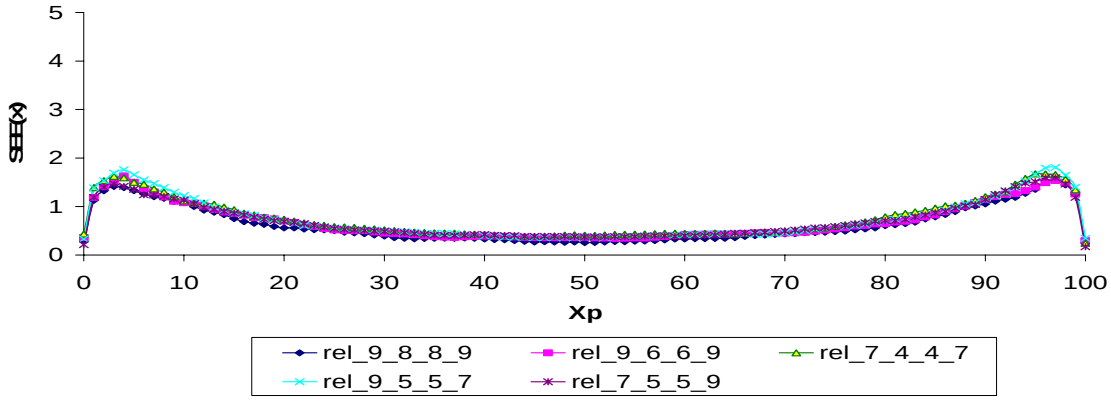

Figure 8. **Raw frequency estimation equipercentile equating method standard errors, population standardized ability difference = 0, $N_P = N_Q = 5{,}000$.**

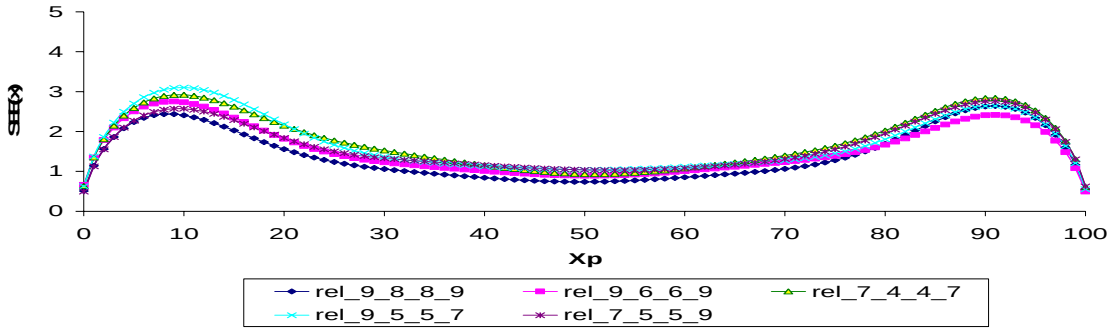

Figure 9. **Smoothed frequency estimation equipercentile equating method standard errors, population standardized ability difference = 0, $N_P = N_Q = 500$.**

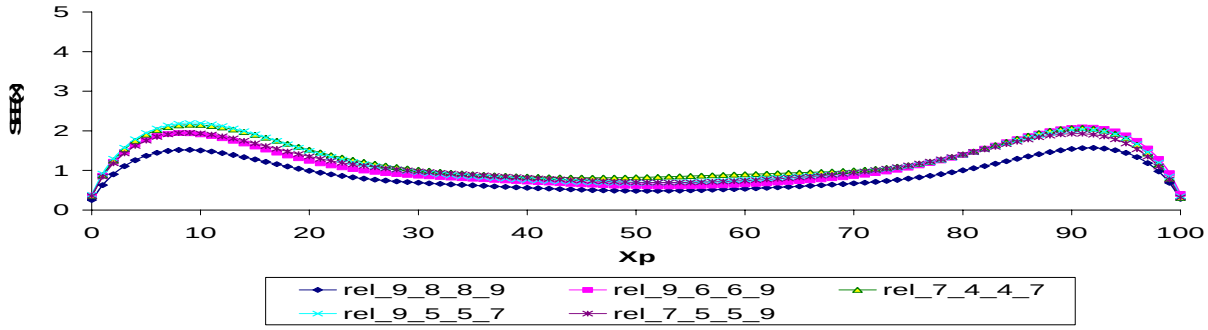

Figure 10. **Smoothed frequency estimation equipercentile equating method standard errors, population standardized ability difference = 0, $N_P = N_Q = 1{,}000$.**

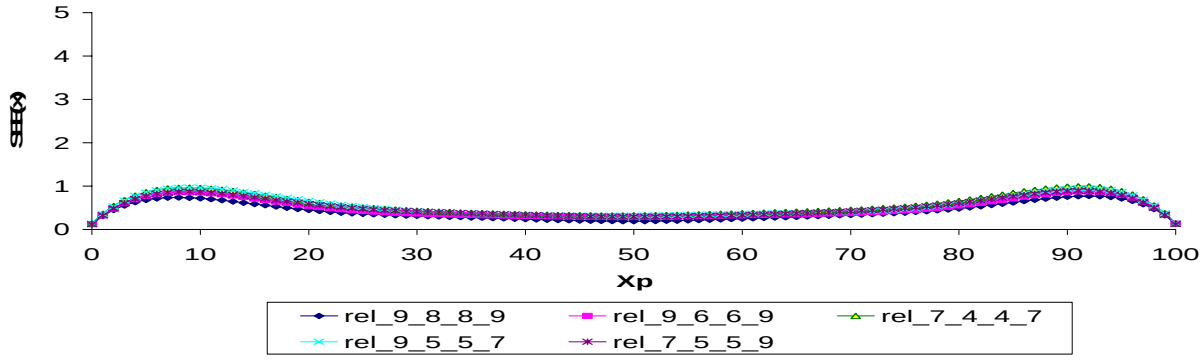

Figure 11. **Smoothed frequency estimation equipercentile equating method standard errors, population standardized ability difference = 0, $N_P = N_Q = 5{,}000$.**

