Research Report

Limits on the Accuracy of Linking

Shelby J. Haberman

October 2010

ETS RR-10-22

Limits on the Accuracy of Linking

Shelby J. Haberman
ETS, Princeton, New Jersey

October 2010

As part of its nonprofit mission, ETS conducts and disseminates the results of research to advance quality and equity in education and assessment for the benefit of ETS's constituents and the field.

To obtain a PDF or a print copy of a report, please visit:

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

Technical Review Editor: Matthias von Davier

Technical Reviewers: Hongwen Guo and Sandip Sinharay

Abstract

Sampling errors limit the accuracy with which forms can be linked. Limitations on accuracy are especially important in testing programs in which a very large number of forms are employed. Standard inequalities in mathematical statistics may be used to establish lower bounds on the achievable inking accuracy. To illustrate results, a variety of equating problems are considered.

Key words: Fisher information, randomized blocks, mean equating, linear equating, two-way layout

Acknowledgments

This report has benefitted from conversations with Neil Dorans and from comments by Matthias von Davier, Sandip Sinharay, and Hongwen Guo.

In practice, the accuracy of equating or linking is limited because estimates used in the process are based on samples. Limitations on accuracy are encountered even under ideal conditions. These limitations can be computed by use of classical statistical inequalities. Bounds on accuracy involve the number of examinees involved in the equating process and the number of forms that must be linked. The limits also involve what assumptions are made concerning forms which contain common items. In typical cases, the most important issue in practice is that the number of examinees available per unit time is affected to only a very limited extent by an increase in the number of forms used within that time interval. Thus more administrations typically leads to fewer examinees per administration. If security considerations limit the number of administrations in which a form can be used, then it necessarily follows that more administrations per unit time results in more forms per unit time, and information concerning each form is based on fewer examinees. Two problems arise simultaneously. Because information concerning a form involves fewer examinees, estimates of form characteristics related to the difficulty of the form become less accurate. In addition, equating and linking involve comparison of different forms. As the number of forms becomes increasingly large due to security constraints, comparisons between different forms must in some cases become increasingly indirect. Two forms to be compared will not have been used together and will share no common items. Even more indirection is involved. Consider the following hypothetical case. A form used on September 1, 2009, may share common items with a form used on January 8, 2009, and with a form used on October 22, 2008. A form used on September 8, 2009, may share common items with a form used on February 22, 2009, and with a form used on July 21, 2008. Thus the September 1, 2009 form can only be linked to the September 8, 2009 form through whatever links are available for the forms used on July 21, 2008, October 22, 2008, January 8, 2009, and February 22, 2009. It may well be the case that none of these forms share any common items, so that further steps are needed to provide linkage. These many steps required to link the two forms used one week apart result in increased equating error due to sampling effects.

In practice, as suggested by some of the results in this report, the standard error associated with equating typically will increase at least in proportion to the square root of the number of forms used in the time interval. When restrictions are placed on the number of times a form can be used, the standard error associated with equating typically will increase in proportion to the number of forms used in the time interval. This point is illustrated for mean equating in

Section 1.2. Thus a testing program with quite satisfactory equating accuracy with six forms per year may have quite unsatisfactory equating accuracy with 60 forms per year.

To illustrate the issues involved, it is helpful to look at some simple examples of equating procedures. In Section 1, some cases of mean equating are explored. In Section 3, linear equating is explored. In this section, application of results of linear equating are also discussed in terms of implications for equating by item response theory. Section 4 considers some consequences of the analysis in this report. A basic knowledge of equating methods is assumed (Kolen & Brennan, 2004; von Davier, Holland, & Thayer, 2004); however, most of the analysis relies on basic statistical theory. The examples are deliberately chosen to be relatively simple, so that the basic issues can be discussed.

