Center for Advanced Studies in Measurement and Assessment

CASMA Research Report

Number 30

First-order and Second-order Equity in Equating

Robert L. Brennan

August 6, 2010

This research was partially supported by a contract with the College Board. The author expresses his appreciation to Won-Chan Lee, Tianyou Wang, Michael Kane, and Neil Dorans for many helpful comments.

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

The notion of equity in equating was introduced by Lord (1980. pp. 195ff). The basic idea is that for any true score on Form Y, the distribution of observed scores on Form Y should be the same as the distribution of converted scores on Form X (see Kolen & Brennan, 2004. pp. 10–11). If this goal is achieved, Lord argued, then it should be a matter of indifference to each and every examinee which form of a test he or she took. Lord showed, however, that scores on fallible forms of a test cannot be equated under this strict definition of equity. The only exception is when the two forms are parallel in the most strict sense (i.e., indistinguishable forms), in which case equating is unnecessary. So, under Lord's definition of equity, it is sometimes stated that equating is either impossible or unnecessary!

Morris (1982) suggested considering a less strict definition of equity, which he called "weak" equity, or first-order equity (FOE), which focuses only on the mean of the distributions of observed scores on Form Y and the converted Form X scores. Today it is rather common to refer to the combination of first-order equity and second-order equity (SOE) as "weak" equity. SOE considers the variance of the distributions of observed scores on Form Y and the converted Form X scores. (We will be more specific about definitions after notation is introduced, below.)

This paper considers requirements for attaining, or nearly attaining, FOE and SOE for true-score and observed-score equating, primarily under certain basic assumptions in classical test theory. Linear equating is considered first, followed by curvilinear equating. Particular consideration is given to the role of reliability in attaining, or nearly attaining, FOE and SOE. As discussed more fully later, equal reliability for forms has long been regarded as a requirement for equating, but high reliability has an ambiguous status in the equating literature. An important purpose of this paper is to provide a firmer ground for considering these matters.

1 Introduction

The following notational conventions are adopted in this paper:

- X and Y are observed-score random variables for Form X and Form Y, respectively, with realizations x and y.
- T_X and T_Y are true-score random variables for Form X and Form Y, respectively, with realizations τ_x and τ_y.
- E_X and E_Y are error-score random variables for Form X and Form Y, respectively, with realizations e_x and e_y.
- $\boldsymbol{E}$ means expected value.
- $l_Y(x)$ is the linear equivalent of x that puts it on the scale of Y. This is the observed-score equating (OSE) relationship.

- $l_{T_Y}(\tau_x)$ is the linear equivalent of τ_x that puts it on the scale of T_Y. This is the true-score equating (TSE) relationship. The notion of equity does not apply to $l_{T_Y}(\tau_x)$ since τ_x and τ_y are infallible scores.
- $l_{T_Y}(x)$ is the linear equivalent of x that puts it on the scale of T_Y; i.e., x replaces τ_x in the TSE relation $l_{T_Y}(\tau_x)$. We will call this applied true-score equating (ATSE) to distinguish it from TSE.[1] There is no a priori theoretical justification for replacing τ_x with x although, as we shall see, doing so sometimes has favorable consequences in terms of FOE.

Sometimes, for OSE we will use $l_Y(x|\tau_y)$ rather than $l_Y(x)$ to focus attention on the subpopulation of examinees who have the same true score on Y. Similarly, sometimes for ATSE we will use $l_{T_Y}(x|\tau_y)$ rather than $l_{T_Y}(x)$. When we want to focus on equivalents for the random variable X, we will replace x with X. In later parts of this paper l will be replaced by q standing for quadratic equivalents, or eq standing for curvilinear equivalents, in general.

Feldt and Brennan (1989) and Haertel (2006) provide extensive discussions of classical test theory. Here, we merely summarize some of the more important results used in this paper. In terms of random variables, the classical test theory model is

$$X = T_X + E_X, \tag{1}$$

where it is assumed that $\boldsymbol{E}(E_X) = 0$, with the expectation is taken over the population of persons. This means that $\mu(X) = \mu(T_X)$, where $\mu(\cdot)$ signifies mean. Under classical test theory, the reliability of scores on Form X is $\rho^2_X = \sigma^2(T_X)/\sigma^2(X)$, where $\sigma^2(\cdot)$ signifies variance.

When we wish to focus on observed scores for examinees with a particular true score (which is often the case in this paper), we replace Equation 1 with

$$X = \tau_x + E_X, \tag{2}$$

where it is assumed that $\boldsymbol{E}(E_X) = 0$, with expectation is taken over examinees from the population who have the same true score τ_x. This means that $\boldsymbol{E}(X) = \tau_x$. Furthermore, $\sigma^2(X|\tau_x) = \sigma^2(E_X|\tau_x)$, which is conditional error variance. Similar results apply to random variables and realizations for Form Y.

Typical treatments of classical test theory use the total-score metric for observed variables (e.g., number of items correct), although sometimes the mean-score metric is used (e.g., proportion of items correct). Since the two metrics are linearly related, reliability is unchanged. True-score variance and error variance are different for the two metrics, however, which has consequences. For example, if test length is increased by a factor of n, then: (i) for the total-score metric, true-score variance increases by a factor of n^2, and error variance *increases* by a factor of n; while (ii) for the mean-score metric, true-score variance is unchanged, and error variance *decreases* by a factor of n. Most of the mathematical developments in this paper do not require specifying what the metric is.

[1] Some authors (e.g., Kolen & Brennan, 2004) designate this as $l_Y(x)$. Here, $l_{T_Y}(x)$ is used to distinguish ATSE from OSE.

However, verbal discussions of results, particularly for SOE, sometimes implicity assume the mean-score metric. The crux of the matter is that understanding some equity issues is much simpler under the mean-score metric in which error variance decreases as test length increases.

2 Linear Equating

In this section, it is assumed that both the TSE and OSE relationships are linear. For linear OSE,

$$\begin{aligned} l_Y(X) &= a + b\,(X) && (3)\\ &= \left[\mu(Y) - \frac{\sigma(Y)}{\sigma(X)}\,\mu(X)\right] + \frac{\sigma(Y)}{\sigma(X)}\,X, && (4)\end{aligned}$$

where a and b are the observed-score intercept and slope respectively.

For TSE, the linear equating relationship is

$$\begin{aligned} l_{T_Y}(T_X) &= a_T + b_T\,(T_X) && (5)\\ &= \left[\mu(T_Y) - \frac{\sigma(T_Y)}{\sigma(T_X)}\,\mu(T_X)\right] + \frac{\sigma(T_Y)}{\sigma(T_X)}\,T_X && (6)\\ &= \left[\mu(Y) - \frac{\sigma(T_Y)}{\sigma(T_X)}\,\mu(X)\right] + \frac{\sigma(T_Y)}{\sigma(T_X)}\,T_X, && (7)\end{aligned}$$

where a_T and b_T are the true-score intercept and slope, respectively. Equation 7 follows from Equation 6 since, under classical test theory assumptions, the mean of observed scores equals the mean of true scores. The left side of Equation 5 could also be designated T_Y.

Since true scores are unknown, TSE cannot be used directly, and equity is an irrelevant consideration. The usual way to circumvent this problem is to replace T_X with X in Equation 7 which gives:

$$\begin{aligned} l_{T_Y}(X) &= a_T + b_T\,(X) && (8)\\ &= \left[\mu(Y) - \frac{\sigma(T_Y)}{\sigma(T_X)}\,\mu(X)\right] + \frac{\sigma(T_Y)}{\sigma(T_X)}\,X. && (9)\end{aligned}$$

This is the equation for linear ATSE.