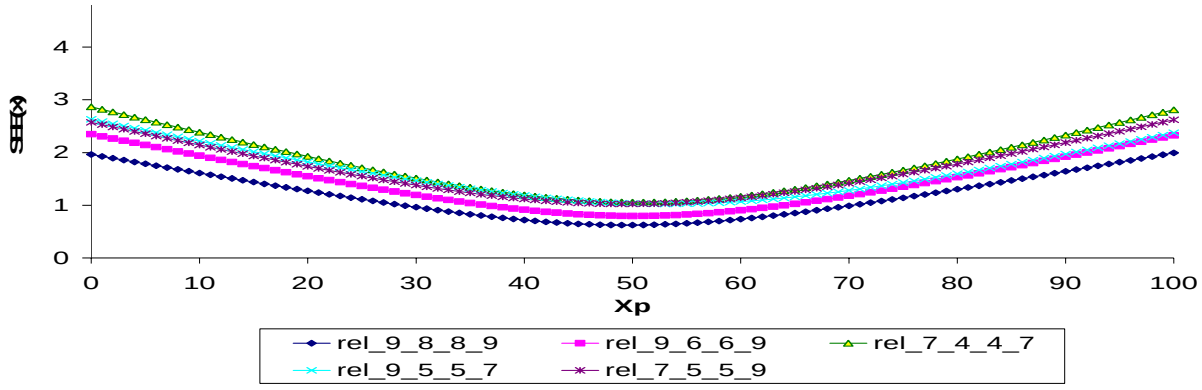

Figure 12. **Chained linear equating method standard errors, population standardized ability difference = 0, $N_P = N_Q = 500$.**

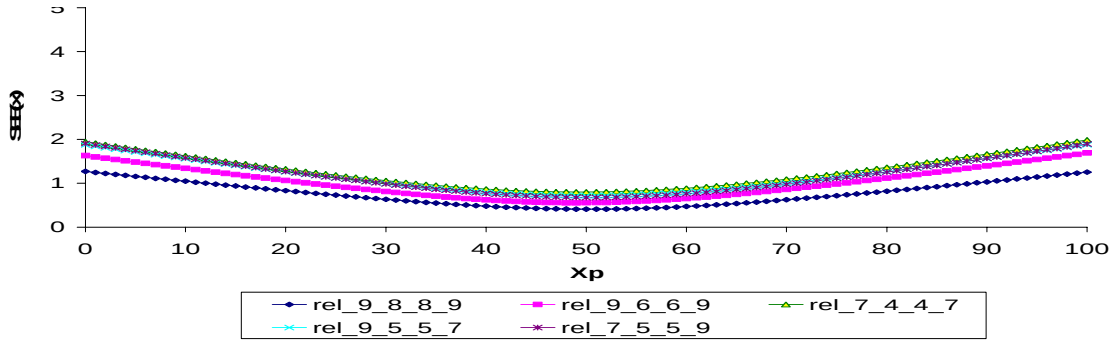

Figure 13. **Chained linear equating method standard errors, population standardized ability difference = 0, $N_P = N_Q = 1{,}000$.**

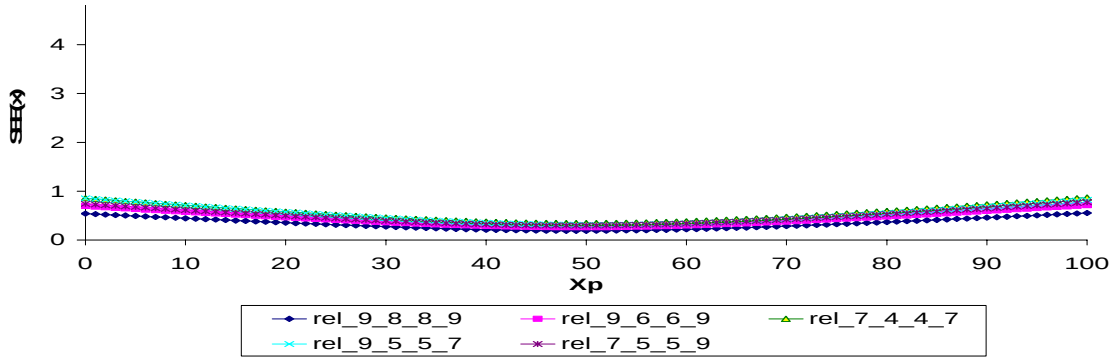

Figure 14. **Chained linear equating method standard errors, population standardized ability difference = 0, $N_P = N_Q = 5{,}000$.**

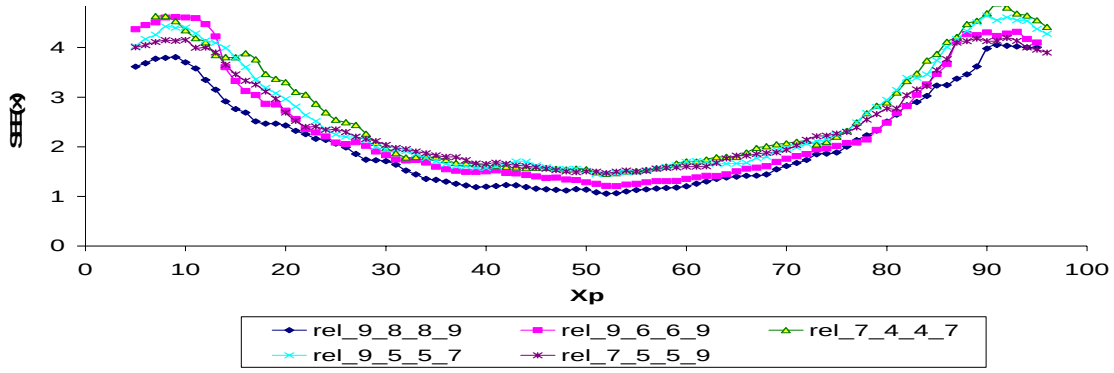

Figure 15. **Raw chained equipercentile equating method standard errors, population standardized ability difference = 0, $N_P = N_Q = 500$.**