1 Mean Equating

Mean equating is a very simple equating procedure in which a constant is added to the raw score to adjust for differences in form difficulty. The approach is most appropriate for observed scores that are normally distributed and have the same variance; however, it can be used more generally. For an initial example, consider the following equating sequence in which randomly equivalent groups of examinees are employed at each administration to link test forms. This example involves a case in which two test forms are used at each administration and a given test form is never used for more than two administrations. Let $T \geq 2$ administrations be considered for N examinees. For example, one might have 24 administrations over a period of two years, with one administration per month, and there might be 240,000 examinees over the two-year period. For simplicity, let $M = N/(2T)$ be an integer. In Administration t, $1 \leq t \leq T$, let N/T examinees be assigned at random to two groups of M examinees. Thus in the hypothetical example, 5,000 examinees are in each of the two randomly equivalent groups at each administration. A total of $U = T + 1$ distinct forms numbered from 1 to U are used for the T administrations. Thus in the hypothetical example, $U = 25$ forms need to be linked. At Administration t, the first randomly equivalent group of examinees, Group 1, receives Form t, and the other group of examinees, Group 2, receives Form $t + 1$. In this design, Form t and Form $t + 1$ can be directly compared, for they are used on equivalent groups of examinees.

With some assumptions, it is possible to compare Forms t and u, $1 \leq t < u \leq T$, even when they are never employed in the same administration. To do so, let the observed score of

Examinee i, $1 \leq i \leq M$, in Group k at Administration t be X_{ikt}. Note that it is assumed for simplicity that no examinee receives more than one form at an administration and no examinee appears in more than one administration. Thus Examinee i of Group k in Administration t has no relationship to Examinee j of Group m at Administration u unless $t = u$, $k = m$, and $i = j$. Let the observed score X_{ikt} have mean μ_{kt} and variance σ^2. Thus the mean of the score X_{ikt} depends on the Group k of the examinee and the Administration t; however, the variance of the score is independent of both group and administration. This assumption, which simplifies analysis, is appropriate for mean equating. Consistent with the assumption that examinees for different groups and administrations are distinct individuals, assume that the X_{ikt} are all independently distributed. To simplify discussion of the impact of equating error, assume that the reliability of each form is the same for each administration. Thus the reliability of Form t or $t + 1$ at Administration t is ρ^2, where $0 < \rho^2 < 1$. In mean equating, $d_t = \mu_{1t} - \mu_{2t}$ provides a measure of the difficulty of Form $t + 1$ relative to Form t. This measure is based on the distributions of raw scores at Administration t for Groups 1 and 2. The fundamental assumption to make for comparisons of forms not used in the same administration is that the difference in difficulty of two forms would be the same were the forms used for other administrations. Thus one may let $D_1 = 0$ and $D_{u+1} = D_u + d_u$ for $u \geq 1$. Then D_u is a measure of the relative difficulty of Form u compared to Form 1. If Form 1 is the base form used in equating or linking, then a raw score of x on Form u would be converted to an equated raw score of $x + D_u$ on Form 1 if D_u were known. This result can be obtained in stages. A score of x on Form 2 is converted to a score of $x + D_2 = x + d_1$ on Form 1. Note that d_1 is greater than 0 if the mean score on Form 2 at Administration 1 is lower relative to the mean score on Form 1 at Administration 1 (Form 2 is more difficult than Form 1 at Administration 1). A score of x on Form 3 is converted to a score of $x + d_2$ on Form 2 based on comparison of Form 2 and Form 3 at Administration 2. In turn, the score of x on Form 3 is converted to a score of $x + d_2 + d1 = x + D_3$ on Form 1. In this manner, a score of x on Form u is eventually converted to a score of $x + D_u$ on Form 1. In the hypothetical example, in the first year, the score x from Form 3, which was used in Group 2 of the February administration in the first year, is thus converted to a score $x + D_3$ on Form 1, which was used in Group 1 of the January administration in the first year of testing. This conversion is also obtained with a series of equipercentile equatings if the X_{ikt} are all normally distributed, the distributions are known, and chained equating is used.