Equations 3–9 are expressed in terms of the variables X or T_X. When we wish to focus on specific observed-score or true-score equivalents, then X and T_X are replaced with x and τ_x, respectively. This does not apply to $\mu(\cdot)$ and $\sigma(\cdot)$, which are parameters for some target population. (It follows that a, b, a_T, and b_T are also parameters.) This paper does not consider issues associated with estimating these parameters.

These linear equating equations are not design specific; i.e., they apply whether the design is the random groups design, the single group design, or the common-item nonequivalent groups (CINEG) design, which is sometimes called the nonequivalent anchor test (NEAT) design. For the CINEG design,

the subscript s (which designates synthetic group) could be added to each of the mean and standard deviation parameters, but for the purposes of this paper, doing so is not necessary.

Under true-score equating (Equations 5–7) there is an increasing monotonic relationship between T_X and T_Y. Among other things, this means that for every τ_x there is a single τ_y, and for every τ_y there is a single τ_x. In other words, for the subpopulation of examinees with the same τ_y, every member of that subpopulation has the same τ_x. Consequently, the distribution of $X|\tau_y$ is equivalent to the distribution of X for some particular τ_x. To simplify notation we will often use $\boldsymbol{E}(X|\tau_y) = \boldsymbol{E}(X|\tau_x)$ and $\sigma^2(X|\tau_y) = \sigma^2(X|\tau_x)$, where τ_x is to be understood as the value of T_X that is associated with τ_y. We make use of these conventions (for both linear and curvilinear equating) throughout this paper.

Strictly speaking, OSE does not recognize the existence of true scores. In order to study equity, however, we must assume a true-score model. Throughout most of this paper, we adopt the classical test theory model.

2.1 First-order Equity

By definition, first order equity holds for ATSE if

$$\boldsymbol{E}[l_{T_Y}(X|\tau_y)] = \boldsymbol{E}(Y|\tau_y) = \tau_y \qquad \text{for all } \tau_y, \tag{10}$$

and first-order equity holds for OSE if

$$\boldsymbol{E}[l_Y(X|\tau_y)] = \boldsymbol{E}(Y|\tau_y) = \tau_y \qquad \text{for all } \tau_y. \tag{11}$$

2.1.1 Applied True-score Equating

To examine first-order equity for linear ATSE, we substitute Equation 8 in the left side of Equation 10, which gives

$$\begin{aligned} \boldsymbol{E}[l_{T_Y}(X|\tau_y)] &= \boldsymbol{E}[a_T + b_T(X|\tau_y)] \\ &= a_T + b_T\boldsymbol{E}(X|\tau_y) \qquad (12) \\ &= a_T + b_T(\tau_x) \qquad (13) \\ &= \tau_y. \qquad (14) \end{aligned}$$

Equation 13 follows from Equation 12 by a two step argument. First, as noted above, for each τ_y there is a unique τ_x, which means that $\boldsymbol{E}(X|\tau_y) = \boldsymbol{E}(X|\tau_x)$. Second, by the assumptions of classical test theory, $\boldsymbol{E}(X|\tau_x) = \tau_x$.

So, in general, FOE holds for linear ATSE. Hanson (1991) proved this for the Levine ATSE method under the CINEG design. In a sense, the proof provided in this section is more general in that it is not design specific.

Recall that there is no a priori theoretical justification for replacing τ_x with x, as is done in ATSE. It is evident, however, that doing so in the context of linear equating leads to FOE, which provides at least a partial justification for what is otherwise an ad hoc procedure.

2.1.2 Observed-score Equating

To examine FOE for linear OSE, we substitute Equation 3 in the left side of Equation 11, which gives

$$\begin{aligned} \boldsymbol{E}[l_Y(X|\tau_y)] &= \boldsymbol{E}[a + b(X|\tau_y)] \\ &= a + b\boldsymbol{E}(X|\tau_y) \quad (15) \\ &= a + b\,(\tau_x). \quad (16) \end{aligned}$$

Equation 16 follows from Equation 15 by the same argument discussed with respect to Equations 13 and 14. Since $a \neq a_T$ and $b \neq b_T$, clearly, FOE does not hold for linear OSE. As shown next, however, a less stringent version of FOE does hold.

For OSE the slope is

$$\begin{aligned} b &= \frac{\sigma(Y)}{\sigma(X)} \quad (17) \\ &= \frac{\sigma(T_Y)\,\rho_X}{\sigma(T_X)\,\rho_Y}, \quad (18) \end{aligned}$$

where ρ_X and ρ_Y are the square roots of the reliabilities for X and Y, respectively. If we assume that $\rho_X^2 = \rho_Y^2$, then $b = b_T$, and from Equations 4 and 7, $a = a_T$. It follows that Equation 16 becomes

$$\boldsymbol{E}[l_Y(X|\tau_y)] = a_T + b_T\,(\tau_x) = \tau_y. \quad (19)$$

That is, FOE is satisfied for linear OSE if $\rho_X^2 = \rho_Y^2$. Occasionally, we will abbreviate this condition as ER (equal reliabilities).

2.2 Second-order Equity

By definition, SOE holds for ATSE if

$$\sigma^2[l_{T_Y}(X|\tau_y)] = \sigma^2(Y|\tau_y) = \sigma^2(E_Y|\tau_y) \qquad \text{for all } \tau_y, \quad (20)$$

and SOE holds for OSE if

$$\sigma^2[l_Y(X|\tau_y)] = \sigma^2(Y|\tau_y) = \sigma^2(E_Y|\tau_y) \qquad \text{for all } \tau_y, \quad (21)$$

where $\sigma^2(E_Y|\tau_y)$ is conditional error variance.

2.2.1 Applied True-score Equating

For linear ATSE,

$$\begin{aligned} \sigma^2[l_{T_Y}(X|\tau_y)] &= \sigma^2[a_T + b_T(X|\tau_y)] \\ &= b_T^2\,\sigma^2(X|\tau_y) \end{aligned}$$

$$
\begin{aligned}
&= \frac{\sigma^2(T_Y)}{\sigma^2(T_X)}\,\sigma^2(X|\tau_x). \\
&= \frac{\sigma^2(T_Y)}{\sigma^2(T_X)}\,\sigma^2(E_X|\tau_x). \qquad (22)
\end{aligned}
$$

Equation 22 can be rewritten as

$$\sigma^2[l_{T_Y}(X|\tau_y)] = \left[\frac{\sigma^2(T_Y)/\sigma^2(E_Y|\tau_y)}{\sigma^2(T_X)/\sigma^2(E_X|\tau_x)}\right]\,\sigma^2(E_Y|\tau_y). \qquad (23)$$

This means that SOE holds if the term in square brackets is 1, which occurs when

$$\frac{\sigma^2(E_Y|\tau_y)}{\sigma^2(E_X|\tau_x)} = \frac{\sigma^2(T_Y)}{\sigma^2(T_X)} \qquad \text{for all } \tau_y. \qquad (24)$$

That is, SOE is satisfied for linear ATSE when, for every τ_y, the ratio of the conditional error variances (CEV) is a constant equal to the ratio of true-score (T) variances. We will call this condition CEVT. It is easy to prove that if Equation 24 holds, then $\rho_X^2 = \rho_Y^2$, but the converse is *not* true; i.e., $\rho_X^2 = \rho_Y^2$ does not guarantee that SOE holds for ATSE.