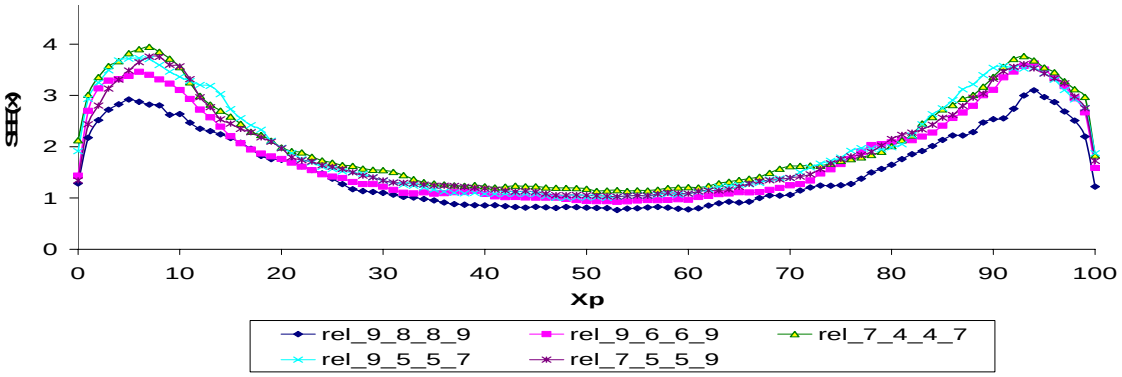

Figure 16. **Raw chained equipercentile equating method standard errors, population standardized ability difference = 0, $N_P = N_Q = 1{,}000$.**

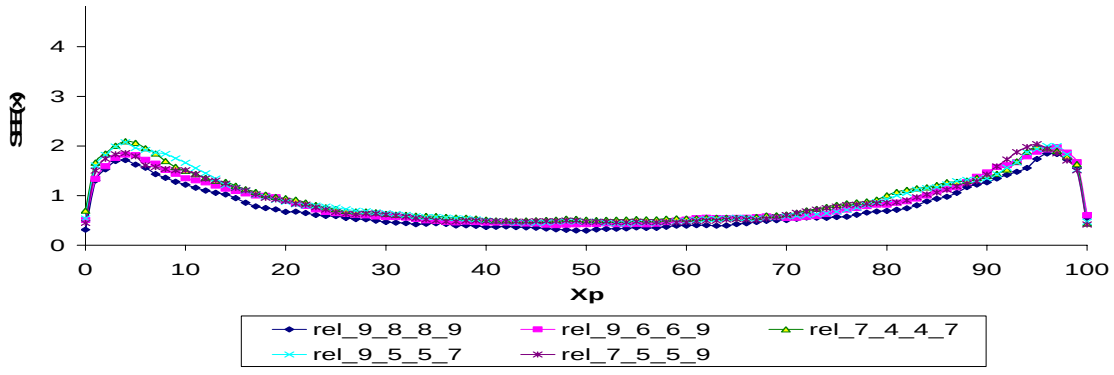

Figure 17. **Raw chained equipercentile equating method standard errors, population standardized ability difference = 0, $N_P = N_Q = 5{,}000$.**

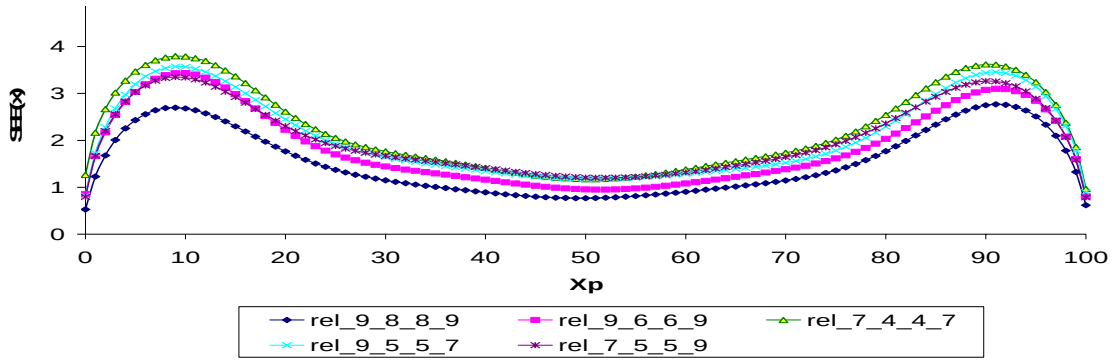

Figure 18. **Smoothed chained equipercentile equating method standard errors, population standardized ability difference = 0, $N_P = N_Q = 500$.**

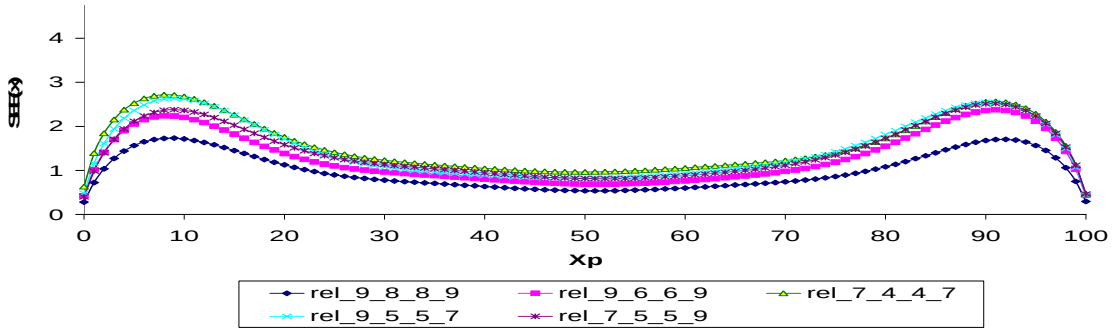

Figure 19. **Smoothed chained equipercentile equating method standard errors, population standardized ability difference = 0, $N_P = N_Q = 1{,}000$.**