In practice, the means required for the conversion of scores are not known and must be estimated. For each Group k from each Administration t, the mean μ_{kt} of the scores X_{ikt} for Group k and Administration t can be estimated by the sample mean

$$\bar{X}_{kt} = M^{-1} \sum_{i=1}^{M} X_{ikt}.$$

Thus the difference d_t, $t \geq 1$, in the difficulty of Forms $t+1$ and t is estimated, based on Administration t, to be the difference

$$\hat{d}_t = \bar{X}_{1t} - \bar{X}_{2t}$$

between the sample means for Group 1, which received Form t, and Group 2, which received Form $t+1$. The estimate $\hat{d}_t$ is unbiased, so that the expectation of $\hat{d}_t$ is d_t, and the variance of $\hat{d}_t = 2\sigma^2/M$. In turn, the conversion size D_u for conversion from Form u to Form 1 is estimated for $u > 1$ by

$$\hat{D}_u = \sum_{t=1}^{u-1} \hat{d}_t.$$

The estimate $\hat{D}_u$ is unbiased, so it has expectation D_u, and the variance of $\hat{D}_u$ is $(u-1)[2\sigma^2/M] = 4(u-1)T\sigma^2/N$. The standard error of $\hat{D}_u$ is then $2[(u-1)T/N]^{1/2}\sigma$. For comparison, the variance of measurement of the score X_{ikt} of Examinee i in Group k of Administration t is $\sigma^2(1-\rho^2)$, so that the ratio of the variance of equating error to the variance of measurement is

$$G_u = 4(u-1)T/[N(1-\rho^2)]$$

in the case of Form u. This ratio increases as the form count u and number of forms $U = T+1$ increase, as the reliability ρ^2 increases, and as the total sample size N decreases. The highest ratio is found for the last form used, for here $u = U = T+1$. In the hypothetical example of 24 administrations over two years for 240,000 examinees, suppose that $\sigma = 10$ and $\rho^2 = 0.9$. In this case, the variance of measurement is 10, and the estimate $\hat{D}_u$ has variance $(u-1)/25$. This variance is only 0.04 for $u = 2$, and $G_1 = 0.004$ is a rather small ratio of variance of equating compared to variance of measurement. Nonetheless, for $u = U = 25$, the variance of $\hat{D}_U$ of 0.96 is not negligible, and $G_U = 0.096$ is large enough that equating error has some effect on the effective reliability of the test. Note that this example involves a substantial number of examinees for a testing program and a number of administrations over two years that is not exceptional.

Some further examples may help provide some perspective. If the total number of examinees is $N = 200,000$, the common reliability coefficient is $\rho^2 = 0.9$, a total of $T = 10$ administrations are used, and equating error is considered for the last form, so that $u = U = 11$, then $G_U = 0.02$ is relatively small, so that equating error is relatively small compared to measurement error for a form. This example applies to a testing program with many examinees and a moderate number of administrations.

If the number of examinees is reduced to $N = 2,000$ but the reliability ρ^2 remains 0.9, the total number T of administrations remains 10, and the form considered remains Form U, then $G_U = 2$ is very large. If the standard deviation of the scores X_{ikt} is $\sigma = 10$, then the variance of measurement of 10 is half the variance 20 of equating. This example applies to a rather small testing program with very small administration sizes of 200 examinees.

If $N = 200,000$ receive the test at some time, the reliability coefficient is still $\rho^2 = 0.9$, the form number is $U = 51$, and the number of administrations is $T = 50$, then the ratio $G_U = 0.10$ is not negligible. The variance of equating error is a tenth of the variance of measurement. Here the number of examinees is fairly large, but the number of administrations is also large.

Assessment of the impact of equating error depends on whether the examinee is regarded as taking a random examination or not. For an examinee who uses Form u at Administration u, the effective variance of measurement is $\sigma^2(1 + G_u)$, the sum of the variance of measurement and the variance of equating. A slight change in the formula occurs for Form $u + 1$ and Administration u because the examinee is part of the data used for estimation of $\hat{D}_{u+1}$. The effective variance of measurement is then $\sigma^2(1 - M^{-1} + G_{u+1})$. If the examinee is regarded as taking a fixed examination, then the equating error $\hat{D}_u - D_u$ has approximate probability 0.05 of benefitting or harming the examinee by more than $1.96\sigma[(1 - \rho^2)G_u]^{1/2}$. This probability is exact if all score distributions are normal. This criteria is somewhat stricter. Relative to the standard error of measurement $\sigma(1 - \rho^2)^{1/2}$, one has

$$1.96\sigma[(1 - \rho^2)G_u]^{1/2}/[\sigma(1 - \rho^2)^{1/2}] = 1.96G_u^{1/2}.$$

Consider the previous examples. In the example with 240,000 examinees, 24 administrations, a standard deviation of scores of $\sigma = 10$, and a reliability coefficient of 0.9, by the last form used ($U = 25$), the effective variance of measurement, 10.96, is appreciably greater than the actual variance of measurement of 10. By the perspective of a fixed administration, the probability is

0.05 that the equating error is at least 1.88 and the standard error of measurement is $10^{1/2} = 3.16$. By this criterion, there is a substantial possibility of a substantial impact of equating error on the reported score.