Suppose conditional error variances for X and Y are both homoscedastic in the sense that

$$\sigma^2(E_Y|\tau_y) = \sigma^2(E_Y) \qquad \text{for all } \tau_y, \qquad (25)$$

and

$$\sigma^2(E_X|\tau_x) = \sigma^2(E_X) \qquad \text{for all } \tau_x. \qquad (26)$$

In this case, SOE is satisfied if

$$\sigma^2[l_{T_Y}(X|\tau_y)] = \sigma^2(E_Y) \qquad \text{for all } \tau_y. \qquad (27)$$

Assuming homoscedasticity, Equation 23 becomes

$$
\begin{aligned}
\sigma^2[l_{T_Y}(X|\tau_y)] &= \frac{\sigma^2(T_Y)/\sigma^2(E_Y)}{\sigma^2(T_X)/\sigma^2(E_X)}\,\sigma^2(E_Y) \\
&= \frac{(S/N)_Y}{(S/N)_X}\,\sigma^2(E_Y),
\end{aligned}
$$

where $(S/N)_X$ and $(S/N)_Y$ are the signal-noise ratios for X and Y, respectively. A signal-noise ratio is functionally related to reliability by the formula $(S/N) = \rho^2/(1-\rho^2)$. It follows that

$$\sigma^2[l_{T_Y}(X|\tau_y)] = \frac{\rho_Y^2(1-\rho_X^2)}{\rho_X^2(1-\rho_Y^2)}\,\sigma^2(E_Y). \qquad (28)$$

If $\rho_X^2 = \rho_Y^2$, then SOE is satisfied. In short, for linear ATSE, SOE is satisfied under the conditions of homoscedastic error variances (HEV) and equal reliabilities (ER).

2.2.2 Observed-score Equating

For linear OSE,

$$\begin{aligned}
\sigma^2[l_Y(X|\tau_y)] &= \sigma^2[a + b(X|\tau_y)] \\
&= b^2\,\sigma^2(X|\tau_y) \\
&= \frac{\sigma^2(Y)}{\sigma^2(X)}\,\sigma^2(X|\tau_x) \\
&= \frac{\sigma^2(Y)}{\sigma^2(X)}\,\sigma^2(E_X|\tau_x). \\
&= \left[\frac{\sigma^2(Y)/\sigma^2(E_Y|\tau_y)}{\sigma^2(X)/\sigma^2(E_X|\tau_x)}\right]\,\sigma^2(E_Y|\tau_y). \qquad (29)
\end{aligned}$$

This means that SOE holds if the term is square brackets is 1, which occurs when

$$\frac{\sigma^2(E_Y|\tau_y)}{\sigma^2(E_X|\tau_x)} = \frac{\sigma^2(Y)}{\sigma^2(X)} \qquad \text{for all } \tau_y. \qquad (30)$$

That is, SOE is satisfied for linear OSE when, for every τ_y, the ratio of the conditional error variances (CEV) is a constant equal to the ratio of observed-score (O) variances. We will refer to this condition as CEVO. It is easy to prove that if Equation 30 holds, then $\rho^2_X = \rho^2_Y$, but the converse is *not* true; i.e., $\rho^2_X = \rho^2_Y$ does not guarantee that SOE holds for OSE.

Suppose conditional error variances for X and Y are both homoscedastic. In this case, SOE is satisfied if

$$\sigma^2[l_Y(X|\tau_y)] = \sigma^2(E_Y) \qquad \text{for all } \tau_y. \qquad (31)$$

Assuming homoscedastic error variances, Equation 29 becomes

$$\begin{aligned}
\sigma^2[l_Y(X|\tau_y)] &= \frac{\rho^2_X\,\sigma^2(T_Y)/\sigma^2(E_Y)}{\rho^2_Y\,\sigma^2(T_X)/\sigma^2(E_X)}\,\sigma^2(E_Y) \\
&= \frac{\rho^2_X\,(S/N)_Y}{\rho^2_Y\,(S/N)_X}\,\sigma^2(E_Y) \\
&= \frac{\rho^2_X\,\rho^2_Y(1-\rho^2_X)}{\rho^2_Y\,\rho^2_X(1-\rho^2_Y)}\,\sigma^2(E_Y) \\
&= \frac{(1-\rho^2_X)}{(1-\rho^2_Y)}\,\sigma^2(E_Y). \qquad (32)
\end{aligned}$$

If $\rho^2_X = \rho^2_Y$, then SOE is satisfied. In short, for linear OSE, SOE is satisfied under the conditions of HEV+ER. Note that these two conditions lead to SOE for *both* linear OSE and linear ATSE (see Section 2.2.1).

3 Curvilinear Equating

Let us consider a simple case of curvilinear equating in which the curvilinearity is completely explained by a quadratic term for both TSE and OSE. Specifically,

let quadratic OSE be defined as

$$q_Y(X) \quad = \quad a + b\,(X) + c\,(X^2), \tag{33}$$

where a and b are the observed-score intercept and slope, respectively, for linear OSE, as defined in Equation 4, and c is a constant.

Similarly, let quadratic TSE be defined as

$$q_{T_Y}(T_X) \quad = \quad a_T + b_T\,(T_X) + c_T\,(T_X^2), \tag{34}$$

where a_T and b_T are the true-score intercept and slope, respectively, for linear TSE, as defined in Equation 7, and c_T is a constant. The left side of Equation 34 is essentially T_Y, provided, of course, that the relationship between T_X and T_Y is truly quadratic. Quadratic ATSE is

$$q_{T_Y}(X) \quad = \quad a_T + b_T\,(X) + c_T\,(X^2). \tag{35}$$

In defining $q_Y(X)$, $q_{T_Y}(T_X)$, and $q_{T_Y}(X)$ in this manner, the last term in each equation is essentially a quadratic deviation from linearity. We assume, as well, that all equating functions are monotonic.

3.1 First-order Equity for Quadratic Equating

FOE is satisfied for ATSE if, for every τ_y, the expected value of Equation 35 equals τ_y. Now,

$$\begin{aligned} \boldsymbol{E}[q_{T_Y}(X|\tau_y)] &= \boldsymbol{E}[a_T + b_T(X|\tau_y) + c_T(X^2|\tau_y)] \\ &= a_T + b_T\boldsymbol{E}(X|\tau_y) + c_T\boldsymbol{E}(X^2|\tau_y) \\ &= a_T + b_T\,(\tau_x) + c_T\boldsymbol{E}(X^2|\tau_x), \end{aligned} \tag{36}$$

which equals τ_y (see Equation 34) only if $c_T = 0$ (in which case the relationship is linear) or $\boldsymbol{E}(X^2|\tau_x) = \tau_x^2$. Note that

$$\sigma^2(X|\tau_x) = \boldsymbol{E}(X^2|\tau_x) - [\boldsymbol{E}(X|\tau_x)]^2 = \boldsymbol{E}(X^2|\tau_x) - \tau_x^2.$$

It follows that the third term in Equation 36 is

$$c_T\boldsymbol{E}(X^2|\tau_x) = c_T\,\{\sigma^2(X|\tau_x) + \tau_x^2\}, \tag{37}$$

which equals $c_T\,\tau_x^2$ only if $\sigma^2(X|\tau_x) = 0$.

