*Center for Advanced Studies in Measurement and Assessment*

*CASMA Research Report*

*Number 37*

# Effects of the Number of Common Items on Equating Precision and Estimates of the Lower Bound to the Number of Common Items Needed*

*Mengyao Zhang and Michael J. Kolen*†

August 2013

---

*The authors thank Robert L. Brennan and Won-Chan Lee for helpful comments on a previous draft.

†Mengyao Zhang is a research assistant in the Center for Advanced Studies in Measurement and Assessment (CASMA), College of Education, University of Iowa (email: mengyao-zhang@uiowa.edu). Michael J. Kolen is Professor, College of Education, University of Iowa (email: michael-kolen@uiowa.edu).

Center for Advanced Studies in
Measurement and Assessment (CASMA)
College of Education
University of Iowa
Iowa City, IA 52242
Tel: 319-335-5439
Web: www.education.uiowa.edu/casma

# Contents

# List of Tables



# List of Figures

# Abstract

The construction of test forms including common items can be challenging. By combining the classical congeneric model with analytic standard errors derived by the delta method, this study develops a process for estimating the numbers of common items that are necessary to provide the desired equating precision indexed by the standard error of equating. The chained linear equating method is studied. Both external and internal sets of common items are considered, along with a variety of real test situations represented by test reliability, sample size available, and score range of interest.

# 1 Introduction

In common item equating, scores on test Form X are equated to scores on test Form Y using scores on a set of items, V, that are in common to the two forms. When scores on the common items contribute to the total score on test forms X and Y, the common items are referred to as being internal. When scores on the common items do not contribute to the total score on test forms X and Y, the common items are referred to as being external.

In common item equating, the groups of examinees taking test forms X and Y can be considered to be equivalent in ability, such as when forms X and Y are randomly assigned to examinees for the purposes of equating using the random groups design (Kolen & Brennan, 2004). Alternatively, the groups can be considered not equivalent in ability using what is sometimes referred to as the common item nonequivalent groups design (CINEG, Kolen & Brennan, 2004) or as the nonequivalent groups anchor test design (NEAT, Holland & Dorans, 2006).

The construction of test forms including common items is one of the most challenging parts of common item equating (Kolen & Brennan, 2004). Some previous studies have empirically shown that larger numbers of common items generally produced greater equating precision (Puhan, 2010; Ricker & von Davier, 2007; Wang, Lee, Brennan, & Kolen, 2006; Yang & Houang, 1996). However, since the findings were based on specific test data and situations by manipulating the numbers of common items in a limited manner, the generalizability of these results is uncertain. Also, no general analytic process exists in the literature for estimating the numbers of common items leading to desired equating precision.

In this study it is also shown that the number of common items included in equating has a direct effect on the precision of the estimates of the equating relationship, with larger numbers of common items leading to greater precision. Furthermore, when designing equating studies, the test developer can base the choice of the number of common items on the degree of equating precision desired. The purpose of this study is to detail a process that can be used to choose the numbers of common items that are necessary to provide the desired degree of equating precision when using chained linear equating procedures. Both external and internal sets of common items are considered.

In this study, under the classical congeneric model, the precision of equating is shown to be related directly to the correlation between the scores on the total test and scores on the common items. The development of the approach begins by showing how this correlation relates to reliability for the total test and the ratio of test lengths for the common items and total numbers of items on the test. The development of the approach then relates this correlation to equating precision as indexed by the standard error of equating (SEE). After specifying reliability of the total test, the sample size available, the score range of interest, and the degree of precision desired, the procedures described in this study allow the test developer to choose the length of the set of common items that will lead to the desired equating precision when the chained linear equating method

is used. In the present study, two major assumptions are made that the groups are equivalent in ability and the score distributions are normal. The simulation provides an empirical check on the extent to which the results hold when these assumptions are violated.

## 2 Classical Congeneric Model Results

Let $X$, $Y$, and $V$ represent the random variable observed scores on test forms X, Y, and the set of common items V, respectively. As assumed in the classical test theory, every observed score on a test form or, more generally, on a set of items is the sum of two exclusive components, true score $T$ and error of measurement $E$ (Feldt & Brennan, 1989; Kolen & Brennan, 2004). Subscripts are needed to specify which test form or item set is considered. For example, $T_X$ and $E_X$ denote true score and error of measurement related to test form X.

In classical test theory, varying degrees of heterogeneity between test forms are studied by using different conceptions of parallel measurements (Feldt & Brennan, 1989). In this study, the classical congeneric model is chosen for test form X and the set of common items V because of its flexibility in reflecting similarity and dissimilarity between a test form and a subset of items on the test. Similar results can be extended to test form Y and the set of common items V.

According to Kolen and Brennan (2004), the following properties hold if the classical congeneric model is assumed.

1. True scores $T_X$ and $T_V$ are linearly related, where $\lambda$'s and $\delta$'s are slopes and intercepts respectively, as

$$X = T_X + E_X = (\lambda_X T + \delta_X) + E_X, \tag{1}$$

and

$$V = T_V + E_V = (\lambda_V T + \delta_V) + E_V. \tag{2}$$

2. Error variances are proportional to effective test lengths $\lambda_X$ and $\lambda_V$ as

$$\sigma^2(E_X) = \lambda_X \sigma^2(E), \tag{3}$$

and

$$\sigma^2(E_V) = \lambda_V \sigma^2(E). \tag{4}$$

3. Score variances and covariances are derived from Equations 1 through 4 as

$$\sigma^2(X) = \lambda_X^2 \sigma^2(T) + \lambda_X \sigma^2(E), \tag{5}$$

$$\sigma^2(V) = \lambda_V^2 \sigma^2(T) + \lambda_V \sigma^2(E), \tag{6}$$

and

$$\sigma(X, V) = \lambda_X \lambda_V \sigma^2(T) + \sigma(E_X, E_V). \tag{7}$$

For convenience, Equations 6 and 7 can be rewritten as equations involving only the observed score variance $\sigma^2(X)$, reliability for the total test $\rho(X, X')$, and effective test lengths $\lambda_X$ and $\lambda_V$, where reliability $\rho(X, X')$ is defined as the ratio $\frac{\sigma^2(T_X)}{\sigma^2(X)} = 1 - \frac{\sigma^2(E_X)}{\sigma^2(X)}$ (Feldt & Brennan, 1989). Specifically, for the classical congeneric model, Equation 6 is rewritten as

$$\begin{aligned}
\sigma^2(V) &= \lambda_V^2\sigma^2(T) + \lambda_V\sigma^2(E) \\
&= \frac{\lambda_V^2}{\lambda_X^2} \cdot \lambda_X^2\sigma^2(T) + \frac{\lambda_V}{\lambda_X} \cdot \lambda_X\sigma^2(E) \\
&= \frac{\lambda_V^2}{\lambda_X^2}\sigma^2(T_X) + \frac{\lambda_V}{\lambda_X}\sigma^2(E_X) \\
&= \frac{\lambda_V^2}{\lambda_X^2}\sigma^2(X)\rho(X, X') + \frac{\lambda_V}{\lambda_X}\sigma^2(X)[1 - \rho(X, X')] \\
&= \frac{\lambda_V}{\lambda_X}\sigma^2(X)\left[1 + \left(\frac{\lambda_V}{\lambda_X} - 1\right)\rho(X, X')\right]. \qquad (8)
\end{aligned}$$

When V is an external set of common items, $\sigma(E_X, E_V) = 0$, and $\sigma(X, V) = \lambda_X\lambda_V\sigma^2(T)$ (Kolen & Brennan, 2004). Thus, Equation 7 can be rewritten as

$$\begin{aligned}
\sigma(X, V) &= \lambda_X\lambda_V\sigma^2(T) \\
&= \frac{\lambda_V}{\lambda_X} \cdot \lambda_X^2\sigma^2(T) \\
&= \frac{\lambda_V}{\lambda_X}\sigma^2(T_X) \\
&= \frac{\lambda_V}{\lambda_X}\sigma^2(X)\rho(X, X'). \qquad (9)
\end{aligned}$$

When V is an internal set of common items, $\sigma(E_X, E_V) = \lambda_V\sigma^2(E)$, and $\sigma(X, V) = \lambda_X\lambda_V\sigma^2(T) + \lambda_V\sigma^2(E)$ (Kolen & Brennan, 2004). Similarly, Equation 7 can be rewritten as

$$\begin{aligned}
\sigma(X, V) &= \lambda_X\lambda_V\sigma^2(T) + \lambda_V\sigma^2(E) \\
&= \frac{\lambda_V}{\lambda_X}[\lambda_X^2\sigma^2(T) + \lambda_X\sigma^2(E)] \\
&= \frac{\lambda_V}{\lambda_X}\sigma^2(X). \qquad (10)
\end{aligned}$$

For an external set of common items, by substituting Equations 8 and 9 in the equation for the Pearson product-moment correlation coefficient, an expression for the squared correlation between the scores on the total test and scores on the common items, $\rho(X, X')$, is developed, where $k$ is defined as the ratio $\frac{\lambda_V}{\lambda_X}$ representing the relative length of the set of common items ($k > 0$), as

$$\rho^2(X,V) = \frac{\sigma^2(X,V)}{\sigma^2(X)\sigma^2(V)} = \frac{[\frac{\lambda_V}{\lambda_X}\sigma^2(X)\rho(X,X')]^2}{\sigma^2(X)\cdot\frac{\lambda_V}{\lambda_X}\sigma^2(X)[1+(\frac{\lambda_V}{\lambda_X}-1)\rho(X,X')]} = \frac{k\rho^2(X,X')}{1+(k-1)\rho(X,X')}. \quad (11)$$

For an internal set of common items, the expression of $\rho^2(X,V)$ is developed by using Equations 8 and 10 in a similar manner ($0 < k \leq 1$) as

$$\rho^2(X,V) = \frac{\sigma^2(X,V)}{\sigma^2(X)\sigma^2(V)} = \frac{[\frac{\lambda_V}{\lambda_X}\sigma^2(X)]^2}{\sigma^2(X)\cdot\frac{\lambda_V}{\lambda_X}\sigma^2(X)[1+(\frac{\lambda_V}{\lambda_X}-1)\rho(X,X')]} = \frac{k}{1+(k-1)\rho(X,X')}. \quad (12)$$

Table 1 shows that when V is an external set of common items, the squared correlation $\rho^2(X,V)$ changes as $\rho(X,X')$ and $k$ change. Figure 1 (upper part) graphically illustrates this relationship. For a fixed $k$, higher test reliability lends to higher $\rho^2(X,V)$. For fixed test reliability, the longer $k$ is, the higher $\rho^2(X,V)$. $\rho^2(X,V)$ would reach 1 only when $\rho(X,X') = 1$.

The internal common items case is shown in Table 2 and Figure 1 (lower part). Note that V and X are actually the same when $k = 1$. The relationship of $\rho^2(X,V)$, $\rho(X,X')$ and $k$ is similar to the external common items case, except that $\rho^2(X,V)$ eventually reaches 1 when $k = 1$.

# 3 Chained Linear Equating Method

In the CINEG design, the chained equating method generally involves two steps (Kolen & Brennan, 2004). First, scores on test form X are converted to scores on the common items V based on the group of examinees taking test form X (Group 1), denoted as $X \rightarrow V$. Next, scores on the common items V are converted to scores on test form Y based on the group of examinees taking test form Y (Group 2), denoted as $V \rightarrow Y$. In this study, the chained linear equating method is considered. As its name suggests, this equating method contains two linear conversions, $X \rightarrow V$ and $V \rightarrow Y$. The chained linear equating method is relatively simple and straightforward compared with other equating methods, and it still can be formulated within the general framework for observed-score equating relationships (Brennan, 2006). In addition, the chained linear equating method often leads to greater random error of equating compared to other linear methods used for common-item equating (Kolen & Brennan, 2004). Thus, estimation of the number of common items based on this equating method might be fairly conservative.