In the case of $N = 200,000$ examinees, a reliability coefficient of $\rho^2 = 0.9$, a standard deviation $\sigma = 10$ of scores, Form $U = 11$, and Administration $T = 10$, the effective variance of measurement of 10.2 is not much more than the variance of measurement of 10. On the other hand, the probability is 0.05 that the equating error changes the score reported by at least $1.96G_U^{1/2} = 0.28$, an amount not negligible relative to a standard deviation of measurement of 3.16.

In the case of a small number $N = 2,000$ of examinees, a reliability coefficient $\rho^2 = 0.9$, a standard deviation of scores of $\sigma = 10$, Form $U = 11$, and Administration $T = 10$, the effective variance of measurement of 30.0 is very large relative to the variance of measurement of 10, and the probability is 0.05 that equating error changes the reported score by at least $1.96G_U^{1/2} = 2.77$, a very large change relative to 3.16, the standard error of measurement.

For an example with many examinees and many administrations, let $N = 200,000$ be the number of examinees, let $\rho^2 = 0.9$ be the reliability coefficient, let the form number be $U = 51$, and let the administration number be $T = 50$. In this case, the effective variance of measurement is 11.0 is substantially greater than the variance of measurement of 10, and the probability is 0.05 that the equating error changes a score by at least $1.96G_u^{1/2} = 0.62$, a fairly large value relative to the standard error of measurement of 3.16.

These computations illustrate a basic issue in terms of assessment design. As long as the total number of examinees under study does not vary, increasing the number of administrations dramatically increases the variability of results. The average variance of equating over all administrations is

$$(2T)^{-1} \sum_{t=1}^{T} [4tT\sigma^2/N + 4(t-1)T\sigma^2/N] = 2T^2\sigma^2/N.$$

For a fixed total number N of examinees, doubling the number of administrations quadruples the average variance of equating and doubles the root mean squared equating error for the N examinees.

1.1 Many Parallel Forms at Each Administration

Modification of the method of data collection yields substantially different results. Suppose that the same $U \geq 2$ forms are used at each administration from 1 to T, and let the total number

of examinees N be a multiple of TU. At Administration t, let examinees be divided randomly into $K = U$ groups of equal size, and let each form be given to the $M = N/(TU)$ examinees in Group u. Let the score of examinee i, $1 \le i \le M$, from Group u at Administration t be X_{iut}. As in Section 1, let X_{iut} have mean μ_{ut} and variance σ^2. Assume that the X_{iut} are all independently distributed, and assume that the reliability of Form u at Administration t is ρ^2, where $0 < \rho^2 < 1$. Assume that μ_{ut} satisfies an additive model $\mu_{ut} = \mu_{1t} - D_u$. In this case, D_u measures the difficulty of Form u relative to Form 1. The assumption is made that the relative difficulty of Form u compared to Form 1 is the same for all administrations. If Form 1 is the base form, then a score x on Form $u \neq 1$ is equated to a score $x + D_u$ on Form 1, provided that D_u is known.

In practice, the relative form difficulty D_u of Form u must be estimated for $u > 1$. For efficient estimation, let equating at Administration t be based on all data available from Administration t and from any prior administrations. Note that this procedure, although statistically appropriate, does lead to a situation in which two examinees with identical raw scores can have different reported scores if they take the same form at different administrations. For each Administration h, $1 \le h \le t$, the form difficulty D_u has an unbiased estimate $\bar{X}_{1h} - \bar{X}_{uh}$, where the sample mean