In other words, FOE is satisfied only if (i) the true equating relationship is linear or (ii) the conditional error variance is 0 for each τ_x (associated with a particular τ_y). Strictly speaking, then, FOE is generally not satisfied for quadratic ATSE.

FOE is more nearly satisfied, however, if overall error variance for X is relatively small or, equivalently, reliability for X is relatively large.[2] We call this

[2]Since true score variances are constant, a decrease in overall error variance is necessarily associated with an increase in reliability.

HRX (high reliability for X). Note that error variance and reliability for Y play no role; hence, strictly speaking there is no requirement for equal reliabilities for both X and Y. Of course, since symmetry is a generally accepted criterion for equating, FOE for putting Y on the scale of X would bring error variance and reliability for Y into play.

When we say that "FOE is more nearly satisfied," we do not mean that, for each and every τ_y, it is necessarily true that Equation 36 get progressively closer to $\boldsymbol{E}(Y|\tau_y)$ as reliability increases. Rather, that phrase "FOE is more nearly satisfied" means that this "typically" happens. The interpretation of the word "typically" should become clearer when we consider an example later.

For quadratic OSE, FOE is satisfied if, for every τ_y, the expected value of Equation 33 equals τ_y. A development similar to that for quadratic ATSE (see, also Section 2.1.2) leads to the conclusion that FOE is not generally satisfied for quadratic OSE. All other things being equal, however, FOE is more nearly satisfied under the conditions of equal reliabilities (ER) for X and Y and relatively high reliabilities (HR). We this joint condition as ER+HR.

Note, in particular, that for both quadratic ATSE and quadratic OSE, equal reliabilities and/or homogeneous error variances are not sufficient to satisfy FOE, because these conditions do not guarantee that $\boldsymbol{E}(X^2|\tau_x) = \tau_x^2$.

3.2 Second-order Equity for Quadratic Equating

SOE is satisfied for quadratic ATSE if, for every τ_y, the variance of Equation 35 equals $\sigma^2(Y|\tau_y)$, which is $\sigma^2(E_Y|\tau_y)$. For quadratic ATSE,

$$\begin{aligned}
\sigma^2[q_{T_Y}(X|\tau_y)] &= \sigma^2[a_T + b_T(X|\tau_y) + c_T(X^2|\tau_y)] \\
&= b_T^2\,\sigma^2(X|\tau_y) + c_T^2\,\sigma^2(X^2|\tau_y) \\
&\quad + 2\,b_T\,c_T\,\sigma(X|\tau_y, X^2|\tau_y) \\
&= b_T^2\,\sigma^2(X|\tau_x) + c_T^2\,\sigma^2(X^2|\tau_x) \\
&\quad + 2\,b_T\,c_T\,\sigma(X|\tau_x, X^2|\tau_x). \qquad (38)
\end{aligned}$$

The first two terms in Equation 38 are positive. It can be shown that the third term is

$$2\,b_T\,c_T\{\boldsymbol{E}(X-\tau_x)^3 + 2\,\tau_x\,\sigma^2(X|\tau_x)\}, \qquad (39)$$

which may be positive or negative.

Given Equation 38 and the developments in Section 2.2.1, SOE holds if:

(a) (i) CEVT holds (for all τ_y, the ratio of conditional error variances is a constant equal to the ratio of true-score variances) or (ii) the conditions of ER+HEV hold—either (i) or (ii) guarantees that the first term of Equation 38 is $\sigma^2(Y|\tau_y)$ [or, equivalently, $\sigma^2(E_Y|\tau_y)$]; *and*

(b) the sum of the second and third terms is zero, which happens, for example, if $c_T = 0$, in which case the equating relationship is linear.

Clearly SOE does not generally hold for quadratic ATSE. All other things being equal, however, as overall error variance for X decreases or, equivalently, reliability for X increases, SOE is likely to be more nearly satisfied. This is evident from the fact that the last two terms in Equation 38 are functions of $(X - \tau_x)$, and $(X - \tau_x) \rightarrow 0$ as error variance decreases. In short, SOE is more nearly satisfied for quadratic ATSE under the conditions of CEVT+HR or HEV+ER+HR.

SOE is satisfied for quadratic OSE if, for every τ_y, the variance of Equation 33 equals $\sigma^2(Y|\tau_y) = \sigma^2(E_Y|\tau_y)$. For quadratic OSE,

$$\sigma^2[q_{T_Y}(X|\tau_y)] \quad = \quad \sigma^2[a + b(X|\tau_y) + c(X^2|\tau_y)],$$

which has the same form as Equation **??** with a_T, b_T, and c_T replaced by a, b, and c, respectively. Using the results in this section and those in Section 2.2.2, it follows that SOE is not generally satisfied by quadratic OSE, but SOE is more nearly satisfied under the conditions of CEVO+HR or HEV+ER+HR. (Recall that RCEVO holds if, for all τ_y, the ratio of conditional error variances is a constant equal to the ratio of observed-score variances.)

3.3 Curvilinear Equating, in General

Strictly speaking, the curvilinear results provided in Sections 3.1 and 3.2 are for a quadratic polynomial, only. It seems clear, however, that extending these results to higher-order polynomials will lead to the conclusions that FOE and SOE do not generally hold, but under certain conditions they may be more nearly satisfied. In particular, high reliability tends to increase the likelihood that FOE and SOE will be more nearly satisfied.

Item response theory (IRT) is clearly a very prevalent example of curvilinear equating. There are two versions of IRT equating, namely IRT ATSE and IRT OSE.

IRT ATSE is based on using test characteristic curves (TCCs) for Forms X and Y, where the TCCs are functions of θ. There are two principal hurdles, therefore, to asserting that the results provided here apply to IRT ATSE: (i) TCCs are not polynomial functions, and (ii) θ is not the classical test theory true score, τ. Note, however, that although TCCs are not defined as polynomials, they could be approximated by polynomials to any desired degree of accuracy. Furthermore, although $\theta \neq \tau$, most investigators seem willing to treat the IRT-based expected number-correct score (or expected proportion-correct score) as τ. In effect, such investigators assert that τ is a function of θ. It follows that there is a unique (τ_y, τ_x) for each (θ_y, θ_x). Under these circumstances, the results provided in Sections 3.1 and 3.2 apply to IRT ATSE.

IRT OSE is an equipercentile equating based on expected observed score distributions, both of which are "built" from the conditional distributions of observed scores given θ. Provided τ is viewed as a function of θ, the results provided in Sections 3.1 and 3.2 apply to IRT OSE, since the equipercentile function can be approximated by a polynomial to any desired degree of accuracy.

3.4 Example

As an illustration, this section considers FOE and SOE for equating with the two-parameter beta-binomial model. This example has the distinct advantages of being analytic and relatively simple, while at the same time being flexible enough to illustrate important results about FOE and SOE with relatively few assumptions. In particular, using the beta-binomial model, we can isolate the contribution of reliability to the approximate achievement of FOE and SOE, without any confounding influences.

Specifically, it is assumed here that for both Forms X and Y, true scores have a beta distribution, errors of measurement conditional on true score are binomially distributed, and marginal observed scores for X and Y have negative hypergeometric distributions (see Lord & Novick, 1968, pp. 515–524). FOE and SOE are examined here when both forms have three different lengths ($n = 20$, $n = 40$, and $n = 60$) and, hence, three different reliabilities. We begin with notational conventions and formulas for FOE and SOE under the beta-binomial model.