Suppose scores on test form X and the common items V in Group 1 satisfy a bivariate normal distribution. Let $\mu(X)$ and $\sigma(X)$ denote the mean and

standard deviation of scores on form X, and let $\mu(V)$ and $\sigma(V)$ denote mean and standard deviation of scores on the set of common items V. Use $N$ to represent the sample size. Subscripts are used to differentiate group membership only when confusion may otherwise occur. For every possible score $x_i$ on test form X, an approximation of random error variance for the single group linear equating $X \rightarrow V$ was originally proposed by Lord (1950) (also see Angoff, 1971; Kolen & Brennan, 2004) as

$$var[\hat{l}_V(x_i)] \cong \frac{\sigma_1^2(V)[1-\rho(X,V)]}{N_1}\left\{2+[1+\rho(X,V)]\left[\frac{x_i-\mu(X)}{\sigma(X)}\right]^2\right\}. \quad (13)$$

A similar equation also holds for the single group linear equating $V \rightarrow Y$ in Group 2 if scores on test form Y and the common items V are assumed to have a bivariate normal distribution as

$$var[\hat{l}_Y(v_i)] \cong \frac{\sigma^2(Y)[1-\rho(Y,V)]}{N_2}\left\{2+[1+\rho(Y,V)]\left[\frac{v_i-\mu_2(V)}{\sigma_2(V)}\right]^2\right\}. \quad (14)$$

According to Braun and Holland (1982), if two equating chains, $X \rightarrow V$ and $V \rightarrow Y$, are statistically independent, the error variance of the entire chained equating, $var[\hat{e}_Y(x_i)]$, could be estimated based on $var[\hat{l}_V(x_i)]$ and $var[\hat{l}_Y(v_i)]$,

$$var[\hat{e}_Y(x_i)] \cong var[\hat{l}_Y(v_i)] + [\hat{l}'_Y(v_i)]^2 var[\hat{l}_V(x_i)], \quad (15)$$

where $\hat{l}'_Y(v_i)$ indicates the slope of the linear conversion from V to Y that is $\frac{\sigma(Y)}{\sigma_2(V)}$ by the definition of linear equating (Kolen & Brennan, 2004). As a result of a linear conversion, two $z$-scores $\frac{v_i-\mu_1(V)}{\sigma_1(V)}$ and $\frac{x_i-\mu(X)}{\sigma(X)}$ should be equal, denoted by $z_i = \frac{v_i-\mu_1(V)}{\sigma_1(V)} = \frac{x_i-\mu(X)}{\sigma(X)}$. By substituting Equations 13 and 14 in Equation 15 and assuming that,

1. groups are equivalent in ability, so that $\mu_1(V) = \mu_2(V)$ and $\sigma_1(V) = \sigma_2(V)$,
2. the correlation between X and V in Group 1 equals the correlation between Y and V in Group 2, $\rho(X,V) = \rho(Y,V)$, and
3. numbers of examinees taking test forms X and Y are equal, namely $N_1 = N_2 = \frac{N_{tot}}{2}$, where $N_{tot}$ represents the total sample size,

an approximation of random error variance for the chained linear equating method is as follows

$$
\begin{aligned}
var[\hat{e}_Y(x_i)] &\cong var[\hat{l}_Y(v_i)] + [\hat{l}'_Y(v_i)]^2 var[\hat{l}_V(x_i)] \\
&= \frac{\sigma^2(Y)[1-\rho(Y,V)]}{N_2}\left\{2 + [1+\rho(Y,V)]\left[\frac{v_i - \mu_2(V)}{\sigma_2(V)}\right]^2\right\} \\
&\quad + \left[\frac{\sigma(Y)}{\sigma_2(V)}\right]^2 \frac{\sigma_1^2(V)[1-\rho(X,V)]}{N_1}\left\{2 + [1+\rho(X,V)]\left[\frac{x_i - \mu(X)}{\sigma(X)}\right]^2\right\} \\
&= \frac{4\sigma^2(Y)[1-\rho(X,V)]}{N_{tot}}\{2 + [1+\rho(X,V)]z_i^2\}. \qquad (16)
\end{aligned}
$$

Letting $Y$ be standardized to having a mean of 0 and a standard deviation of 1,

$$
var[\hat{e}_Y(x_i)] \cong \frac{4[1-\rho(X,V)]}{N_{tot}}\{2 + [1+\rho(X,V)]z_i^2\}. \qquad (17)
$$

This result is also consistent with the equation presented by Lord (1950) for "Case IV" in which test forms X and Y are both equated to the set of common items V (also see Angoff, 1971).

## Example

By substituting the expressions in Equation 11 or 12 for $\rho(X,V)$ in Equation 16 or 17, error variance for equating can be viewed as a function of reliability, $\rho(X,X')$, relative effective test length, $k$, sample size, $N_{tot}$, and standardized score $z_i$. Suppose that there is a desire to estimate the length of the set of common items necessary for the equating to have a certain level of precision over a range of $z$-scores. Assume that reliability of the test is known and that the sample size available for equating is known as well. Equations 16 and 17 can be used to find the approximate number of common items needed to achieve the desired equating precision.

Consider the following example. An external set of common items is to be used. The test contains 50 multiple-choice items. Test reliability is 0.8 and the available sample size for equating is $N_{tot} = 2,000$. Also assume that the target equating precision is a standard error of equating of 0.1 or below over the range of $z$-scores from -3 to 3. Table 3 provides standard errors of equating (square root of error variance) for this situation at various values of $k$ and at various $z$-scores. Based on this table, approximately $k = 0.50$ or greater is necessary to achieve the precision target. Because the test length is 50 items, the common items length should be at least 25 items to achieve the precision target. Note that when using external common items, the relative length of the set of common items can be even longer than the total test.

Now consider that all of the same characteristics hold, except that an internal set of common items is to be used. Based on Table 4, approximately $k = 0.20$ or greater would be needed to achieve the target precision. Thus, an internal set

of common items of at least 10 items would be needed to achieve the precision target.

The values in Table 3 and 4 are shown graphically in Figure 2. As can be seen, for a given value of $k$ the minimum standard error of equating is at a $z$-score of 0, and the more the $z$-scores deviate from 0 the greater is the standard error of equating. For the external common items, the standard errors of equating are clustered together more than for the internal common items. In addition, using the internal set of common items leads to smaller standard errors than using the external set of common items.

# 4 Direct Estimates of Lower Bound of the Number of Common Items Needed

As mentioned in the previous example, in practice, the test developer may need to decide on the relative length of the set of common items that are necessary to provide the desired degree of equating precision. In the previous section, two tables were created that were used to provide an approximate procedure for finding the length of the external and internal set of common items respectively. In this section, a procedure that can be used to directly estimate the length of the common items is developed. It is based on test reliability, $\rho(X, X')$, the sample size available, $N_{tot}$, target SEE in terms of numbers of standard deviation units, $u$, and standardized test scores of interest, $z_i$. Some tables are also created so the test developer can easily deal with a variety of test construction situations.

In general, this process involves four steps.

*Step 1. Specify $N_{tot}$ and $u$.*

It is not surprising that, when the sample size used for equating is large or the test developer has a high tolerance of equating error or both, the target equating precision can be achieved even when relatively shorter sets of common items are used. However, when the sample size available is limited or the desired equating precision is strict, constraints placed on the number of common items become stringent. Then there is a need to estimate the lower bound of the number of common items necessary.

*Step 2. Determine $z_i$.*

Every test is designed to fulfill some specific purposes. As a result, the score range of interest varies from test to test. For example, a test that provides information for selecting scholarship recipients tends to focus more on better-than-average performances, whereas a test that is to be used for a variety of purposes might need precision for a wide range of scores. Accordingly, in terms of standardized score $z_i$, the range of interest might be $1.5 \leq z_i \leq 3$ for the former, whereas it could be $-3 \leq z_i \leq 3$ for the latter. As shown in Figure 2 in the previous example, random equating error reaches its lowest value at $z_i = 0$, and increases as it deviates from the middle scores. Consequently, especially when the score range of interest covers some extreme score values, the number of common items should be large enough to provide the desired equating precision

at these extreme values.

*Step 3. Choose the type of common items, internal or external.*

Type of common items, internal or external common items, may also be a factor when deciding the length of the set of common items needed.

*Step 4. Specify test reliability.*

If test reliability is too low, it might be impossible to find a lower bound of the number of common items needed. Note that under the classical congeneric model, reliability for test form X can be estimated by Feldt's internal consistency coefficient (Feldt & Brennan, 1989) as

$$_F\hat{\rho}_{XX'} = \frac{S_X^2(S_X^2 - \sum S_{X_f}^2)}{S_X^4 - \sum S_{X_f X}^2},$$

where $S_X^2$ is the total score variance, $S_{X_f}^2$ is the variance for individual item $X_f$, and $S_{X_f X}$ is the covariance between individual item $X_f$ and total test score $X$. In practice, however, Cronbach's alpha is routinely reported as an index of test reliability, which is under the essentially tau-equivalent model. Feldt and Brennan (1989) provided an example that compared different reliability coefficient estimates based on the same variance-covariance matrix (see pp. 114–116).

The process for choosing $k$ is represented in the flowchart in Figure **??**, and detailed analytic derivations are provided in the Appendix. As seen in the flowchart, there are five different results regarding the choice of the number of common items necessary to provide the target SEE, and sometimes a result can be reached without going through all four steps. The following set of examples is used to demonstrate the use of the flowchart in practice.

## Example

Consider the example described in the previous section. The sample size available for equating is $N_{tot} = 2,000$, and the target SEE is assumed to be $u = 0.1$ standard deviation units. At Step 1,

$$N_{tot}u^2 = 2000(0.1)^2 = 20 > 8,$$

so branch left and move to Step 2. At Step 2, squared $z$-scores of interest are compared with a criterion,

$$\frac{N_{tot}u^2 - 8}{4} = \frac{2000(0.1)^2 - 8}{4} = 3.$$

Suppose the $z$-score range of interest is from -1.5 to 1.5. Note that every possible $z_i^2 \leq (1.5)^2 = 2.25 < 3$, so branch left again and move to Result 1. That is, under this situation, the target SEE will always be achieved regardless of the value of $k$ assuming that the test satisfies the requirements of the classical congeneric model and randomly equivalent groups are administered Forms X and Y.

Suppose that all of the same characteristics hold, except that the $z$-score range of interest now is from -3 to 3. As a result, at Step 2, some $z_i^2$ can exceed the criterion, so branch right instead, and move to Step 3. Now, the type of common items directly affects the estimation of the lower bound of the number of common items necessary to achieve the target SEE. If external common items are to be used, then test reliability needs to be higher than a criterion,

$$\rho_H^2 = \left( \frac{-1 + \sqrt{z_i^4 - \frac{N_{tot}u^2 - 8}{4z_i^2} + 1}}{z_i^2} \right)^2 \cong 0.51.$$

Otherwise, it is impossible to achieve the target SEE regardless of the choice of the number of common items included (Result 3). Assume that test reliability is 0.8 which is higher than 0.51, and then go to Result 2; that is, the relative length of the set of common items is

$$k \geq \frac{\rho_H^2[1 - \rho(X, X')]}{\rho(X, X')[\rho(X, X') - \rho_H^2]} \cong 0.44.$$

Thus, if the test contains 50 items, at least $0.44(50) = 22$ external common items should be used. If internal common items are to be used, go to Result 4; that is, the relative length of the set of common items is

$$k \geq \frac{\rho_H^2[1 - \rho(X, X')]}{1 - \rho_H^2\rho(X, X')} \cong 0.17.$$

Similarly, for a test including 50 items, at least $0.17(50) = 8.5 \cong 9$ internal common items are necessary. Tables 5 and 6 provide estimates of $k$ using external and internal sets of common items at various combinations of sample sizes, $N_{tot}$, and degree of precision, $u$, where reliability is 0.8, and $z$-scores range from -3 to 3.