$$\bar{X}_{uh} = M^{-1} \sum_{i=1}^{M} X_{iuh}$$

estimates the population mean μ_{uh} for any Form u and Administration h. The estimates $\bar{X}_{1h} - \bar{X}_{uh}$, $1 \le h \le t$, are independent and have common variance variance $2\sigma^2/M$. Thus the estimate of D_u at Administration t is obtained by averaging the estimates of form difficulty from the first t administrations. The resulting estimate is

$$\hat{D}_{u \cdot t} = -t^{-1} \sum_{h=1}^{t} (\bar{X}_{uh} - \bar{X}_{1h}).$$

This estimate is unbiased, so that the expectation of $\hat{D}_{u \cdot t}$ is D_u, and the variance of $\hat{D}_{u \cdot t}$ is $2\sigma^2/(tM) = 2TU\sigma^2/(tN)$. For comparison, the variance of measurement of X_{kt} is $\sigma^2(1 - \rho^2)$, so that the ratio of the equating error to the variance of measurement is

$$G_{u \cdot t} = 2TU/[tN(1 - \rho^2)]$$

for each Form u. This ratio decreases as the Administration t increases and as the total sample size N increases, but the ratio increases as the number T of administrations, the number U of forms, and the reliability increase. The ratio $G_{u \cdot T} = 2U/[N(1 - \rho^2)]$.

For comparison with the equating design presented at the start of Section 1, consider an example with $N = 200,000$ examinees, a reliability coefficient $\rho^2 = 0.9$, $U = 10$ forms, $T = 10$ administrations, and Administration $t = 10$. Then, for each Form $u > 1$, $G_{u \cdot T} = 0.001$ is quite small. Note how much smaller $G_{u \cdot T}$ is than the value $G_U = 0.02$ in Section 1 achieved for Form $U = 11$ for $N = 200,000$ examinees and $T = 10$ administrations. Even for the much less favorable case of the initial administration, $G_{u \cdot 1}$ is 0.01 for each Form $u > 1$. Thus the use of many parallel forms permits much more accurate estimation of the conversion constants D_u than was the case in the design in Section 1. Nonetheless, much smaller sample sizes are much less satisfactory. Consider the case of $N = 2,000$ examinees. Let the reliability coefficient remain $\rho^2 = 0.9$, let the number T of administrations and the number U of forms both remain 10. In this case, only 20 examinees in an administration receive the same form. Not surprisingly, $G_{u \cdot 1}$ is 1, so that the variance of the equating conversion is as large as the variance of measurement. By the last administration, $G_{u \cdot T}$ is 0.1, a figure which is not negligible but obviously much better than for the first administration. For a case with many administrations but a moderate number of forms, consider $N = 200,000$ examinees, a reliability coefficient $\rho^2 = 0.9$, $T = 50$ administrations, and $U = 10$ forms. By the final administration, $G_{u \cdot T} = 0.001$ is the same as in the previous example with 200,000 examinees. On the other hand, the situation for the initial administration is rather less satisfactory, for $G_{u \cdot 1}$ is then 0.05. Thus the variance of the examinee's reported score due to equating is 0.05 times as great as the variance of measurement.

In practice, despite the favorable results, the design with many parallel forms used in each administration can be difficult to apply, both due to limitations in the ability of testing programs to administer a large number of forms in the same administration and due to security considerations. Even if a very large number of forms can be regarded as only a limited security risk due to the difficulty of determining in advance the answers for the very large number of items in all the forms, there remains the problem of starting out. For initial administrations, equating accuracy can be quite limited. Some delay in initial reporting until more administrations are completed can alleviate the problem, but pressure to report scores promptly may render this design impractical. As a consequence, it is appropriate to consider other alternatives.

1.2 Design Limits

Somewhat more complex equating designs based on randomly equivalent groups may be developed. These designs do not result in improvements in results when the design in Section 1.1 can actually be used, but the designs can be employed to indicate inherent limitations in equating accuracy once security considerations and reporting deadlines restrict form reuse and restrict the ability to delay reporting until data are available from more administrations. The fundamental issue is that under a restriction that no form can appear in more than a specified number of administrations, as the same number of examinees is divided into more administrations, the average equating error, as measured in mean squared error, across administrations becomes proportional to at least the square of the number of administrations. In terms of root mean squared error, this measure of accuracy is at least proportional to the number of administrations, so that multiplying the number of administrations by 10 decreases accuracy by the criterion of root mean squared error by a factor of 10.