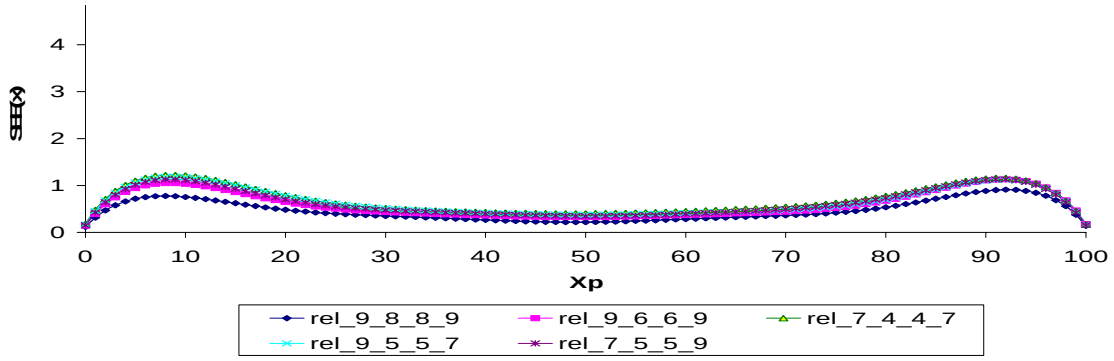

Figure 20. **Smoothed chained equipercentile equating method standard errors, population standardized ability difference = 0, $N_P = N_Q = 5{,}000$.**

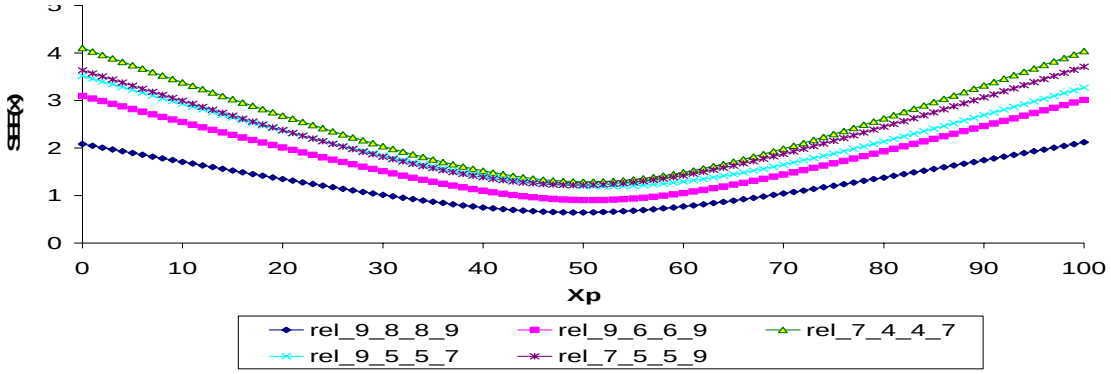

Figure 21. **Levine observed with correct reliabilities equating method standard errors, population standardized ability difference = 0, $N_P = N_Q = 500$.**

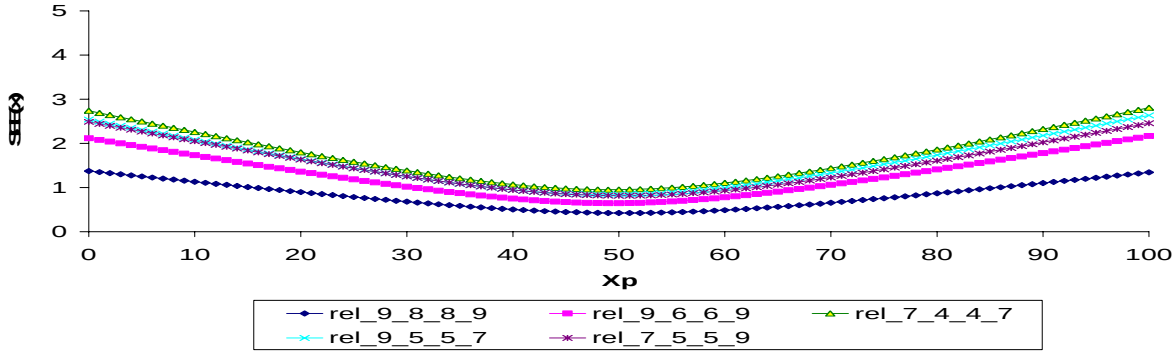

Figure 22. **Levine observed with correct reliabilities equating method standard errors, population standardized ability difference = 0, $N_P = N_Q = 1{,}000$.**

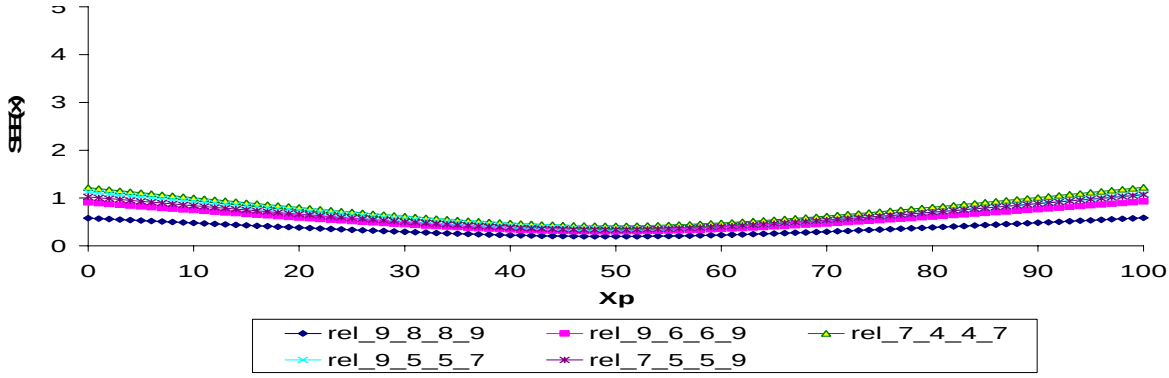

Figure 23. **Levine observed with correct reliabilities equating method standard errors, population standardized ability difference = 0, $N_P = N_Q = 5{,}000$.**