# 5 Simulation

In this study, two major assumptions are made in order to derive a simplified form for estimating SEE for the chained linear equating method and a practical process for directly estimating the number of common items needed to achieve desired equating precision. These assumptions are discussed in more detail in this section. Two separate simulation studies are presented. Simulation Study 1 focuses on the accuracy of simplified random error estimation for the chained linear equating when two major assumptions are violated to varying degrees. Simulation Study 2 provides some modifications for using Tables 5 and 6 to choose appropriate numbers of common items when the two major assumptions are violated.

The first assumption is that the two groups used for equating are equivalent in ability, and in particular, that the means and variances of scores on the common items in the two groups are identical. Namely, $\mu_1(V) = \mu_2(V)$ and

$\sigma_1^2(V) = \sigma_2^2(V)$. This assumption is very useful in deriving a simplified form for estimating random error variance for the chained linear equating based on two equating links $X \rightarrow V$ and $V \rightarrow Y$. However, in many situations where common items are involved, the groups taking different forms differ from one another by varying amounts. The group equivalence assumption can be violated slightly or dramatically. In the simulation studies, an effect size parameter is defined to reflect the group difference,

$$ES = \frac{\mu_1(V) - \mu_2(V)}{\sqrt{\frac{N_1\sigma_1^2(V)+N_2\sigma_2^2(V)}{N_1+N_2}}}, \tag{18}$$

where all the parameters involved in the equation are defined in previous sections. When groups are equivalent in terms of equal means, $ES$ is exactly zero. Otherwise, the larger $ES$ is associated with greater group differences. When the sample sizes taking two forms are identical, Equation 18 is simplified to Equation 19 as

$$ES = \frac{\mu_1(V) - \mu_2(V)}{\sqrt{\frac{\sigma_1^2(V)+\sigma_2^2(V)}{2}}}. \tag{19}$$

The second assumption is that score distributions are normal. Specifically, scores on test form X and the common items V in Group 1 follow a bivariate normal distribution, and similarly, scores on test form Y and the common items V in Group 2 follow a bivariate normal distribution. The normality assumption is important for developing a simplified form for estimating random error variance for the chained linear equating. In practice, however, the normality assumption might be violated. In the simulation studies, a lognormal distribution and its translated mirror image are applied to simulate positively and negatively skewed score distributions to assess the impact of violation of the normality assumption.

## 5.1 Simulation Study 1

A crucial equation in this study is a simplified form for estimating the SEE for the chained linear equating method, as shown in Equation 16 or 17. The only difference between Equations 16 and 17 is whether $Y$ is standardized to have a mean of 0 and a standard deviation of 1.

Continue using the scenario that was introduced and discussed in previous sections. That is, the test contains 50 multiple-choice items, test reliability is 0.8, and the available sample size for equating is $N_{tot} = 2,000$, where 1,000 examinees take each form. In Simulation Study 1, the number of common items is fixed at 20, which is 40% of the total test length. Both external and internal sets of common items are considered. The following steps are used to evaluate estimation accuracy of the analytic SEE for the chained linear equating:

1. Take a random sample of size 1,000 from a bivariate distribution of scores on test Form X and common items V in Group 1. Take a random sample

of size 1,000 from a bivariate distribution of scores on test Form Y and common items V in Group 2.

2. Equate Form X to Form Y using the chained linear equating method based on these random samples $r$. The estimated equating relationship is denoted as $\hat{e}_Y^{(r)}(x_i)$.

3. Use Equation 16 to estimate the SEE with statistics based on these random samples $r$ replacing parameters, referred to as $\widehat{SEE}^{(r)}(x_i)$.

4. Repeat steps 1-3 $R$ times ($R = 10,000$). There are two ways of estimating SEE at every possible Form X raw score, $x_i$, denoted as $\widehat{SEE}_a(x_i)$ and $\widehat{SEE}_b(x_i)$ respectively.

   (a) Compute the standard deviation of equated scores as

   $$\widehat{SEE}_a(x_i) = \sqrt{\frac{\sum_{r=1}^{R}[\hat{e}_Y^{(r)}(x_i) - \bar{\hat{e}}_Y(x_i)]^2}{N-1}}, \tag{20}$$

   where $\bar{\hat{e}}_Y(x_i)$ is the average equated score over $R$ replications.

   (b) Average $\widehat{SEE}^{(r)}(x_i)$ over $R$ replications, where $\widehat{SEE}^{(r)}(x_i)$ is computed at Step 3,

   $$\widehat{SEE}_b(x_i) = \frac{1}{R}\widehat{SEE}^{(r)}(x_i). \tag{21}$$

For the two estimates of SEE, $\widehat{SEE}_a(x_i)$ is a straightforward estimated standard error of equating obtained by simulation that is not affected by the analytic estimation process. $\widehat{SEE}_b(x_i)$ represents how Equation 16 might be used in practice when values of population parameters are unknown.

Different degrees of violation of group equivalence and normality assumptions are reflected in the characteristics of population score distributions from which Steps 1 and 2 draw samples. For potential group differences, eight levels of $ES$ are considered: 0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.75, and 1.0. For potential score distribution shapes, three levels are considered: normal, lognormal, and translated mirror image of lognormal, representing normal, positively skewed, and negatively skewed distributions, respectively. In addition, two types of common items, external and internal, are included. In total, there are $8 \times 3 \times 2 = 48$ combinations of conditions. The replication process runs separately for each condition by using R (R Development Core Team, 2005).

## 5.2 Simulation Results 1

Parameters for the simulation are summarized in Table 7. When a positively skewed distribution is used, the skewness of scores on Form X in Group 1 is approximately 0.77, and skewness of scores on Form Y in Group 2 varies from

0.77 to 1.10 as the difference between two populations increases from 0 to 1.0 in terms of $ES$. When a negatively skewed distribution is used, only the direction of skewness changes to negative.

The accuracy of analytic SEE estimation is evaluated by examining the difference between the empirical SEE, $\widehat{SEE}_a(x_i)$, and the analytic SEE, $\widehat{SEE}_b(x_i)$, along the Form X $z$-score scale, where $z_i = \frac{x_i - \mu(X)}{\sigma(X)}$. The difference is reflected in the number of standard deviation units. The ideal value is 0, suggesting that the analytic SEE estimation led to the same result as the empirical SEE. Positive values indicate "overestimation" using the analytic procedure, and negative values are related to "underestimation" using the analytic procedure. The greater the deviation from zero, the less the accuracy of the analytic SEE estimation.

As shown in Figures 4 to 6, within each distribution condition, larger $ES$ is associated with more "bias" in analytic estimation compared to empirical estimation. When the normality assumption holds, the effect of group difference on the accuracy of analytic SEE estimation is not very large. Even when the group difference is as large as $ES = 1.0$, the difference between analytic SEE and empirical SEE is still within -0.02 to 0.02 standard deviation units. However, when score distributions are positively or negatively skewed, the accuracy of analytic SEE is reduced especially when groups are largely different in ability. Holding level of group difference and shapes of score distributions constant, the use of an internal set of common items tends to produce more accurate analytic SEE estimation compared to the use of an external set of common items.

Figure 7 focuses on the effect of shapes of score distributions on the accuracy of analytic SEE estimation. For the six conditions as shown in the figure, groups taking forms X and Y are equivalent, because $ES = 0$. Score distributions are either normal, positively skewed, or negatively skewed. Both external and internal common items are considered. When score distributions are normal, the analytic SEE is very close to the SEE estimated empirically. However, when score distributions are skewed, analytic estimation tends to overestimate the SEE for scores near the middle and underestimate the SEE for scores near either extreme. Using an internal set of common items generally produces more accurate analytic estimation than using an external set of common items.

## 5.3 Simulation Study 2

Tables 5 and 6 in the previous section can provide test developers some practical guidance in choosing the number of common items under certain conditions. Estimates in these tables are obtained by following the direct estimation procedure that is also based on group equivalence and normality assumptions.

Simulation Study 2 intends to examine and modify estimates of relative length of the set of common items, $k$, when two assumptions are violated by different amounts. Three levels of group difference are considered: $ES = 0$, $ES = 0.2$, and $ES = 0.5$. Three shapes of score distribution are considered: normal, moderately positively skewed, and extremely positively skewed. A log-normal transformation is used to simulate the extremely skewed condition where

the skewness is approximately from 0.7 to 0.8. A hybrid lognormal and normal transformation is used to simulate the moderately skewed condition where the skewness is around 0.2 to 0.3. Results for positively skewed distributions are generalizable to those for negatively skewed distributions. Both external and internal common items are used. In total, $3 \times 3 \times 2 = 18$ combinations of conditions are displayed.

Similar to Simulation Study 1, the test still consists of 50 multiple-choice items, test reliability is 0.8, and the available sample size for equating is $N_{tot} = 2,000$. The target SEE, $u$, in terms of numbers of standard deviation units is fixed to be 0.10, and $z$-scores of interest range from -3 to 3. The following steps are followed:

1. Initialize $k$ using the value from Table 5 or 6 depending on which types of common items are used. [1]
2. Run $R$ replications ($R = 10,000$) as described in Simulation Study 1. Only compute the empirical SEE, $\widehat{SEE}_a(x_i)$, where $z_i = \frac{x_i - \mu(X)}{\sigma(X)}$ ranges from -3 to 3.
3. Compare $\widehat{SEE}_a(x_i)$ with $u\sigma(Y)$. If Table 5 or 6 works perfectly, the SEE at the $z$-scores of interest, which is from -3 to 3, should fall below $u$ standard deviation units, which is $u\sigma(Y)$. Otherwise, because as $k$ increases, the SEE will decrease, a larger value of $k$ will be considered.
   (a) If $max_{-3 \le z_i \le 3} \widehat{SEE}_a(x_i) < u\sigma(Y)$, stop and report $k$ as the modified relative length of the set of common items.
   (b) if $max_{-3 \le z_i \le 3} \widehat{SEE}_a(x_i) \ge u\sigma(Y)$, add 0.02 to current $k$ (i.e., use $50 \times 0.02 = 1$ additional common item). For an internal set of common items, if the updated $k$ exceeds 1.0, stop and report an error message. Otherwise, go back to Step 2.
4. Repeat steps 1-3 $T$ times ($T = 20$) to stabilize the estimation of $k$.

The replication process was conducted using R (R Development Core Team, 2005).

## 5.4 Simulation Results 2

Tables 8 and 9 contain summaries of modified $k$ values and corresponding numbers of common items needed under the 18 different simulation conditions. Figure 8 directly illustrate the result as shown in Table 9. As expected, when two major assumptions both hold, values from Tables 5 and 6 lead to the target SEE. When a set of external common items is used, as the condition moves from

[1] To decide the initial value of $k$ for some conditions, instead of using Table 5 or 6 that might be inefficient and time-consuming, several additional simulation runs were done first to gather a rough idea of the relationship between $k$ and empirical SEE (similar to Simulation Study 1).

ideal (i.e., $ES = 0$, normal) to extreme (i.e., $ES = 0.5$, extremely skewed), the number of common items that is necessary to meet the target SEE steadily increases. When a set of internal common items is used, violation of the normality assumption tends to affect the number of common items needed more dramatically than violation of the group equivalence assumption. For example, when the normality assumption holds, even as two groups differ from each other as much as $ES = 0.5$, only two additional common items are needed to achieve the target SEE. In general, compared to the use of a set of external common items, the use of a set of internal common items leads to more stable estimation of the number of common items necessary, although violation of both assumptions still requires the number of common items needed to be approximately twice as many as to achieve the same target SEE.