3.4.1 Notational Conventions

The following notational conventions are employed:

- $\beta(u_X, v_X)$ is the continuous two-parameter beta probability density function (pdf) for Form X with $0 < \tau_x < 1$;
- $I_{T_X}(u_X, v_X)$ is the continuous incomplete beta distribution, or the beta cumulative distribution function (cdf), for Form X with $0 < \tau_x < 1$;
- $\mathrm{B}(n, x|\tau_x)$ is the discrete binomial pdf conditional on τ_x for $x = 0(1)n$ (i.e., x ranging from 0 to n in increments of 1); and
- $\mathrm{NH}(u_X, v_X, n)$ is the discrete negative hypergeometric (or beta-binomial) pdf for $x = 0(1)n$.

By convention, $\beta(u_X, v_X)$ uses true-score in the proportion-correct metric, while $\mathrm{B}(n, x|\tau_x)$ and $\mathrm{NH}(u_X, v_X, n)$ use observed-score in the number-correct metric. For quantifying FOE and SOE, we will transform number-correct scores x to proportion-correct scores $\overline{x} = x/n$ and denote the random variable as $\overline{X}$. With obvious changes in notation, these same conventions apply to Y.

When both T_X and T_Y have beta distributions, the T_Y-equivalent of any τ_x score is given by

$$e_{T_Y} = G^{-1}[F(\tau_x)], \tag{40}$$

where $F(\tau_x) = I_{\tau_x}(u_X, v_X)$ and $G(\tau_y) = I_{\tau_y}(u_Y, v_Y)$. Equation 40 is the TSE relationship in the proportion-correct metric. For an n item test, the ATSE equivalents in the proportion-correct metric are obtained by using Equation 40 with the $n + 1$ values of τ_x that equal the $\overline{x}$ values $0(x/n)1$.

3.4.2 Computational Formulas for FOE and SOE

For ATSE, FOE is satisfied for τ_y if $\boldsymbol{E}(\overline{Y}|\tau_y)$ equals

$$\boldsymbol{E}[e_{T_Y}(\overline{X})|\tau_y] = \sum_{x=0}^{n} e_{T_Y}(\overline{x})\,\mathrm{B}(n,x|\tau_x), \tag{41}$$

where $\tau_x = F^{-1}[G(\tau_y)]$ (i.e., the inverse of Equation 40). The right side of Equation 41 provides a computational formula. Since $\boldsymbol{E}(\overline{Y}|\tau_y) = \tau_y$, deviations from FOE (i.e., FOE bias) are quantified here as

$$\mathrm{DFOE}(\tau_y) = \tau_y - \boldsymbol{E}[e_{T_Y}(\overline{X})|\tau_y]. \tag{42}$$

$|\mathrm{DFOE}(\tau_y)|$ will be used to focus on the absolute magnitude of the bias.

For ATSE, SOE is satisfied for τ_y if $\sigma^2(\overline{Y}|\tau_y)$ equals

$$\sigma^2[e_{T_Y}(\overline{X})|\tau_y] = \boldsymbol{E}[e_{T_Y}(\overline{X})|\tau_y]^2 - \{\boldsymbol{E}[e_{T_Y}(\overline{X})|\tau_y]\}^2 \tag{43}$$

$$= \sum_{x=0}^{n}[e_{T_Y}(\overline{x})]^2\,\mathrm{B}(n,x|\tau_x) - \left\{\sum_{x=0}^{n} e_{T_Y}(\overline{x})\,\mathrm{B}(n,x|\tau_x)\right\}^2, \tag{44}$$

where Equation 44 provides a computational formula. For a particular τ_y, deviations from SOE (i.e., SOE bias) are quantified here as

$$\mathrm{DSOE}(\tau_y) = \sigma^2(\overline{Y}|\tau_y) - \sigma^2[e_{T_Y}(\overline{X})|\tau_y]. \tag{45}$$

$|\mathrm{DSOE}(\tau_y)|$ will be used to focus on the absolute magnitude of the bias.

Using the mean-score metric, the negative hypergeomentric equating relationship puts $\overline{x}$ on the scale of $\overline{Y}$, rather than the scale of T_Y. Therefore, for negative hypergeometric OSE, the equations for $\mathrm{DSOE}(\tau_y)$ and $\mathrm{DSOE}(\tau_y)$ are obtained by replacing $e_{T_Y}(\overline{x})$ and $e_{T_Y}(\overline{X})$ in Equations 41–45 with $e_{\overline{Y}}(\overline{x})$ and $e_{\overline{Y}}(\overline{X})$, respectively.

3.4.3 Results

Suppose $u_X = 4$, $v_X = 3$, $u_Y = 3$, and $v_Y = 3$. Then, the beta pdf for X, $\beta(4,3)$, is given by the top left sub-figure of Figure 1, and the beta pdf for Y, $\beta(3,3)$, is given by the middle left sub-figure of Figure 1. (All figures and tables for this example are in the Appendix.) The T_Y equivalents of τ_x are given by the difference plot in the lower left sub-figure of Figure 1. Specifically, for any given τ_x, $e_{T_Y}(\tau_x)$ is obtained by adding τ_x and the corresponding value on the vertical axis. This type of difference plot has the advantage of visually revealing departures from linearity, even when a direct plot of $e_{T_Y}(\tau_x)$ looks nearly linear.

The three sub-figures on the right of Figure 1 provide the negative hypergeometric distributions for $\overline{X}$ and $\overline{Y}$ with $n = 20$; i.e., $\mathrm{NH}(4,3,20)$ and $\mathrm{NH}(3,3,20)$, respectively. The bottom right sub-figure provides the negative-hypergeometric mean-score equivalents, $e_{\overline{Y}}(\overline{x})$, in terms of a difference plot. Figure 2 provides corresponding figures for $n = 40$ and $n = 60$.

Figure 3 provides plots of DFOE(τ_y) and DSOE(τ_y) for both ATSE and OSE. Table 2 provides weighted (by the density of T_Y) and unweighted average values of |DFOE(τ_y)| and |DSOE(τ_y)|. It is evident from Figure 3 and Table 2 that:

- |DFOE(τ_y)| tends to be considerable smaller under ATSE than OSE;
- |DSOE(τ_y)| tends to be considerable larger under ATSE than OSE;
- on average (see Table 2) FOE and SOE are more nearly satisfied as the number of items increases; and
- for nearly every τ_y, under both ATSE and OSE, FOE and SOE are more nearly satisfied as the number of items increases.

The last two points are consistent with the observation that FOE and SOE are more nearly satisfied as reliability increases.

4 Discussion

The principal results derived in this paper for *linear* equating are:

1. FOE is satisfied for ATSE;
2. FOE is satisfied for OSE under the condition of ER (equal reliabilities);
3. SOE is satisfied for ATSE if
   (a) CEVT holds (for all τ_y, the ratio of conditional error variances is a constant equal to the ratio of true-score variances), or
   (b) the conditions of HEV+ER hold (homogeneous error variances and equal reliabilities); and
4. SOE is satisfied for OSE if
   (a) CEVO holds (for all τ_y, the ratio of conditional error variances is a constant equal to the ratio of observed-score variances), or
   (b) the conditions of HEV+ER hold.