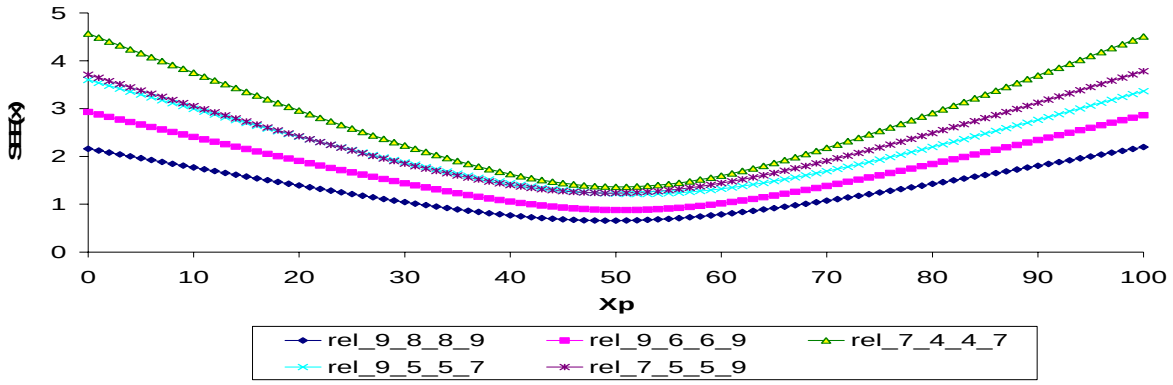

Figure 24. **Levine observed with Angoff reliabilities equating method standard errors, population standardized ability difference = 0, $N_P = N_Q = 500$.**

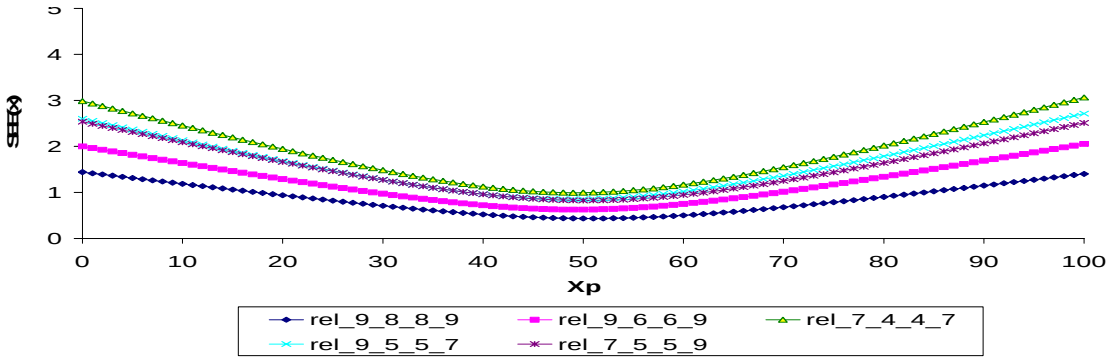

Figure 25. **Levine observed with Angoff reliabilities equating method standard errors, population standardized ability difference = 0, $N_P = N_Q = 1{,}000$.**

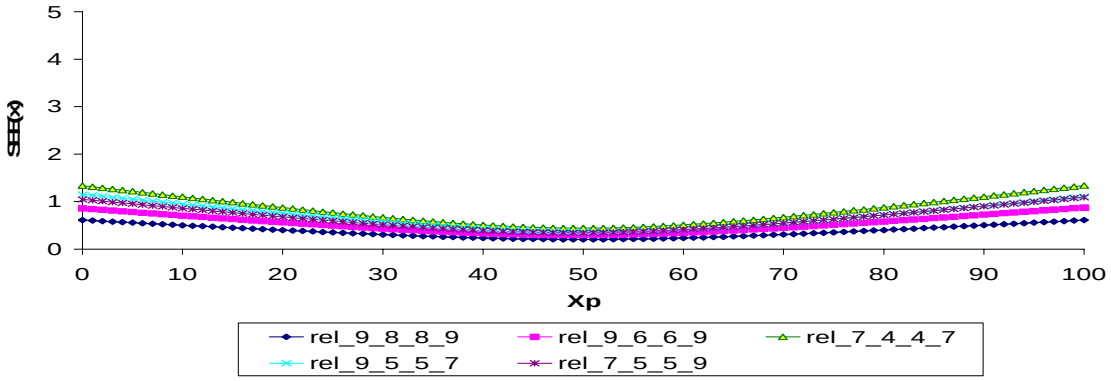

Figure 26. **Levine observed with Angoff reliabilities equating method standard errors, population standardized ability difference = 0, $N_P = N_Q = 5{,}000$.**

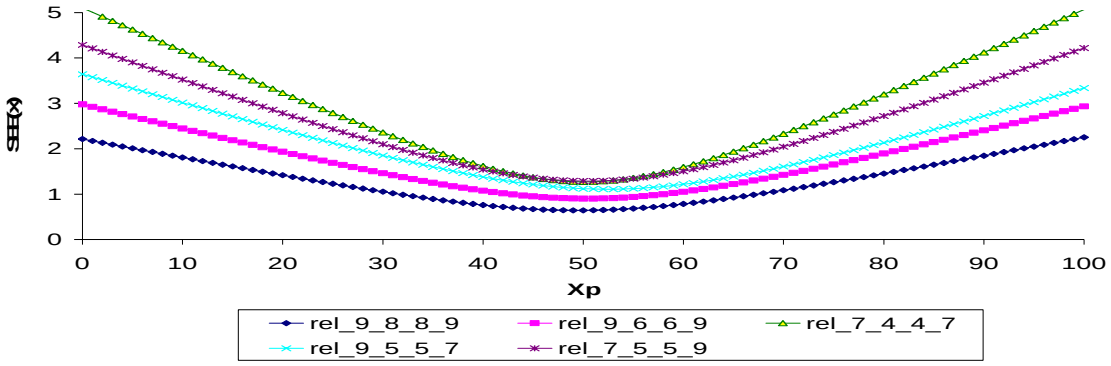

Figure 27. **Levine true with correct reliabilities equating method standard errors, population standardized ability difference = 0, $N_P = N_Q = 500$.**