# 6 Discussion

Most previous studies on numbers of common items only provided exploratory results concerning the relationship between the numbers of common items and equating precision. These results were applicable to specific tests, situations, and limited conditions of lengths for the common items. In this study a novel way of understanding the relationship between number of common items and equating precision is provided, by combining the classical congeneric model with analytic standard errors derived by the delta method. This study describes a process along with some figures and tables that can be used by the test developers to choose the length of the common item set that leads to the desired equating precision under various real test situations.

For both external and internal common items, the relationship between test score reliability for the total score on a test, the effective test length of the common items, and the correlation between test scores and scores on common items was derived analytically in this study using the classical congeneric model. These relationships show clearly that as reliability and effective test length for the common items increase, the correlation between total test and common item scores increases. These derivations were used to illustrate how the standard error of equating for chained linear equating is related directly to test reliability and to the effective test length of the common items. In addition, a process was developed to estimate the number of common items needed for a specified degree of equating precision for chained linear equating.

Two points are worth noting when the estimated lower bound for the number of common items is used in practice. Theoretically, the estimated lower bound of the relative length of common items provided in this study is always a ratio of effective test lengths, $k = \frac{\lambda_V}{\lambda_X}$. In some situations, this ratio can be viewed, approximately, as the ratio of actual number of items, as described in the examples in this study. In other situations, the relationship between the ratio of effective test lengths and ratio of actual test lengths needs careful consideration, such as when a test contains both multiple-choice and constructed-response questions. In addition, as $k$ approaches 0, the two test forms X and Y share few, if any,

items in common. However, small values of $k$ might lead to some problems in applying these methods. One such problem is that, when the number of common items is very small, content and statistical representativeness are difficult to maintain. In practice, the number of common items should, at a minimum, be large enough to adequately represent the content of the total test.

Two simulation studies were used to empirically check the accuracy of the simplified form for estimating SEE for the chained linear equating method and the process for directly estimating the number of common items needed to achieve desired equating precision when the group equivalence and normality assumptions are violated. It appears that violation of the normality assumption can lead to the analytic SEE being a substantial underestimate of the SEE and number of common item required to meet target precision being substantially underestimated by the simplified procedure. When the normality assumption holds and the group differences are greater than zero, the analytic SEE is a slight underestimate of the SEE and the number of common items required to meet target precision is slightly underestimated. Overall, it appears that the simplified process described in this study for estimating the SEE and the number of common items needed to meet the target precision is reasonably accurate when the scores are close to being normally distributed and the group differences are not large.

# 7 Appendix

Larger numbers of common items generally provide greater equating precision. Thus, estimation of the lower bound of the relative length of common items required by the specified SEE is important. The following derivation consists of two parts. First, the expected correlation between the scores on the total test and scores on the common items, $\rho(X,V)$, given the specified situation is evaluated. Next, the lower bound of the relative length of the common item set, $k$, under the classical congeneric model is estimated. Various equating methods may perform differently in terms of random error. The chained linear method is examined in this study.

## 7.1 Evaluating the Expected Correlation, $\rho(X,V)$

Let $u$ index the target SEE in terms of numbers of standard deviation unit such that $var[\hat{e}_Y(x_i)] \leq u^2\sigma^2(Y)$. Specifically, according to Equation 17, when $Y$ is standardized to have a mean of 0 and a standard deviation of 1,

$$\begin{aligned} var[\hat{e}_Y(x_i)] \leq u^2 &\Leftrightarrow \frac{4[1-\rho(X,V)]}{N_{tot}}\{2+[1+\rho(X,V)]z_i^2\} \leq u^2 \\ &\Leftrightarrow \frac{1}{N_{tot}}\{-4z_i^2\rho^2(X,V) - 8\rho(X,V) + (8+4z_i^2)\} \leq u^2 \\ &\Leftrightarrow \frac{z_i^2}{2}\rho^2(X,V) + \rho(X,V) + \frac{(N_{tot}u^2-8)-4z_i^2}{8} \geq 0. \end{aligned} \quad (22)$$

The final expression in Equation 22 is a quadratic inequality of $\rho(X,V)$ when $z_i \neq 0$. The purpose of this step is to determine possible values of $\rho(X,V)$ which make the inequality in Equation 22 true. The relationship between the determinant and the roots of a quadratic function plays an important role.

The determinant of the inequality in Equation 22 is

$$\Delta_1 = 1^2 - 4 \cdot \frac{z_i^2}{2} \cdot \frac{(N_{tot}u^2 - 8) - 4z_i^2}{8} = z_i^4 - \frac{N_{tot}u^2 - 8}{4} z_i^2 + 1. \tag{23}$$

Note that $\Delta_1$ is another quadratic function of $z_i^2$, and its determinant is

$$\Delta_2 = \left(\frac{N_{tot}u^2 - 8}{4}\right)^2 - 4 \cdot 1 \cdot 1 = \frac{N_{tot}u^2(N_{tot}u^2 - 16)}{16}. \tag{24}$$

Three conditions are discussed to explore possible values of $\rho(X,V)$.

**Condition 1: $N_{tot}u^2 > 16$.** Under this condition, $\Delta_2$ is always positive, and consequently, $\Delta_1$ has two distinct roots as

$$(z_i^2)_L \equiv \frac{\frac{N_{tot}u^2-8}{4} - \sqrt{\Delta_2}}{2} = \frac{(N_{tot}u^2 - 8) - 4\sqrt{\Delta_2}}{8}, \tag{25}$$

and

$$(z_i^2)_H \equiv \frac{\frac{N_{tot}u^2-8}{4} + \sqrt{\Delta_2}}{2} = \frac{(N_{tot}u^2 - 8) + 4\sqrt{\Delta_2}}{8}. \tag{26}$$

It can be shown that both $(z_i^2)_L$ and $(z_i^2)_H$ are positive. Furthermore, if $(z_i^2)_L \leq z_i^2 \leq (z_i^2)_H$, then $\Delta_1 \leq 0$ and the inequality in Equation 22 is always true. Otherwise, if $0 < z_i^2 < (z_i^2)_L$ or $z_i^2 > (z_i^2)_H$, then $\Delta_1 > 0$ and the inequality in Equation 22 has two distinct roots as

$$\rho_L \equiv \frac{-1 - \sqrt{\Delta_1}}{z_i^2}, \tag{27}$$

and

$$\rho_H \equiv \frac{-1 + \sqrt{\Delta_1}}{z_i^2}. \tag{28}$$

The inequality in Equation 22 holds only if $\rho(X,V) \leq \rho_L$ or $\rho(X,V) \geq \rho_H$. Since $\rho_L$ is always negative whereas the correlation between the scores on the total test and scores on the common items, $\rho(X,V)$, is expected to be positive for a well-designed test, possible values of $\rho(X,V)$ should be no less than $\rho_H$. The sign of $\rho_H$ decides whether this requirement is trivial. Specifically,

$$\begin{aligned}
\rho_H &\equiv \frac{-1+\sqrt{\Delta_1}}{z_i^2} > 0 \\
&\Leftrightarrow -1+\sqrt{\Delta_1} > 0 \\
&\Leftrightarrow \Delta_1 > 1 \\
&\Leftrightarrow z_i^4 - \frac{N_{tot}u^2-8}{4} z_i^2 > 0 \\
&\Leftrightarrow z_i^2 > \frac{N_{tot}u^2-8}{4}. \qquad (29)
\end{aligned}$$

Thus, if $z_i^2 > \frac{N_{tot}u^2-8}{4}$, then $\rho(X,V)$ is expected to be larger than $\rho_H$. Otherwise, because $\rho(X,V)$ is expected to be positive, the inequality $\rho(X,V) > 0 \geq \rho_H$ always holds. It can be shown that either $(z_i^2)_L$ or $(z_i^2)_H$ is smaller than $\frac{N_{tot}u^2-8}{4}$, so the above results with regard to the intervals of $z_i^2$ can be combined and simplified.

In sum, under Condition 1, the specified precision is obtained for $-\frac{\sqrt{N_{tot}u^2-8}}{2} \leq z_i \leq \frac{\sqrt{N_{tot}u^2-8}}{2}$ as long as the test is well developed. For either $z_i < -\frac{\sqrt{N_{tot}u^2-8}}{2}$ or $z_i > \frac{\sqrt{N_{tot}u^2-8}}{2}$, the specified precision can only be achieved as the expected correlation between the scores on the total test and scores on the common items, $\rho(X,V)$, exceeds $\rho_H$.

**Condition 2:** $8 < N_{tot}u^2 \leq 16$**.** Under this condition, $\Delta_2$ is no longer positive. As a result, $\Delta_1 \geq 0$ and there always exist two distinct roots of the inequality in Equation 22, $\rho_L$ and $\rho_H$. Again, for $-\frac{\sqrt{N_{tot}u^2-8}}{2} \leq z_i \leq \frac{\sqrt{N_{tot}u^2-8}}{2}$, the equating precision does not depend heavily on slight changes in $\rho(X,V)$ as long as the test is well developed, whereas for $z_i < -\frac{\sqrt{N_{tot}u^2-8}}{2}$ or $z_i > \frac{\sqrt{N_{tot}u^2-8}}{2}$, the specified precision required $\rho(X,V) \geq \rho_H$.

**Condition 3:** $0 < N_{tot}u^2 \leq 8$**.** Same as Condition 2, there exist two distinct roots of the inequality in Equation 22, and the sign of $\rho_H$ directly affect the possible values of $\rho(X,V)$. However, as $N_{tot}u^2 \leq 8$, the final inequality in Equation 29 is always true with $z_i \neq 0$, such that $\rho_H > 0$. Thus, to provide the specified equating precision, $\rho(X,V)$ needs to exceed $\rho_H$.

Next, corresponding relative lengths of the common item set are estimated.

## 7.2 Estimating the relative length of common items, $k$

According to the discussion in previous subsection, sometimes the target equating precision can be achieved as long as the test forms are well constructed, whereas other times, $\rho(X,V)$ needs to be no less than $\rho_H$ which is a positive value after test reliability, the sample size available, the degree of precision desired, and the standardized score range have been clearly specified. Estimation

of $k$ for three different conditions defined in previous subsection are very similar. Using external or internal common items always leads to different estimates.