Given these results, it seems reasonable to assert that most linear equating procedures are not likely to meet the strict criteria of FOE and SOE, although FOE and SOE may be satisfied if certain stringent conditions hold. The assumption of homogeneous error variances seems particularly unlikely to be met in most circumstances, although an arcsine transformation may facilitate achieving it (see Kolen & Brennan, 2004, pp. 348ff).

One notable fact about the principal linear results is that none of them directly depend upon the magnitude of reliability. By contrast, for curvilinear equating, FOE and SOE are not generally satisfied, but the magnitude of reliability matters in the sense that, all other things being equal, FOE and SOE are more nearly satisfied when reliabilities are high. Specifically, for curvilinear equating:

1. FOE for ATSE is more nearly satisfied under HRX (high reliability for X);
2. FOE for OSE is more nearly satisfied under ER+HR (equal and high reliabilities for X and Y);
3. SOE for ATSE is more nearly satisfied under CEVT+HR or HEV+ER+HR; and
4. SOE for OSE is more nearly satisfied under CEVO+HR or HEV+ER+HR.

These results for curvilinear equating are generally consistent with statements made by Morris (1982, p. 177), although his development differs considerable from the approach taken in this paper.

The fact that FOE and SOE (or one of their weaker versions) are more likely to be satisfied under linear equating than curvilinear equating does not mean that linear equating is preferable. These properties of linear equating are realized only when the true equating function is linear, which is seldom the case. Indeed, if satisfying FOE and SOE were taken as the sole criterion for choosing an equating method, then the "best" equating would be no equating!

4.1 Reliability and Equity

This paper suggests the following conclusions about the role of reliability in achieving, or facilitating the approximate achievement of, FOE and SOE:

- reliability has no bearing on FOE for linear ATSE;
- ER guarantees that FOE is satisfied for linear (but not curvilinear) OSE;
- ER alone does not guarantee that SOE holds for linear ATSE or OSE, or for curvilinear ATSE or OSE;
- almost always ER facilitates achieving approximate FOE and SOE; and
- HR facilitates achieving approximate FOE and SOE when equating is curvilinear.

By definition, FOE and SOE are conditional on specific values of τ_y, whereas reliability is defined over a population. In this sense, the above summary of results cannot possibly capture all of the relevant issues surrounding FOE and SOE. Still, this summary seems helpful. Note, also, that true-score variances are parameters in equating contexts, which means that any changes in reliability are attributable to changes in error variance. In many contexts, then, an increase/decrease in reliability means a decrease/increase in test length.

In practice, the ER condition needs to be viewed with a degree of common sense. For example, Morris (1982, p. 175), who summarizes an argument made by Beaton (1976), notes that, even if reliabilities are unequal, they can be made equal artificially by adding random noise to scores for the form with the

higher reliability. Doing so will lead to (or facilitate) satisfying FOE and/or SOE. Satisfying equity, however, is a poor justification for such an obviously inappropriate strategy.

This paper provides a theoretical justification for the conclusion that high reliabilities are not required to achieve FOE or SOE when the true equating function is linear, which is a rather rare condition. In such situations, however, high reliabilities might well increase the likelihood that FOE (and sometimes SOE) will be nearly satisfied. Consider, for example, the following two situations.

First it is obvious from Equations 13 and 16 that FOE is satisfied for linear OSE when $a = a_T$ and $b = b_T$. As $\rho_X^2 \rightarrow 1$ and $\rho_Y^2 \rightarrow 1$, $a \rightarrow a_T$ and $b \rightarrow b_T$, whether or not it is strictly true that $\rho_X^2 = \rho_Y^2$. In this sense, all other things being equal, the closer we get to perfect reliability, the more likely it is that FOE will be satisfied for linear OSE.

Second, in at least some equating contexts, it is difficult to make a compelling argument that $\rho_X^2 = \rho_Y^2$ is necessarily true. Suppose, for example, that the intended domain of coverage for a test is very broad, but the actual test forms are rather short. Under this circumstance, it seems very unlikely that scores on Forms X and Y will have equal reliability, since, among other things, the forms are unlikely to sample the same content. Longer test forms, however, may better approximate the condition that $\rho_X^2 = \rho_Y^2$ and, of course, longer test forms are likely to be more reliable. In this sense, in at least some contexts, it seems more likely that FOE and SOE will be more nearly satisfied when reliabilities are reasonably high.

For both linear and curvilinear equating, all other things being equal, as form scores get more reliable, $\sigma^2(E_X|\tau_y)$ and $\sigma^2(E_Y|\tau_y)$ become smaller for all τ_y.[3] In the limit, of course, they are zero. If $\sigma^2(E_X|\tau_y)$ and $\sigma^2(E_Y|\tau_y)$ are quite small, logically examinees should have little preference for one form over the other, even if reliabilities are not precisely equal. From this "matter of indifference" perspective, forms with highly reliable scores are to be preferred for equating.

When Lord (1980, pp. 195ff) introduced the notion of equity, he concluded that fallible scores on Forms X and Y cannot be equated under a strict definition of equity. He did not conclude, however, that the magnitude of reliability had no bearing on the degree to which equity might be achieved approximately. The results in this paper suggest that, all other things being equal, if approximate equity is a desirable goal, then almost always test forms with scores that are highly reliable are to be preferred over test forms with scores that have low reliability.

Various lists of criteria, requirements, or properties have been proposed for equating. (See Harris & Crouse, 1993 for a somewhat dated but still relevant review.) Most of the entries in most of these lists have an "aspirational" characteristic, since they are not perfectly achievable in most contexts. Some lists

[3]For purposes of this discussion, is is assumed that observed scores are in the proportion-correct metric, not the total score metric.

specifically include equal reliability (see, for example, Holland & Dorans, 2006, p. 194), which is certainly justified by the results in this paper. Other lists effectively include equal reliability indirectly through a "same specifications" property (e.g., Kolen & Brennan, 2004, p. 10).

None of the most frequently cited lists include high reliability, although high reliability is occasionally cited as a desirable characteristic for meaningful equating (see, for example, Dorans & Walker, p. 183). Also, Dorans and Holland (2000) and Holland and Dorans (2006, p. 216) argue that high reliability facilitates achieving population invariance (another often-cited equating criterion or property), although Brennan (2008) argues that high reliability is not a strict requirement for population invariance.

By contrast, this paper has demonstrated that relatively high reliability is necessary for approximating the requirements of equity, which is universally viewed as a desirable property for equating. Strictly speaking, this high-reliability "requirement" applies only when the true equating function is curvilinear. This is not much of a restriction, however, since almost always the true equating function is curvilinear in real-world situations. Linear estimation procedures are often used, but almost always such procedures are employed for practical reasons, not because there is any assumption that the true equating function is linear. In short, the results in this paper suggest that relatively high reliability deserves consideration as a desirable property for equating.

4.2 Limitations

This paper is primarily a theoretical treatment of FOE and SOE. As such, it does not treat issues concerning the estimation of parameters in equating relationships for various designs. These are important matters, of course, that must be addressed in practical applications.

While this paper emphasizes the role of reliabilities of X and Y in attaining, or approximately attaining, FOE and SOE, this paper does not address issues associated with estimating reliability. It is certainly not necessarily the case, for example, that Coefficient α is the appropriate estimate of reliability. In fact, in most equating contexts, a generalizability coefficient (see Brennan, 2001) that explicitly incorporates multiple, potential sources of error is likely to be much more defensible than Coefficient α.