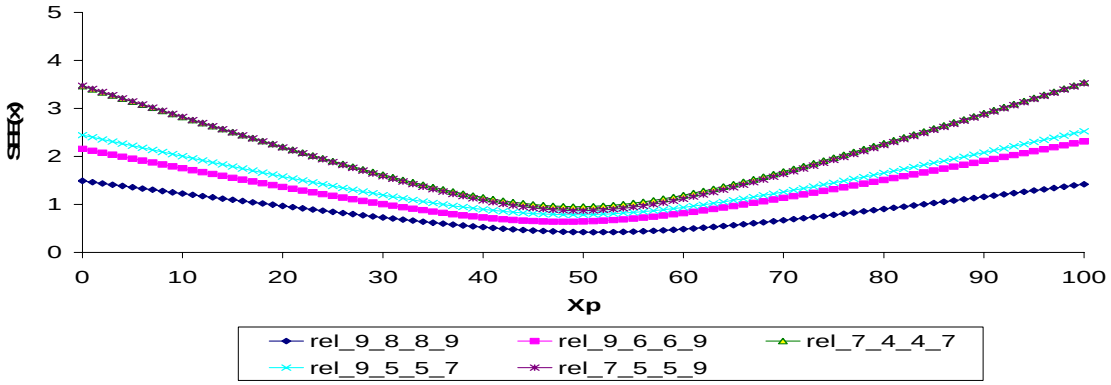

Figure 28. **Levine true with correct reliabilities equating method standard errors, population standardized ability difference = 0, $N_P = N_Q = 1{,}000$.**

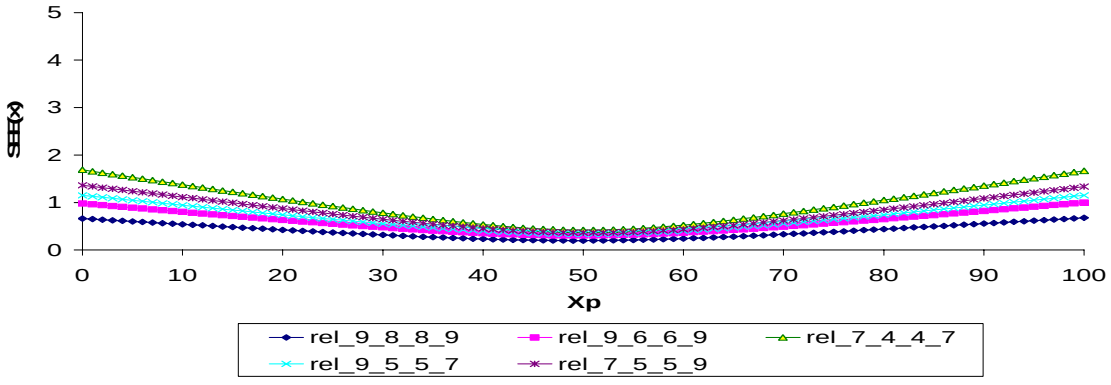

Figure 29. **Levine true with correct reliabilities equating method standard errors, population standardized ability difference = 0, $N_P = N_Q = 5{,}000$.**

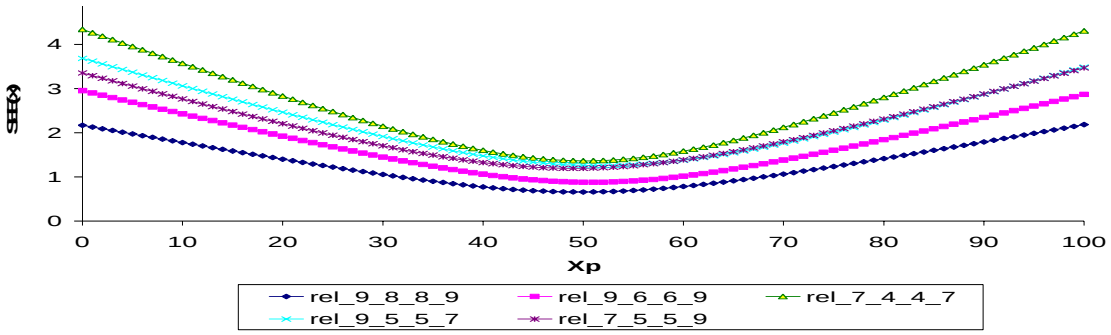

Figure 30. **Levine true with Angoff reliabilities equating method standard errors, population standardized ability difference = 0, $N_P = N_Q = 500$.**

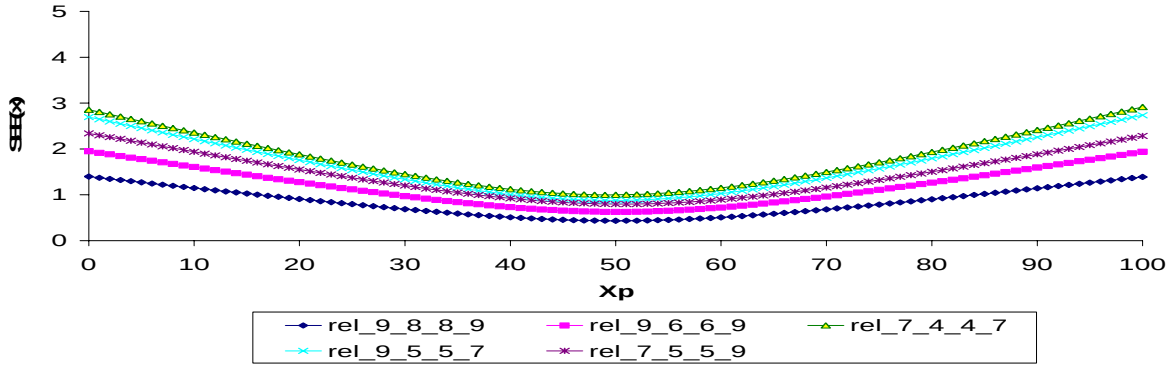

Figure 31. **Levine true with Angoff reliabilities equating method standard errors, population standardized ability difference = 0, $N_P = N_Q = 1{,}000$.**