For an external set of common items, substitute Equation 11 in $\rho(X,V) \geq \rho_H$,

$$\begin{aligned} \rho(X,V) \geq \rho_H &\Leftrightarrow \rho^2(X,V) \geq \rho_H^2 \\ &\Leftrightarrow \frac{k\rho^2(X,X')}{1+(k-1)\rho(X,X')} \geq \rho_H^2 \\ &\Leftrightarrow \rho(X,X')[\rho(X,X') - \rho_H^2]k \geq \rho_H^2[1-\rho(X,X')]. \end{aligned} \quad (30)$$

Note that if $\rho(X,X') \leq \rho_H^2$, the final equality in Equation 30 never holds. In other words, in some situations where $\rho(X,X') \leq \rho_H^2$, the specified equating precision cannot be obtained no matter how many common items are included. The test developer needs to redesign the test or may be able to modify the situation such as increasing the sample size. If $\rho(X,X') > \rho_H^2$, the lower bound of the relative length of the set of common items is

$$k \geq \frac{\rho_H^2[1-\rho(X,X')]}{\rho(X,X')[\rho(X,X') - \rho_H^2]}. \quad (31)$$

For an internal set of common items, substitute Equation 12 in $\rho(X,V) \geq \rho_H$,

$$\begin{aligned} \rho(X,V) \geq \rho_H &\Leftrightarrow \rho^2(X,V) \geq \rho_H^2 \\ &\Leftrightarrow \frac{k}{1+(k-1)\rho(X,X')} \geq \rho_H^2 \\ &\Leftrightarrow k \geq \frac{\rho_H^2[1-\rho(X,X')]}{1-\rho_H^2\rho(X,X')}. \end{aligned} \quad (32)$$

In theory, the internal set of common items can be lengthened to be the total test form eventually, so it is not surprising that any specified equating precision can be provided.

Although the mathematical procedures dealing with Condition 1 and Condition 2 are different, the final results turn out to be identical, so these two conditions are combined into a single condition, $N_{tot}u^2 > 8$, for simplicity.

If $z_i = 0$, the inequality in Equation 22 is linear rather than quadratic. The whole procedure is analogous but can be simplified. The same conditions are considered.

**Condition 1: $N_{tot}u^2 > 8$.** Under this condition, the equating precision does not heavily depend on varying numbers of common items for a well-developed test.

**Condition 2:** $0 < N_{tot}u^2 \leq 8$**.** Under this condition, if external common items are used and reliability of the total test exceeds $\frac{(8-N_{tot}u^2)^2}{64}$, the specified precision is satisfied by choosing

$$k \geq \frac{(8 - N_{tot}u^2)^2[1 - \rho(X, X')]}{\rho(X, X')[64\rho(X, X') - (8 - N_{tot}u^2)^2]}. \tag{33}$$

If internal common items are used, the precision is always achieved by choosing

$$k \geq \frac{(8 - N_{tot}u^2)^2[1 - \rho(X, X')]}{64 - (8 - N_{tot}u^2)^2\rho(X, X')}. \tag{34}$$

# 8 References


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

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

Brennan, R. L. (2006). *Chained linear equating* (CASMA Technical Note No. 3). Iowa City, IA: Center for Advanced Studies in Measurement and Assessment, The University of Iowa.

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

Holland, P. W., & Dorans, N. J. (2006). Linking and equating. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., 187–220). Westport, CT: Praeger.

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

Lord, F. M. (1950). *Notes on comparable scales for test scores* (Research Bulletin 5048). Princeton, NJ: Educational Testing Service.

Puhan, G. (2010). A comparison of chained linear and poststratification linear equating under different testing conditions. *Journal of Educational Measurement, 47*, 54–75.

R Development Core Team (2005). R: A language and environment for statistical computing, reference index version 2.1.1. [Computer software]. Vienna, Austria: R Foundation for Statistical Computing.

Ricker, K. L., & von Davier, A. A. (2007). *The impact of anchor test length on equating results in a nonequivalent groups design* (ETS Research Report 07-44). Princeton, NJ: Educational Testing Service.

Wang, T., Lee, W., Brennan, R. L., & Kolen, M. J. (2008). A comparison of the frequency estimation and chained equipercentile methods under the common-item nonequivalent groups design. *Applied Psychological Measurement, 32*, 632–651.

Yang, W., & Houang, R. T. (April, 1996). *The effect of anchor length and equating method on the accuracy of test equating: Comparisons of linear and IRT-based equating using an anchor-item design.* Paper presented at the annual meeting of the American Educational Research Association, New York.

Table 1: $\rho^2(X, V)$ as a function of $\rho(X, X')$ and $k$ using external common items

| | $\rho(X, X')$ | | | | | | |
|---|---|---|---|---|---|---|---|
| $k = \frac{\lambda_V}{\lambda_X}$ | 0.70 | 0.75 | 0.80 | 0.85 | 0.90 | 0.95 | 0.99 |
| 0.10 | 0.1324 | 0.1731 | 0.2286 | 0.3075 | 0.4263 | 0.6224 | 0.8992 |
| 0.20 | 0.2227 | 0.2813 | 0.3556 | 0.4516 | 0.5786 | 0.7521 | 0.9424 |
| 0.30 | 0.2882 | 0.3553 | 0.4364 | 0.5352 | 0.6568 | 0.8082 | 0.9578 |
| 0.40 | 0.3379 | 0.4091 | 0.4923 | 0.5898 | 0.7044 | 0.8395 | 0.9656 |
| 0.50 | 0.3769 | 0.4500 | 0.5333 | 0.6283 | 0.7364 | 0.8595 | 0.9704 |
| 0.60 | 0.4083 | 0.4821 | 0.5647 | 0.6568 | 0.7594 | 0.8734 | 0.9736 |
| 0.70 | 0.4342 | 0.5081 | 0.5895 | 0.6789 | 0.7767 | 0.8836 | 0.9759 |
| 0.80 | 0.4558 | 0.5294 | 0.6095 | 0.6964 | 0.7902 | 0.8914 | 0.9777 |
| 0.90 | 0.4742 | 0.5473 | 0.6261 | 0.7107 | 0.8011 | 0.8975 | 0.9790 |
| 1.00 | 0.4900 | 0.5625 | 0.6400 | 0.7225 | 0.8100 | 0.9025 | 0.9801 |

Table 2: $\rho^2(X, V)$ as a function of $\rho(X, X')$ and $k$ using internal common items

| | $\rho(X, X')$ | | | | | | |
|---|---|---|---|---|---|---|---|
| $k = \frac{\lambda_V}{\lambda_X}$ | 0.70 | 0.75 | 0.80 | 0.85 | 0.90 | 0.95 | 0.99 |
| 0.10 | 0.2703 | 0.3077 | 0.3571 | 0.4255 | 0.5263 | 0.6897 | 0.9174 |
| 0.20 | 0.4546 | 0.5000 | 0.5556 | 0.6250 | 0.7143 | 0.8333 | 0.9615 |
| 0.30 | 0.5882 | 0.6316 | 0.6818 | 0.7407 | 0.8108 | 0.8955 | 0.9772 |
| 0.40 | 0.6897 | 0.7273 | 0.7692 | 0.8163 | 0.8696 | 0.9302 | 0.9852 |
| 0.50 | 0.7692 | 0.8000 | 0.8333 | 0.8696 | 0.9091 | 0.9524 | 0.9901 |
| 0.60 | 0.8333 | 0.8571 | 0.8824 | 0.9091 | 0.9375 | 0.9677 | 0.9934 |
| 0.70 | 0.8861 | 0.9032 | 0.9211 | 0.9396 | 0.9589 | 0.9790 | 0.9957 |
| 0.80 | 0.9302 | 0.9412 | 0.9524 | 0.9639 | 0.9756 | 0.9877 | 0.9975 |
| 0.90 | 0.9677 | 0.9730 | 0.9783 | 0.9836 | 0.9890 | 0.9945 | 0.9989 |
| 1.00 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 |

Table 3: Estimated standard errors of equating using external common items ($\rho(X, X') = 0.8$, $N_{tot} = 2,000$)

| | $z_i$ | | | | | | |
|---|---|---|---|---|---|---|---|
| $k = \frac{\lambda_V}{\lambda_X}$ | 0 | ±0.5 | ±1.0 | ±1.5 | ±2.0 | ±2.5 | ±3.0 |
| 0.10 | 0.0457 | 0.0497 | 0.0603 | 0.0746 | 0.0909 | 0.1083 | 0.1264 |
| 0.20 | 0.0402 | 0.0440 | 0.0539 | 0.0672 | 0.0823 | 0.0983 | 0.1150 |
| 0.30 | 0.0369 | 0.0405 | 0.0499 | 0.0624 | 0.0766 | 0.0917 | 0.1073 |
| 0.40 | 0.0346 | 0.0380 | 0.0470 | 0.0590 | 0.0725 | 0.0868 | 0.1017 |
| 0.50 | 0.0329 | 0.0362 | 0.0449 | 0.0564 | 0.0694 | 0.0831 | 0.0974 |
| 0.60 | 0.0315 | 0.0348 | 0.0432 | 0.0543 | 0.0669 | 0.0802 | 0.0940 |
| 0.70 | 0.0305 | 0.0337 | 0.0418 | 0.0527 | 0.0649 | 0.0779 | 0.0912 |
| 0.80 | 0.0296 | 0.0328 | 0.0407 | 0.0513 | 0.0633 | 0.0759 | 0.0889 |
| 0.90 | 0.0289 | 0.0320 | 0.0398 | 0.0502 | 0.0619 | 0.0742 | 0.0870 |
| 1.00 | 0.0283 | 0.0313 | 0.0390 | 0.0492 | 0.0607 | 0.0728 | 0.0853 |

Table 4: Estimated standard errors of equating using internal common items ($\rho(X, X') = 0.8$, $N_{tot} = 2,000$)

| | $z_i$ | | | | | | |
|---|---|---|---|---|---|---|---|
| $k = \frac{\lambda_V}{\lambda_X}$ | 0 | ±0.5 | ±1.0 | ±1.5 | ±2.0 | ±2.5 | ±3.0 |
| 0.10 | 0.0401 | 0.0439 | 0.0538 | 0.0671 | 0.0822 | 0.0982 | 0.1148 |
| 0.20 | 0.0319 | 0.0352 | 0.0437 | 0.0549 | 0.0676 | 0.0811 | 0.0950 |
| 0.30 | 0.0264 | 0.0293 | 0.0365 | 0.0461 | 0.0569 | 0.0684 | 0.0802 |
| 0.40 | 0.0222 | 0.0246 | 0.0309 | 0.0391 | 0.0484 | 0.0581 | 0.0682 |
| 0.50 | 0.0187 | 0.0208 | 0.0261 | 0.0331 | 0.0410 | 0.0493 | 0.0579 |
| 0.60 | 0.0156 | 0.0174 | 0.0219 | 0.0278 | 0.0344 | 0.0414 | 0.0486 |
| 0.70 | 0.0127 | 0.0142 | 0.0179 | 0.0227 | 0.0282 | 0.0339 | 0.0398 |
| 0.80 | 0.0098 | 0.0110 | 0.0138 | 0.0176 | 0.0219 | 0.0263 | 0.0309 |
| 0.90 | 0.0066 | 0.0074 | 0.0093 | 0.0119 | 0.0148 | 0.0178 | 0.0209 |
| 1.00 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 |

Table 5: Estimated lower bound of relative length of the external set of common items ($\rho(X, X') = 0.8$, $-3 \leq z_i \leq 3$)

| | $u$ | | | | | |
|---|---|---|---|---|---|---|
| $N_{tot}$ | 0.05 | 0.10 | 0.15 | 0.20 | 0.25 | $\geq 0.30$ |
| 500 | | | | 0.44 | 0.11 | $< .01$ |
| 550 | | | | 0.34 | 0.07 | $< .01$ |
| 600 | | | | 0.27 | 0.04 | $< .01$ |
| 650 | | | 0.99 | 0.21 | 0.02 | $< .01$ |
| 700 | | | 0.81 | 0.17 | $< .01$ | $< .01$ |
| 750 | | | 0.67 | 0.13 | $< .01$ | $< .01$ |
| 800 | | | 0.57 | 0.10 | $< .01$ | $< .01$ |
| 850 | | | 0.49 | 0.07 | $< .01$ | $< .01$ |
| 900 | | | 0.42 | 0.05 | $< .01$ | $< .01$ |
| 950 | | | 0.37 | 0.03 | $< .01$ | $< .01$ |
| 1000 | | | 0.32 | 0.02 | $< .01$ | $< .01$ |
| 1200 | | | 0.19 | $< .01$ | $< .01$ | $< .01$ |
| 1400 | | | 0.11 | $< .01$ | $< .01$ | $< .01$ |
| 1600 | | 0.78 | 0.05 | $< .01$ | $< .01$ | $< .01$ |
| 1800 | | 0.57 | 0.02 | $< .01$ | $< .01$ | $< .01$ |
| 2000 | | 0.44 | $< .01$ | $< .01$ | $< .01$ | $< .01$ |
| 2200 | | 0.34 | $< .01$ | $< .01$ | $< .01$ | $< .01$ |
| 2400 | | 0.27 | $< .01$ | $< .01$ | $< .01$ | $< .01$ |
| 2600 | | 0.21 | $< .01$ | $< .01$ | $< .01$ | $< .01$ |
| 2800 | | 0.17 | $< .01$ | $< .01$ | $< .01$ | $< .01$ |
| 3000 | | 0.13 | $< .01$ | $< .01$ | $< .01$ | $< .01$ |

*Note.* Blank areas indicate that the precision target can never been achieved regardless of the numbers of common items included. Some bounds may be too low to maintain the content and statistical representativeness and need to be used with caution.