This paper does not explicitly consider FOE and SOE under non-linear transformations of observed scores. Many such transformations are nearly linear throughout much of the range, and, if so, the results in this paper should apply approximately. Otherwise, however, there are subtle and potentially difficult issues that need to be addressed with non-linear transformations. For example, with ATSE we need the transformation in terms of true scores, but with OSE we need the transformation in terms of observed scores. There is no necessary reason why the two transformations must be the same.

5 References

Beaton, A. E. (1976). Comments on equating on the Cohort Charge Study. *Educational Testing Service Memorandum*, December 14. Princeton, NJ: ETS.

Brennan, R. L. (2008). A discussion of population invariance. —emphApplied Psychological Measurement, 32, 102-113.

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

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

Dorans, N. J., & Walker, M. E.. (2007). Sizing up linkages. In N. J. Dorans, M. Pommerich, & P. W. Holland (Eds.), Linking and aligning scores and scales (pp. 179–198). New York: Springer-Verlag.

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

Haertel, E. H. (2006). Reliability. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 65-110). Westport, CT: American Council on Education/Praeger.

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.

Harris, D. J., & Crouse, J. D. (1993). A study of criteria used in equating. *Applied Measurement in Education, 6*, 195–240.

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

Kolen, M. J., & Brennan, R. L. (1987). Linear equating models for the common-item nonequivalent-populations design. *Applied Psychological Measurement, 11*, 263–277.

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

Lord, F. M. (1980). *Applications of item response theory to practical testing*

problems. Hillsdale, NJ: Erlbaum.

Lord, R. M., & Novick, M. R. (1968). *Statistical theories of mental test scores.* Reading, MA: Addison-Wesley.

Morris, C. N. (1982). On the foundations of test equating. In P. W. Holland and D. B. Rubin (Eds.), *Test equating* (pp. 169–191). New York: Academic Press.

6 Appendix: Tables and Figures for Example

Table 1: Moments and KR21 for Distributions

Form	Distribution	Mean[a]	SD[a]	Skew[b]	Kurt[b]	KR21
X	$\beta(4,3)$	.571	.175	-.181	2.127	
X	NH(4,3,20)	.571 (11.429)	.203 (4.066)	-.183	2.424	.741
X	NH(4,3,40)	.571 (22.857)	.190 (7.586)	-.182	2.439	.851
X	NH(4,3,60)	.571 (34.286)	.185 (11.093)	-.182	2.441	.900
Y	$\beta(3,3)$	.500	.189	.000	2.333	
Y	NH(3,3,20)	.500 (10.000)	.215 (4.309)	.000	2.315	.769
Y	NH(3,3,40)	.500 (20.000)	.203 (8.106)	.000	2.328	.870
Y	NH(3,3,60)	.500 (30.000)	.198 (11.892)	.000	2.331	.909

[a]Moments for total-score metric are in parentheses.

[b]Skewness and kurtosis are the same for both the mean-score and total-score metrics.

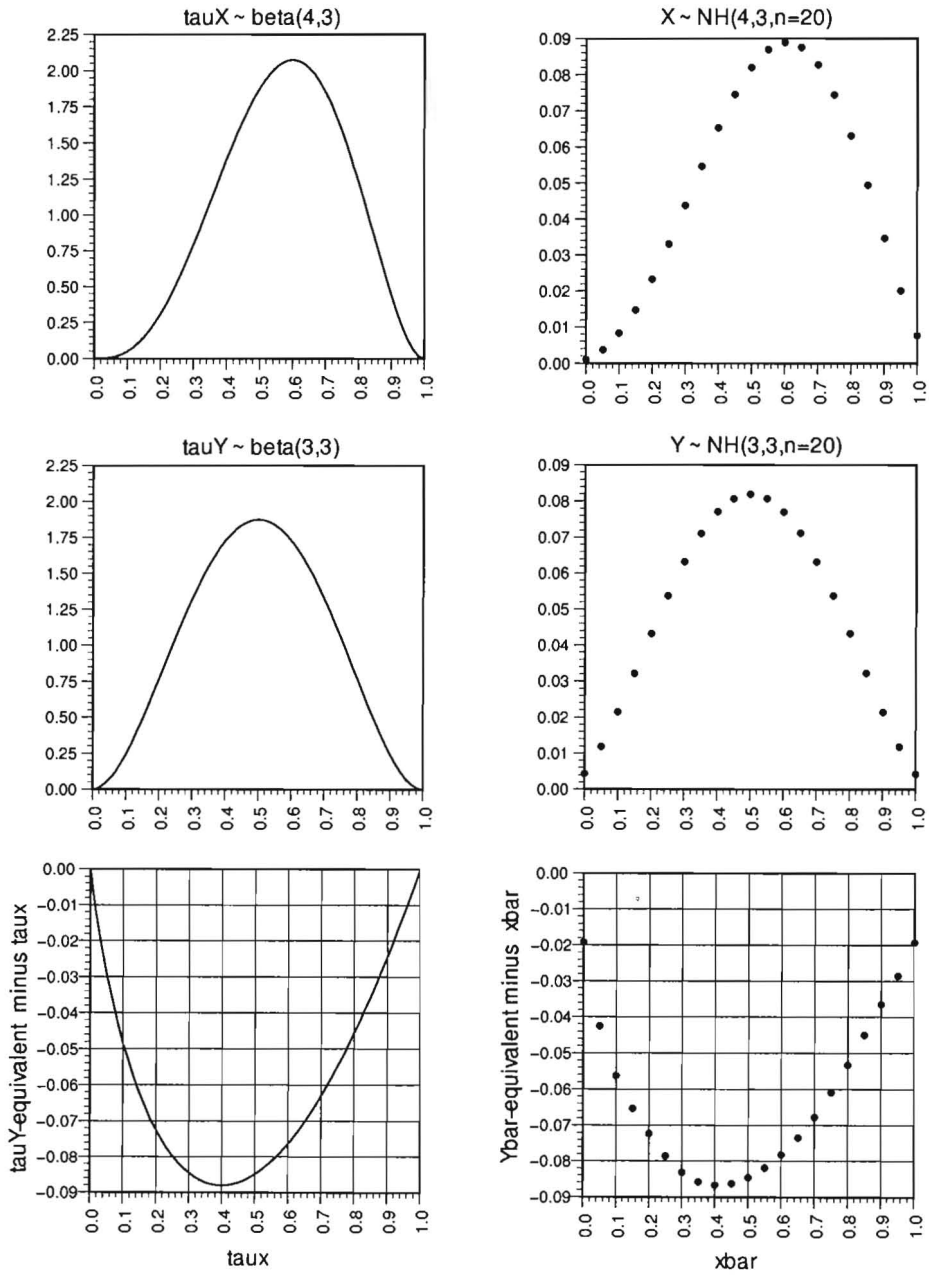

Figure 1: Beta densities and $X \to Y$ equating; negative hypergeometric densities and $X \to Y$ equating for $n = 20$.