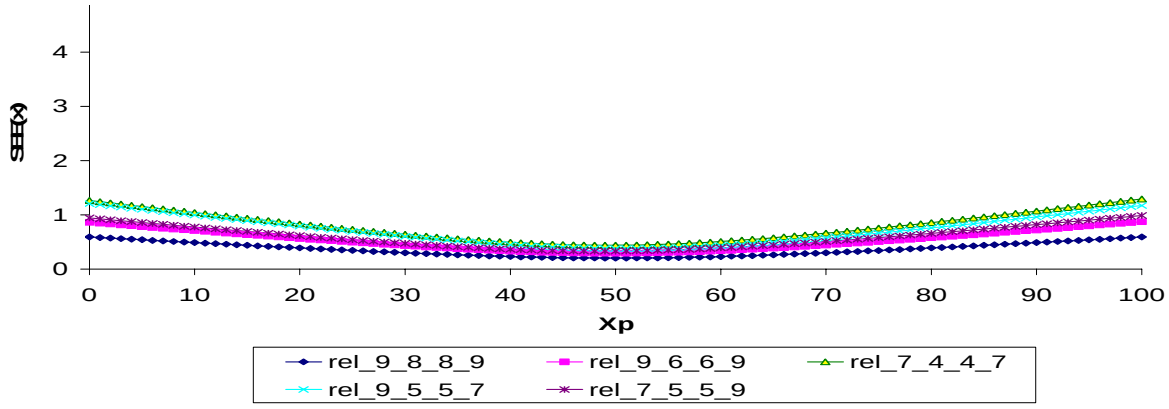

Figure 32. **Levine true with Angoff reliabilities equating method standard errors, population standardized ability difference = 0, $N_P = N_Q = 5{,}000$.**

The results of the simulation are consistent extensions of what is known about the performance of equating methods. When reliabilities become lower and abilities differ, the chained, conditioning, and Levine methods disagree more with each other. In terms of their average performance and the data generation model used in this study, Levine true and observed overmatched on ability differences relative to the chained methods, while the conditioning methods undermatched on ability differences (also described in Holland, 2004; Livingston, 2004; MacCann, 1990). The use of a different data generation model would not be expected to change the relative ordering of equated methods' equated scores, though it would change each method's accuracy (see Wang et al., 2006, for a comparison based on an item response theory (IRT) data generation model where, like this study's results, the conditioning methods were more biased and less variable than the chained methods).

Changes in reliability had a visible effect on equating method standard errors that was relatively small when compared to the effect of changes in sample size. In general, the equating methods that use the anchor as a conditioning variable tend to exhibit smaller standard errors than do chained and Levine methods (von Davier et al., 2004; von Davier & Kong, 2005; Kolen & Brennan, 2004). The results of this study's simulations showed that the standard errors of the conditioning methods are less influenced by levels of reliability compared to the chained and Levine methods. The equipercentile equating functions (e.g., raw and smoothed frequency estimation equipercentile and chained equipercentile) are more variable than their linear counterparts (e.g., Tucker and chained linear), but they exhibit responses to reliability changes that are similar to their linear counterparts.

Levine Results

There were subtle, but understandable, results noted for the Levine methods from the simulation. Levine true's slope varied much more than the slopes of other equating methods when the test reliabilities differed because it was the only considered equating method that built reliability into its slope (16). As test reliabilities differed but observed score variances remained constant, scaling in terms of true score variability was very different from scaling in terms of observed score variability. Varying reliabilities while holding observed score variances constant is an unrealistic feature of this study's generation model, however it is a potential explanation for the large difference in the Levine true method's slope relative to other equating methods' slopes (a phenomenon that is often observed and mulled over in equating practice). The more

sophisticated description for this difference is that true score methods such as the Levine true method are built to satisfy requirements such as second-order equity (i.e., the error variances of $e_y(x)$ and Y are equal at given true scores), and these methods can produce very different results from observed score methods that are built to match observed score variances (Tong & Kolen, 2005).

The incorporation of Angoff (1953) reliability estimates with the Levine methods had important effects on the slope for the Levine true method. Angoff reliability estimates are based on the classical congeneric model's assumptions of perfectly correlated anchor and test true scores and effective test lengths. The assumptions about variances being proportional to reliability were not closely followed in this study, where observed score variances were held constant as reliabilities were altered. Table 4 shows this study's generated (correct) reliabilities and Angoff estimates for the five reliability combinations. When a more reliable (X_P=.9) test is equated to a less reliable (Y_Q=.7) test and the test and anchor observed score standard deviations of this study are used (Table 2), the Levine true function's slope (from 17) is

$$\frac{.\sqrt{.7}(18)}{\sqrt{.5}(5.4)}\frac{\sqrt{.5}(5.4)}{\sqrt{.9}(18)} = \frac{.\sqrt{.7}(18)}{\sqrt{.9}(18)} = \frac{.\sqrt{.7}}{\sqrt{.9}} = .88$$ using correct reliabilities (Figure 1) and

$$\frac{.\sqrt{.78}(18)}{\sqrt{.45}(5.4)}\frac{\sqrt{.54}(5.4)}{\sqrt{.83}(18)} = \frac{\sqrt{.78}\sqrt{.54}}{\sqrt{.45}\sqrt{.83}} = 1.06$$ using Angoff reliabilities (Figure 2). The reversed and smaller magnitude slopes of the Levine true method with Angoff reliabilities rather than correct reliabilities is directly attributable to the extent of inaccuracy in the Angoff reliabilities.

Equating method standard errors were affected by whether correct reliabilities were over or underestimated by the Angoff (1953) reliabilities (Table 4). Levine observed's equating functions were generally more variable with Angoff reliabilities than with correct reliabilities, except when both total tests' reliabilities were underestimated (the reliability combination of .9_.6_.6_.9). the Levine true method's equating functions became more variable when the Angoff reliability estimates were used and they overestimated Y_Q*'s* reliability (reliability combinations of .9_.8_.8_.9, .7_.4_.4_.7, and .9_.5_.5_.7) and became less variable when Angoff estimates underestimated Y_Q*'s* reliability (reliability combinations of .9_.6_.6_.9, and .7_.5_.5_.9). The reliability estimation method influences the variability of Levine equating functions through the extent of magnification in μ_{AP}-μ_{AQ} , setting $\gamma_{ey} = \frac{\sqrt{rel_{YQ}}}{\sqrt{rel_{AQ}}}$ in (14) for (16) and

(17). An underestimated Y_Q reliability and/or an overestimated A_Q reliability resulted in less magnification of μ_{AP}-μ_{AQ} and its sampling variability on the final Levine functions, whereas an overestimated Y_Q reliability and/or an underestimated A_Q reliability resulted in more magnification of μ_{AP}-μ_{AQ} and its sampling variability on the final Levine functions.