Table 6: Estimated lower bound of relative length of the internal set of common items ($\rho(X, X') = 0.8$, $-3 \leq z_i \leq 3$)

| | $u$ | | | | | |
|---|---|---|---|---|---|---|
| $N_{tot}$ | 0.05 | 0.10 | 0.15 | 0.20 | 0.25 | $\geq 0.30$ |
| 500 | 0.86 | 0.58 | 0.34 | 0.17 | 0.06 | $< .01$ |
| 550 | 0.85 | 0.56 | 0.31 | 0.15 | 0.04 | $< .01$ |
| 600 | 0.84 | 0.53 | 0.28 | 0.12 | 0.02 | $< .01$ |
| 650 | 0.83 | 0.51 | 0.26 | 0.10 | 0.01 | $< .01$ |
| 700 | 0.81 | 0.49 | 0.24 | 0.09 | $< .01$ | $< .01$ |
| 750 | 0.80 | 0.47 | 0.22 | 0.07 | $< .01$ | $< .01$ |
| 800 | 0.79 | 0.45 | 0.20 | 0.06 | $< .01$ | $< .01$ |
| 850 | 0.78 | 0.43 | 0.18 | 0.04 | $< .01$ | $< .01$ |
| 900 | 0.77 | 0.41 | 0.17 | 0.03 | $< .01$ | $< .01$ |
| 950 | 0.76 | 0.39 | 0.15 | 0.02 | $< .01$ | $< .01$ |
| 1000 | 0.75 | 0.38 | 0.14 | 0.01 | $< .01$ | $< .01$ |
| 1200 | 0.71 | 0.32 | 0.09 | $< .01$ | $< .01$ | $< .01$ |
| 1400 | 0.68 | 0.27 | 0.06 | $< .01$ | $< .01$ | $< .01$ |
| 1600 | 0.64 | 0.23 | 0.03 | $< .01$ | $< .01$ | $< .01$ |
| 1800 | 0.61 | 0.20 | 0.01 | $< .01$ | $< .01$ | $< .01$ |
| 2000 | 0.58 | 0.17 | $< .01$ | $< .01$ | $< .01$ | $< .01$ |
| 2200 | 0.56 | 0.15 | $< .01$ | $< .01$ | $< .01$ | $< .01$ |
| 2400 | 0.53 | 0.12 | $< .01$ | $< .01$ | $< .01$ | $< .01$ |
| 2600 | 0.51 | 0.10 | $< .01$ | $< .01$ | $< .01$ | $< .01$ |
| 2800 | 0.49 | 0.09 | $< .01$ | $< .01$ | $< .01$ | $< .01$ |
| 3000 | 0.47 | 0.07 | $< .01$ | $< .01$ | $< .01$ | $< .01$ |

*Note.* Some bounds may be too low to maintain the content and statistical representativeness and need to be used with caution.

Table 7: Parameters for Simulation Study 1

| Parameter Name | Value |
|---|---|
| **Test information** | |
| Number of items on Form X | 50 |
| Number of items on Form Y | 50 |
| Relative length of the common item set V, $k$ | 0.4 |
| Number of common items | 20 |
| Reliability coefficient, $\rho(X, X') = \rho(Y, Y')$ | 0.8 |
| **Population where Group 1 is from** | |
| Mean score on common items, $\mu_1(V)$ | 10 |
| Variance of scores on common items, $\sigma_1^2(V)$ | 9 |
| Mean score on Form X, $\mu(X) = \frac{\mu_1(V)}{k}$ [a] | 25 |
| Variance of scores on Form X, $\sigma^2(X)$ [Equation 8] | 43.26923 |
| Covariance between X and V, $\sigma(X, V)$ [Equation 9 or 10] | 13.84615 or 17.30769[b] |
| **Population where Group 2 is from** | |
| Mean score on common items, $\mu_2(V)$ [Equation 19] | Varies[c] |
| Variance of scores on common items, $\sigma_2^2(V)$ | 9 |
| Mean score on Form Y, $\mu(Y) = \frac{\mu_2(V)}{k}$ | Varies |
| Variance of scores on Form Y, $\sigma^2(Y)$ [Equation 8] | 43.26923 |
| Covariance between Y and V, $\sigma(Y, V)$ [Equation 9 or 10] | 13.84615 or 17.30769 |
| **Sample Size** | |
| Number of examinees taking Form X | 1,000 |
| Number of examinees taking Form Y | 1,000 |
| **Group Differences** | |
| Effect size, $ES$ | Varies |
| **Non-Normality** | |
| Lognormal transformation | |

[a] The equation holds when the ratio of effective test lengths equals the ratio of actual test lengths.
[b] Covariance between $X$ and $V$ varies when different types of common items are used. For an external common items, it is 13.84615 by using Equation 9, and for an internal common items, it is 17.30769 by using Equation 10. Similar results apply to $Y$ and $V$.
[c] Amount of group difference is reflected by mean score difference between two populations on common items. To obtain $ES$ levels of 0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.75, and 1.0, the mean for Group 2 is lower than the mean for Group 1 by 3 $ES$ score points according to Equation 19.

Table 8: Modified $k$ for Tables 5 and 6 based on Simulation Study 2 ($\rho(X, X') = 0.8$, $N_{tot} = 2,000$, $u = 0.10$)

| | $ES = 0$ | $ES = 0.2$ | $ES = 0.5$ |
|---|---|---|---|
| **Using a set of external common items** | | | |
| Normal | 0.4530 (0.0149) | 0.5250 (0.0128) | 0.6490 (0.0165) |
| Moderately Skewed | 0.5920 (0.0136) | 0.6540 (0.0114) | 0.7670 (0.0149) |
| Extremely Skewed | 0.7140 (0.0131) | 0.7750 (0.0089) | 0.8630 (0.0163) |
| **Using a set of internal common items** | | | |
| Normal | 0.1820 (0.0062) | 0.2000 (0.0000) | 0.2210 (0.0045) |
| Moderately Skewed | 0.2800 (0.0000) | 0.3000 (0.0000) | 0.3240 (0.0082) |
| Extremely Skewed | 0.3620 (0.0062) | 0.3800 (0.0000) | 0.4090 (0.0102) |

*Note.* Because the test in simulation studies contains 50 items, $k = 0.18$ which would give an integer solution for the number of common items needed to provide the target SEE is used as initial value when internal common items are considered. Values in parenthesis are standard deviations of $k$ over 20 replications.

Table 9: Modified numbers of common items needed based on Simulation Study 2 (total number of items on either Form X or Y is 50, $\rho(X, X') = 0.8$, $N_{tot} = 2,000$, $u = 0.10$)

| | $ES = 0$ | $ES = 0.2$ | $ES = 0.5$ |
|---|---|---|---|
| **Using a set of external common items** | | | |
| Normal | 23 | 27 | 33 |
| Moderately Skewed | 30 | 33 | 39 |
| Extremely Skewed | 36 | 39 | 44 |
| **Using a set of internal common items** | | | |
| Normal | 10 | 10 | 12 |
| Moderately Skewed | 14 | 15 | 17 |
| Extremely Skewed | 19 | 19 | 21 |

*Note.* Numbers of common items needed are calculated by $50 \times k$, where $k$'s are displayed in Table 8. If the resulting number is not an integer, find the closest integer that is bigger than it.