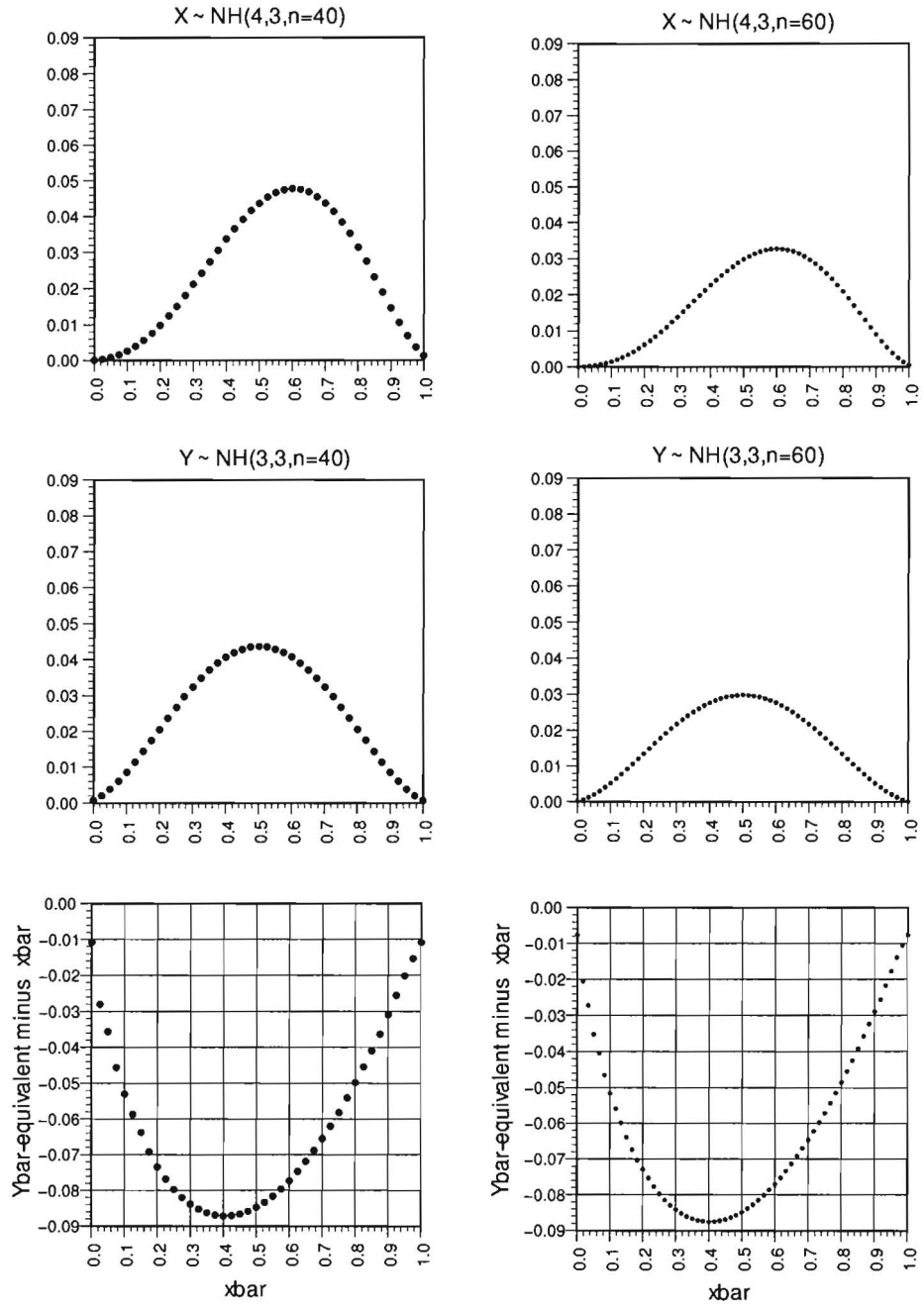

Figure 2: Negative hypergeometric densities and $X \rightarrow Y$ equatings for $n = 40$ and $n = 60$.

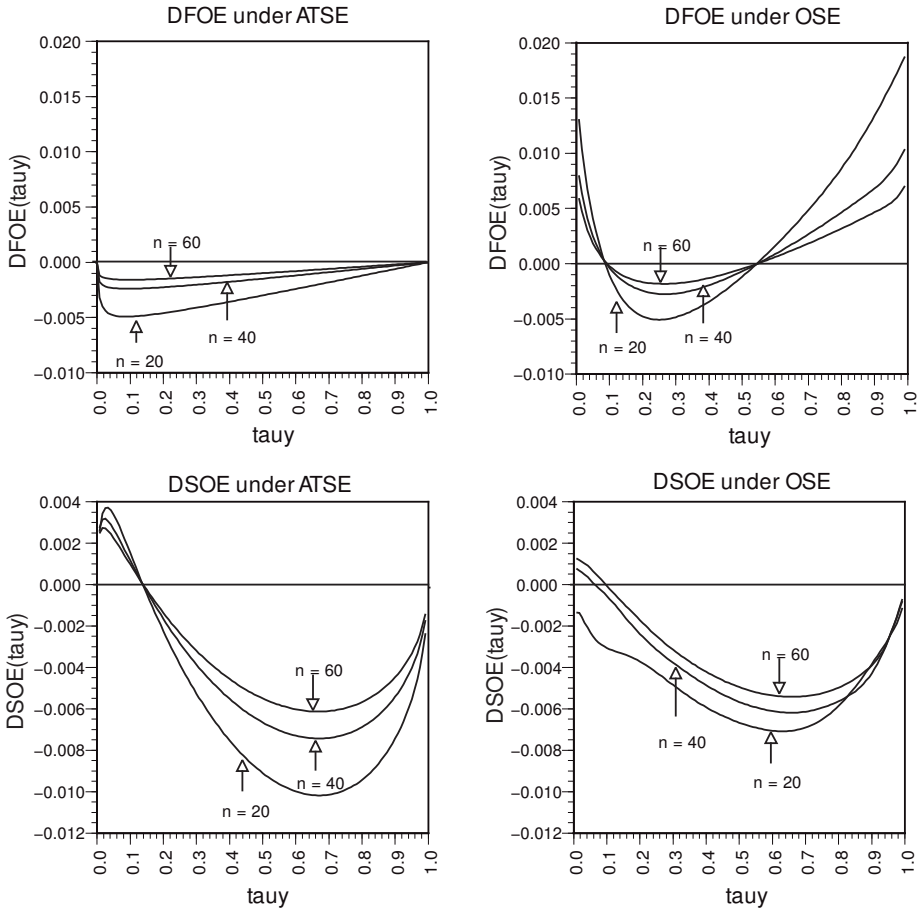

Figure 3: FOE and SOE, for ATSE and OSE, for $n = 20$, $n = 40$, and $n = 60$.

Table 2: Average Values for |DFOE(τ_y)| and |DSOE(τ_y)|

			$n = 20$		$n = 40$		$n = 60$
TSE	Weighted[a] Ave.	|DFOE(τ_y)|	.00296	>	.00146	>	.00097
TSE	Unweighted[b] Ave.	|DFOE(τ_y)|	.00285	>	.00141	>	.00093
TSE	Weighted[a] Ave.	|DSOE(τ_y)|	.00787	>	.00576	>	.00475
TSE	Unweighted[b] Ave.	|DSOE(τ_y)|	.00663	>	.00488	>	.00303
OSE	Weighted[a] Ave.	|DFOE(τ_y)|	.00374	>	.00204	>	.00138
OSE	Unweighted[b] Ave.	|DFOE(τ_y)|	.00570	>	.00307	>	.00209
OSE	Weighted[a] Ave.	|DSOE(τ_y)|	.00595	>	.00500	>	.00433
OSE	Unweighted[b] Ave.	|DSOE(τ_y)|	.00497	>	.00405	>	.00354

[a] Weighted by the density of true scores on Form Y, $\beta(3, 3)$.

[b] τ_y having values .008(.008).992.