Implications

There are important implications for studying reliability as a relationship between score models and equating methods. When data are unreliable, the effect of population ability differences on test scores depends on the assumed score model, and different score models are compatible with some equating methods but not others. In unreliable data, an equating practitioner may have to make a nonempirical choice among models based on how reliability impacts the test scores, whether unreliability reduces (the Tucker and frequency estimation methods), magnifies (the Levine method) or does not affect (the chained linear method) the extent to which examinee ability influences test scores. The major basis for this choice may be some interpretative evaluation of the quality of the anchor scores for estimating ability effects on test scores.

Relationships between equating methods and score models not considered in this paper can potentially be understood in terms of this paper's discussion. Manipulating reliability in the classical congeneric model has a somewhat analogous effect in terms of manipulating reliability in a two-parameter IRT model, essentially that reliability reductions reduce the extent to which examinee ability is visible on observed scores. For example, if reliability were reduced in a two-parameter logistic model through reducing the α_i parameter in

$$P(X_{it}=1\,|\,\theta_t,\beta_i,\alpha_i)=\frac{\exp[\alpha_i(\theta_t-\beta_i)]}{1+\exp[\alpha_i(\theta_t-\beta_i)]}, \qquad (25)$$

where θ_t, β_i, and α_i have their usual meanings as test-taker trait level and item difficulty and discrimination parameters, the result would be that the difference between test taker ability and item difficulty would be less visible in the IRT-based observed item and test characteristic curves. Levine's magnification of anchor score mean differences may therefore be somewhat appropriate for IRT-generated data in the same way that Levine is appropriate for classical congeneric models. This suggestion has some support from results showing that IRT and Levine

equating methods cluster together when there are population ability differences (Livingston, Dorans, & Wright, 1990), and other IRT-simulated results (Wang et al., 2006) show the same bias orderings between chained linear and conditioning equating methods described in the introduction.

Another implication of this study is that more complicated interactions of reliability with test score characteristics can potentially be studied with respect to test equating through the use of more complex versions of the score models considered in this paper. This paper was concerned with the very simple case of tests and anchors with perfectly correlated true scores and examinee populations with no systematic subpopulations. In actual data, low and/or unequal reliability coincides with lack of population invariance (Dorans & Holland, 2000; Flanagan, 1951; Holland, Liu, & Thayer, 2005; Kolen, 2004) and imperfectly correlated true scores. Group effects and construct differences could be built into many different score models, and then these effects could be studied in terms of their implications for equating. Such effects violate equating requirements other than the requirement of equal test reliabilities. The study of equating methods' behavior with respect to combinations of equating requirement violations is an important way of relating degrees of equating violations to degrees of equating inadequacy.

References

Allen, M. J., & Yen, W. M. (1979). *Introduction to measurement theory.* City, CA: Wadsworth, Inc.

Angoff, W. H. (1953). Test reliability and effective test length. *Psychometrika, 18,* 1–14.

Angoff, W. H. (1971). Scales, norms, and equivalent scores. In R. L. Thorndike (Ed.), *Educational measurement* (2nd ed., pp. 508–600). Washington, DC: American Council on Education.

Brennan, R. L. (1990). *Congeneric models and Levine's linear equating procedures* (ACT Research Rep. No. 90-12). Iowa City, IA: American College Testing.

Brennan, R. L. (2001). *Generalizability theory.* New York: Springer-Verlag.

Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. (1972). *The dependability of behavioral measurements: Theory of generalizability for scores and profiles.* New York: Wiley.

von Davier, A. A., Holland, P. W., & Thayer, D. T. (2004). *The kernel method of test equating.* New York: Springer-Verlag.

von Davier, A. A., & Kong, N. (2005). A unified approach to linear equating for the non-equivalent group design. *Journal of Educational and Behavioral Statistics, 30,* 313–342.

Dorans, N. J., & Holland, P. W. (2000). Population invariance and the equitability of tests: Basic theory and the linear case. *Journal of Educational Measurement, 37*(4), 281–306.

Feldt, L. S., & Brennan, R. L. (1989). Reliability. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 105–146). New York: Macmillan.

Flanagan, J. C. (1951). Units, scores and norms. In E. F. Lindquist (Ed.), *Educational measurement* (pp. 695–763). Washington DC: American Council on Education.

Hanson, B. A. (1991). A note on Levine's formula for equating unequally reliable tests using data from the common item nonequivalent groups design. *Journal of Educational Statistics, 16,* 93–100.

Holland, P. W. (2004). *Three methods of linear equating for the NEAT design.* Unpublished manuscript.

Holland, P. W., Liu, M., & Thayer, D. (2005). *Exploring the population sensitivity of linking functions to differences in test constructs and reliability using the Dorans-Holland*

measures, kernel equating, and data from the LSAT. Paper presented at the National Council for Measurement and Education, Montreal, Canada.

Holland, P. W., & Thayer, D. T. (1987). *Notes on the use of log-linear models for fitting discrete probability distributions* (ETS Research Rep. No. RR-87-31). Princeton, NJ: ETS.

Holland, P. W. & Thayer, D. T. (2000). Univariate and bivariate loglinear models for discrete test score distributions. *Journal of Educational and Behavioral Statistics, 25,* 133–183.

Kolen, M. J. (2004). Population invariance in equating and linking: Concept and history. *Journal of Educational Measurement, 41*(1), 3–14.

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

Livingston, S. A. (2004). *Equating test scores (without IRT).* Princeton, NJ: ETS.

Livingston, S. A., Dorans, N. J., & Wright, N. K. (1990). What combination of sampling and equating methods works best? *Applied Measurement in Education, 3*(1), 73–95.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems.* Hillsdale, NJ: Lawrence Erlbaum Associates.

MacCann, R. G. (1990). Derivations of observed score equating methods that cater to populations differing in difficulty. *Journal of Educational Statistics, 15,* 146–170.

Petersen, N. S., Kolen, M. J., & Hoover, H. D. (1989). Scaling, norming, and equating. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 221–262). New York: Macmillan.

Tong, Y., & Kolen, M. J. (2005). Assessing equating results on different equating criteria. *Applied Psychological Measurement, 29*(6), 418–432.

Wang, T., Lee, W., Brennan, R. L., & Kolen, M. J. (2006). *A comparison of the frequency estimation and chained equipercentile methods under the common-item non-equivalent groups design.* Paper presented at the National Council on Measurement in Education, San Francisco, CA.