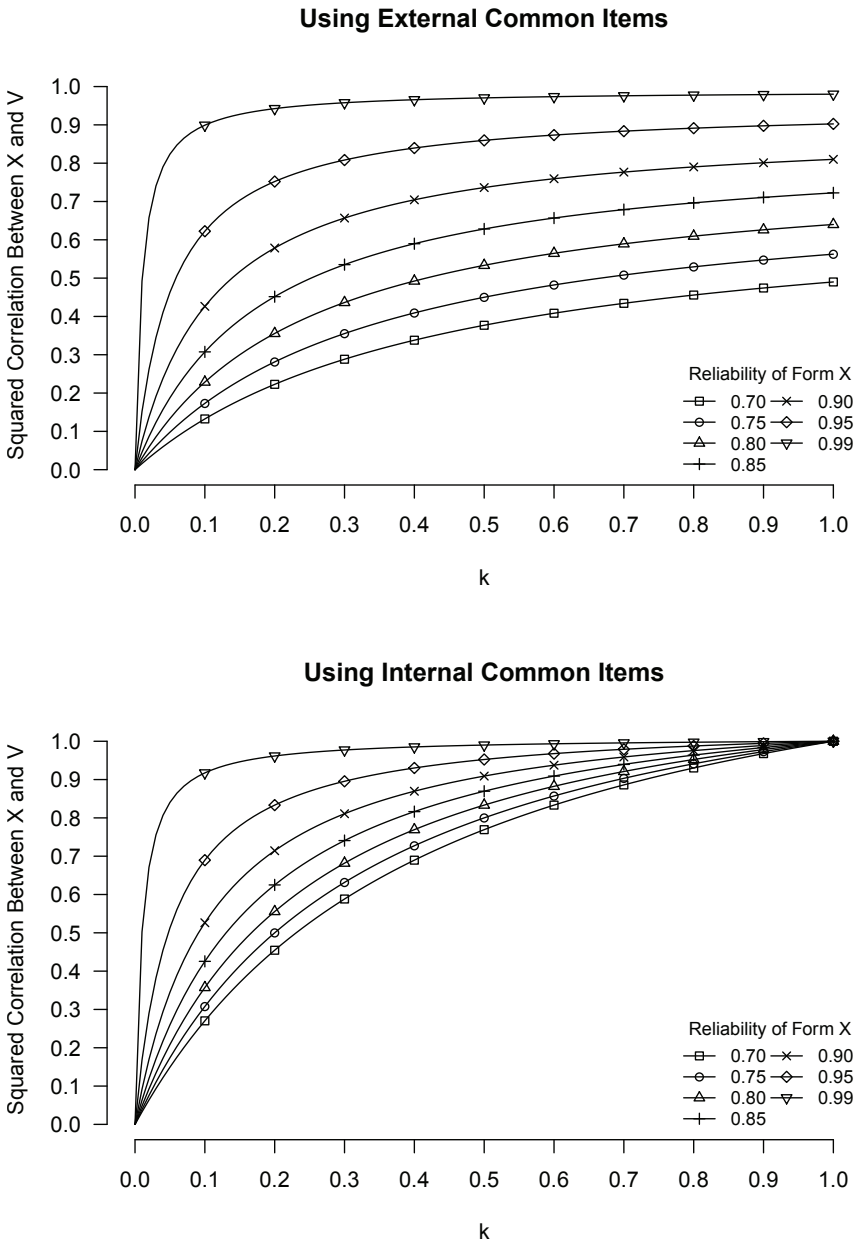


Figure 1: $\rho^2(X, V)$ as a function of $\rho(X, X')$ and $k$

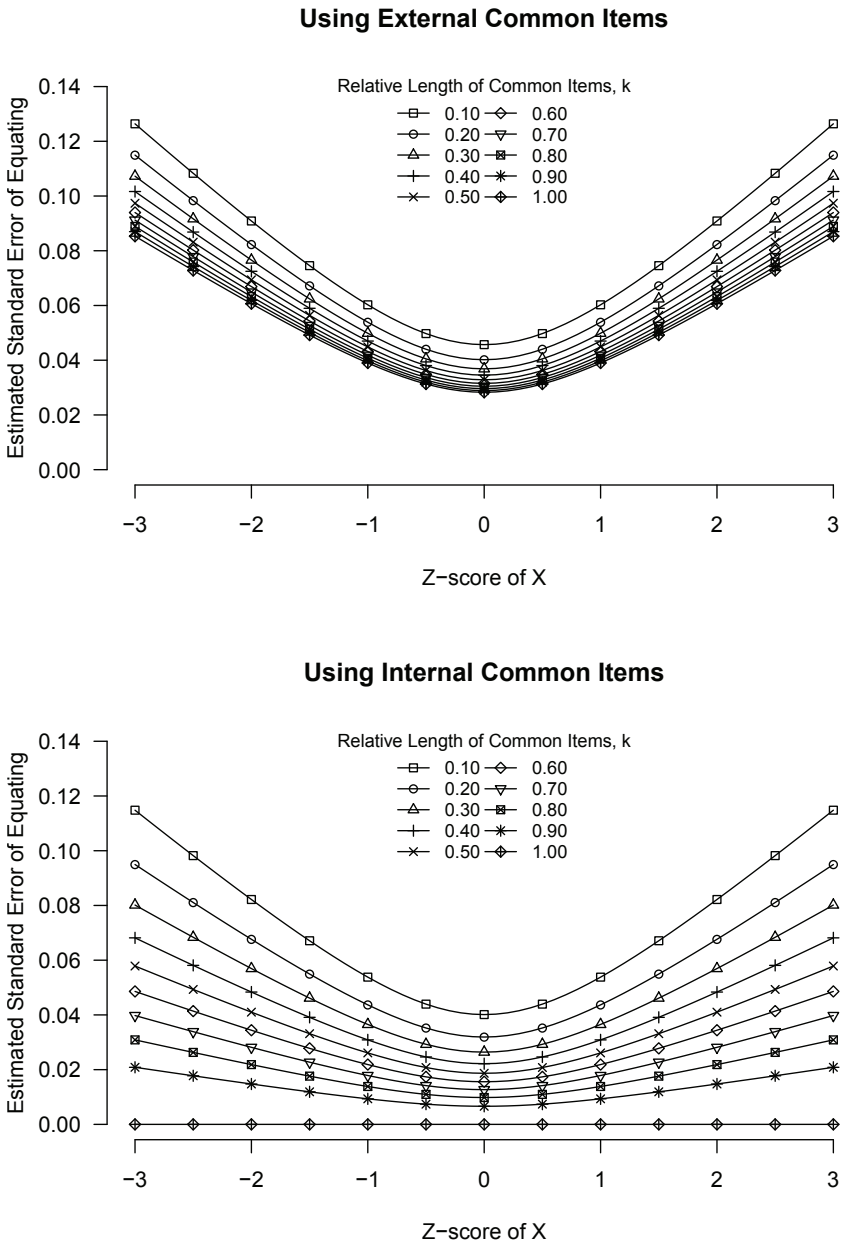


Figure 2: Estimated SEE ($\rho(X, X')$=0.8, $N_{tot}$=2,000)

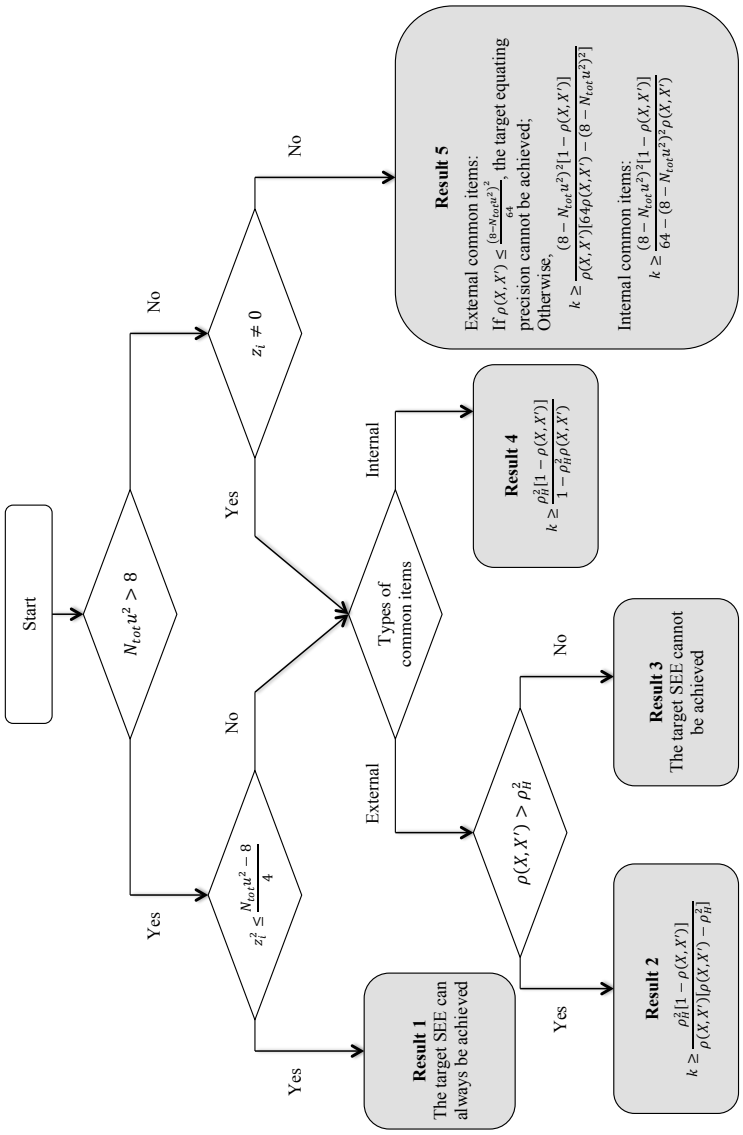


Figure 3: Flowchart of a process for estimating lower bound of the number of common items needed

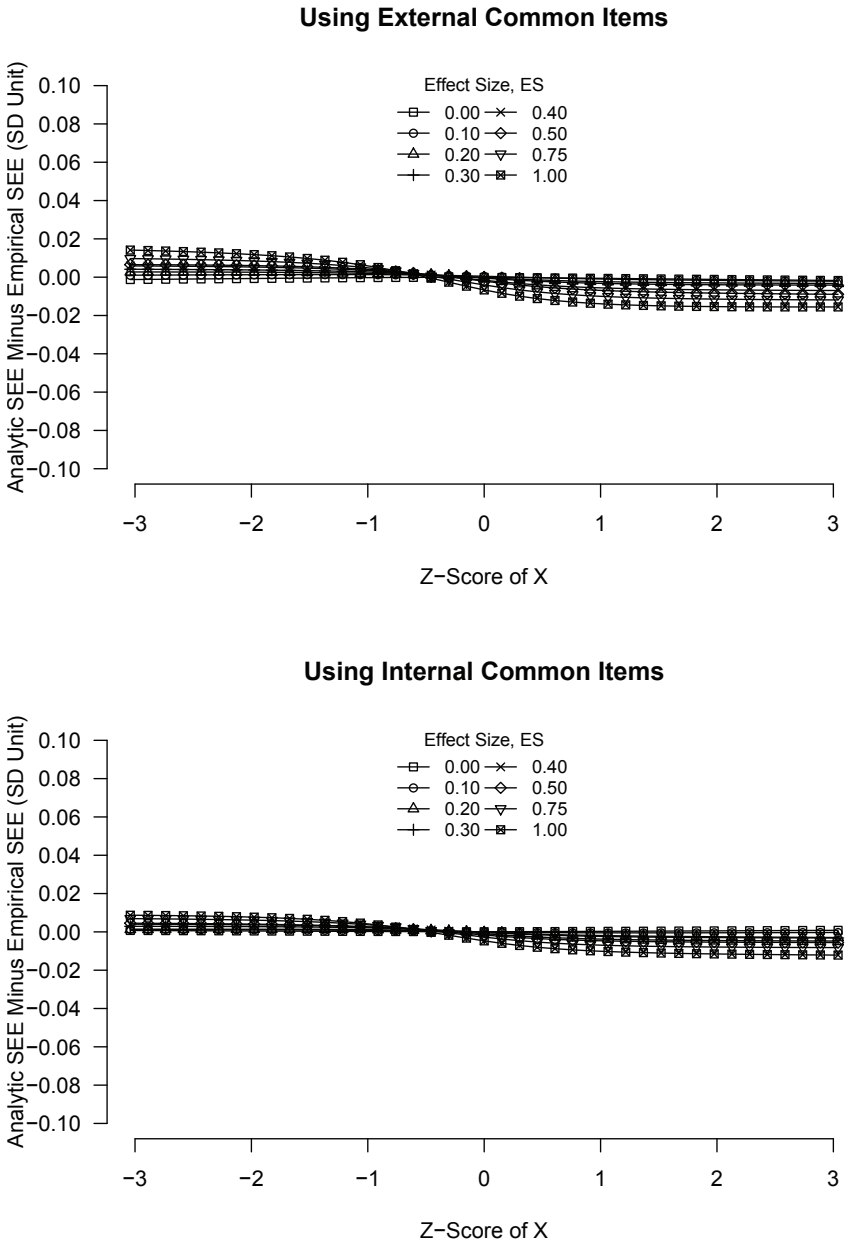


Figure 4: Difference between analytic SEE and empirical SEE (normal distribution, $\rho(X, X') = 0.8$, $N_{tot} = 2,000$)

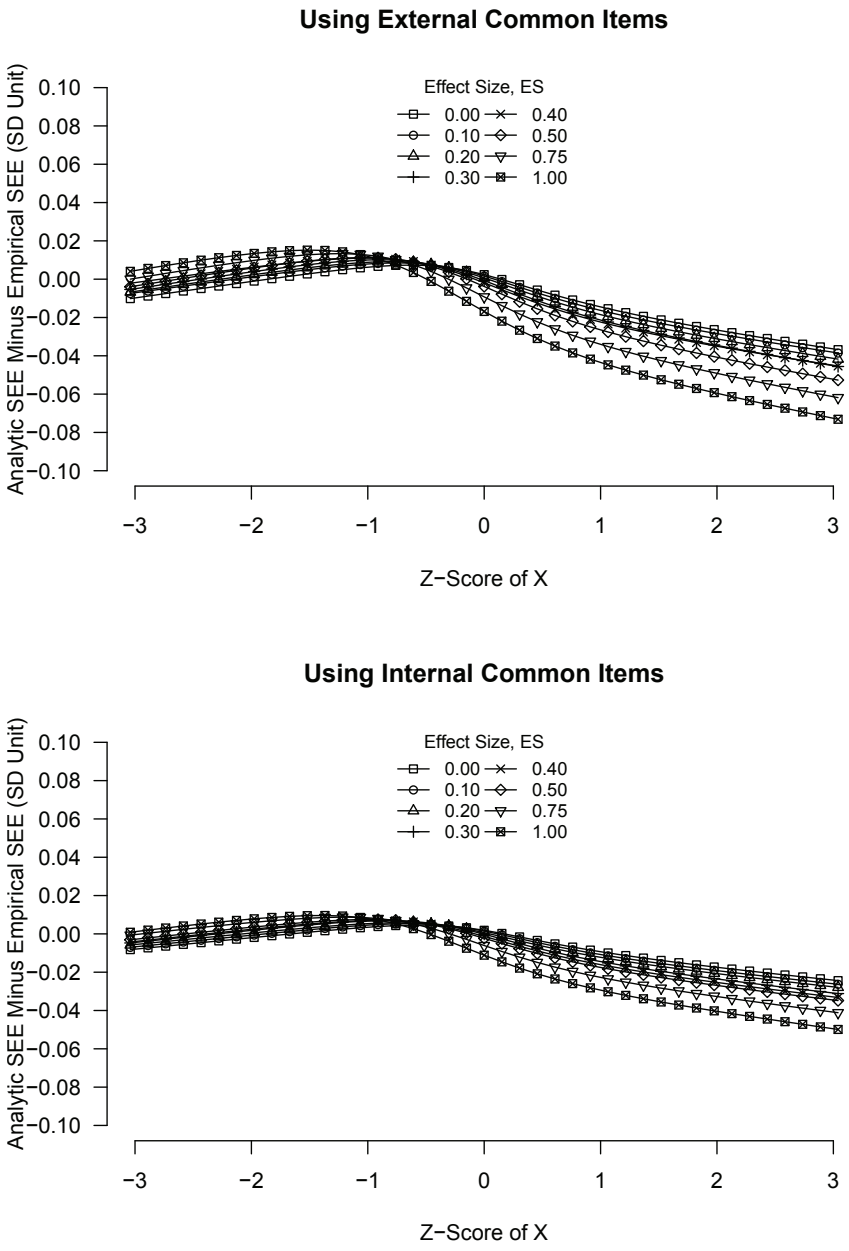


Figure 5: Difference between analytic SEE and empirical SEE (positively skewed distribution, $\rho(X, X') = 0.8$, $N_{tot} = 2,000$)

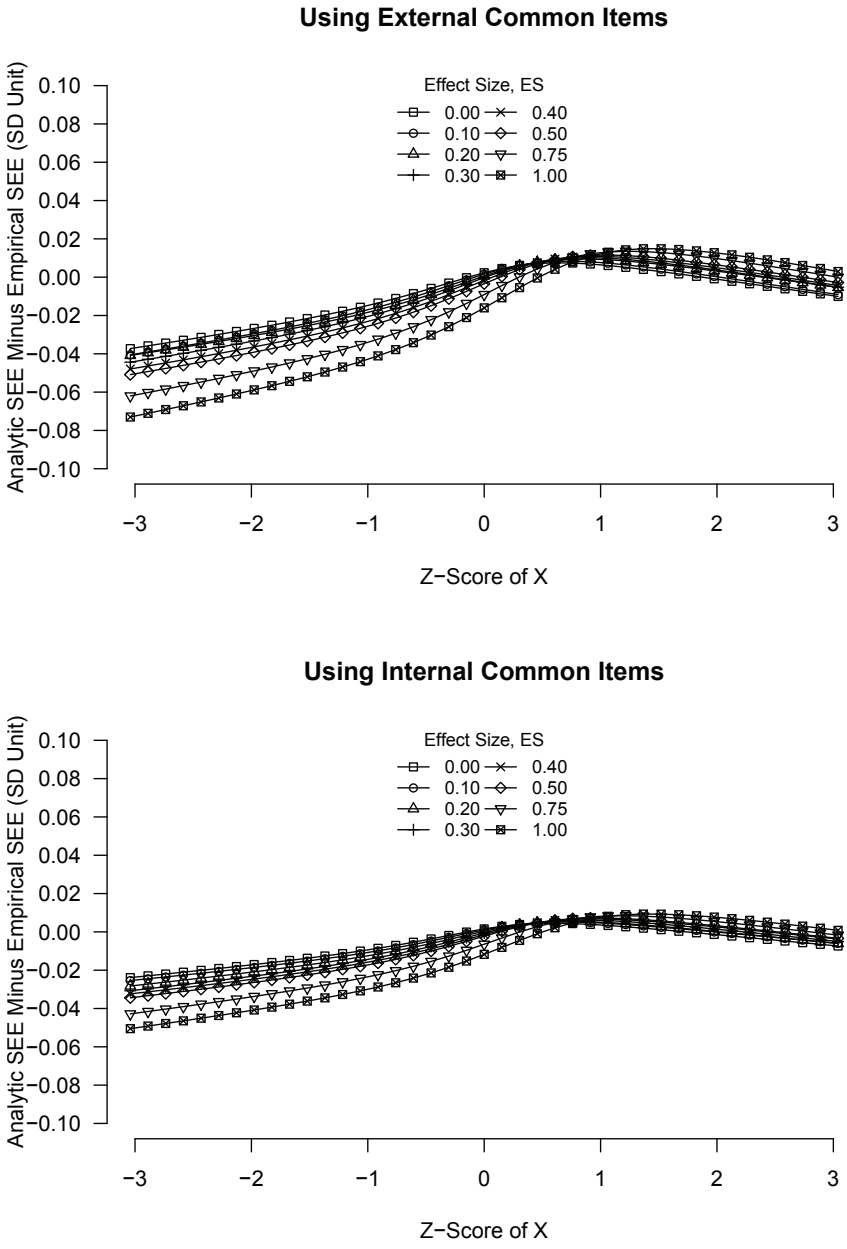


Figure 6: Difference between analytic SEE and empirical SEE (negatively skewed distribution, $\rho(X, X') = 0.8$, $N_{tot} = 2,000$)

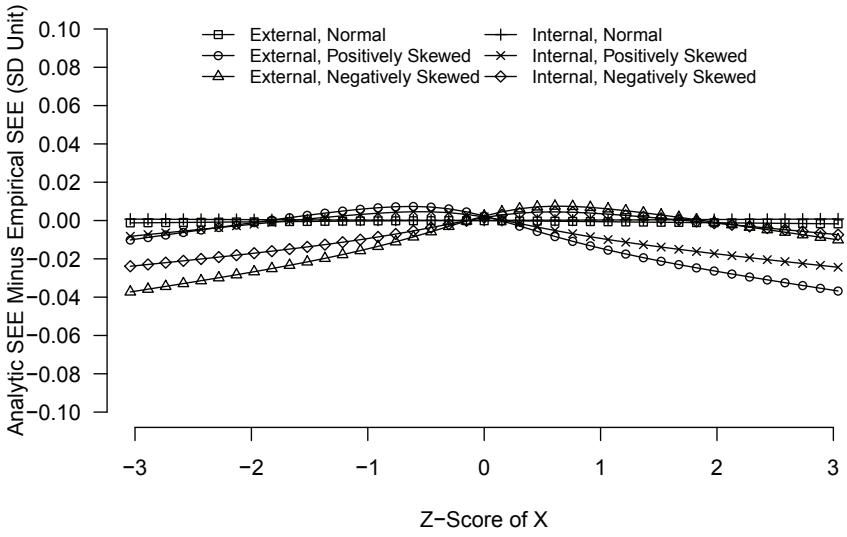


Figure 7: Difference between analytic SEE and empirical SEE when $ES = 0$ ($\rho(X, X') = 0.8$, $N_{tot} = 2,000$)

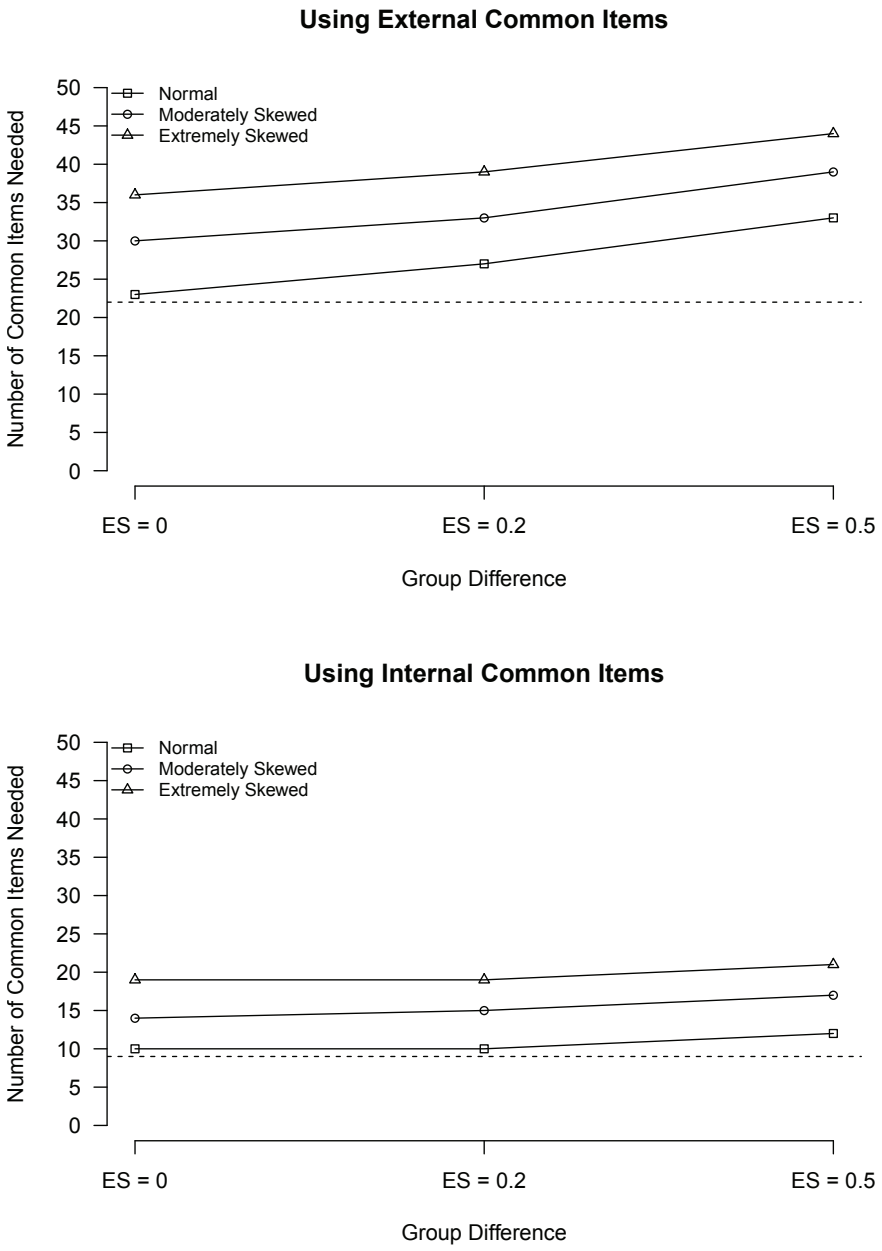


Figure 8: Modified numbers of common items needed (total number of items on either Form X or Y is 50, $\rho(X, X') = 0.8$, $N_{tot} = 2,000$, $u = 0.10$)