To discuss equating designs for randomly equivalent groups, the following design is introduced to generalize the equating designs of Sections 1 and 1.1. Consider $T \geq 2$ administrations, N examinees, and $U \geq 2$ forms. At each Administration t, there are $K \geq 2$ equivalent groups k, $1 \leq k \leq K$, of $M = N/(KT) \geq 1$ examinees, and Group k is administered Form u_{kt}. For simplicity, assume that $u_{11} = 1$, so that Form 1, the base form, is administered in Administration 1 to the M examinees in Group 1. Note the implicit assumption that N is an integer multiple of KT. It is still assumed that the raw score X_{ikt} of Examinee i from Group k at Administration t is a random variable with mean μ_{kt} and variance $\sigma^2 > 0$, and it is assumed that the reliability coefficient is ρ^2 for each combination of form and administration. The examinee scores X_{ikt} are still assumed to be mutually independent. The assumption on the mean μ_{kt} of the scores for Group k of Administration t is that μ_{kt} is additive in administration and form, so that

$$\mu_{kt} = \alpha_t - D_{u_{kt}}$$

for some real number constants α_t, $1 \leq t \leq T$, and some real constants D_u, $1 \leq u \leq U$. To identify parameters, it is assumed that $D_1 = 0$, so that α_{11} is the expectation of the score of examinees at Administration 1 who receive the base form, Form 1.

This additive model is consistent with additive models previous employed in Section 1 and 1.1. In Section 1, the number of groups in an administration is $K = 2$, the number of administrations

is T, and the number of forms is $U = T + 1$. At Administration t, $1 \leq t \leq T$,the administered forms are $u_{1t} = t$ and $u_{2t} = t + 1$. The parameter difference $d_u = D_{u+1} - D_u = \mu_{uu} - \mu_{(u+1)u}$ for Form u, $1 \leq u \leq T$, so that D_1 is 0 and $D_{u+1} = D_u + d_u$ for $1 \leq u \leq T$. It follows that $\alpha_t = \mu_{tt} + D_t$ for $1 \leq t \leq T$.

In Section 1.1, $K = U$ groups are used in each administration, and $u_{kt} = k$ for $1 \leq k \leq U$ and $1 \leq t \leq T$, so that Group u, $1 \leq u \leq U$, is administered Form u at Administration t. Here $D_u = \mu_{1t} - \mu_{ut}$ for each Form u and Administration t, and $\alpha_t = \mu_{1t}$ for $1 \leq t \leq T$ is the score mean for the examinees in Group 1 who received the base form, Form 1, at Administration t.

In general, the basic feature of the additive model is that the difference

$$\mu_{kt} - \mu_{k't} = D_{u_{k't}} - D_{u_{kt}}$$

in means is a function of the Forms u_{kt} and $u_{k't}$ administered at Administration t to Groups k and k'. The difference has no further dependence upon the Administration t. The difference $\beta_t = \alpha_t - \alpha_1$ provides a measure of the relative proficiency of examinees at Administration t compared to examinees at Administration 1. This proficiency difference is assumed independent of the Form u. The parameter D_u provides a measure of the difficulty of Form u relative to the difficulty of Form 1. This parameter is assumed not to depend on the administration.

As in sections 1 and 1.1, if the parameter D_u is known for Form $u > 1$, then a score x on Form u is converted to a score $x + D_u$ on Form 1. To estimate the parameters D_u for $u > 1$, least squares may be applied to obtain least-squares estimate $\hat{D}_u$ of D_u for $1 < u \leq U$. The constraint is imposed that $\hat{D}_1 = D_1 = 0$. Given the estimate $\hat{D}_u$ for a Form $u > 1$, a raw score of x on Form u can be equated to a raw score of $x + \hat{D}_u$ on Form u. Computation of the least-squares estimates of the D_u is a familiar task from the study of two-way analysis of variance with unequal numbers of observations in cells (Scheffé, 1959, p. 114), although conditions are required to ensure that all the parameters D_u, $2 \leq u \leq U$, are estimable.

To obtain least-squares estimates, a three-dimensional array m_{ktu}, $1 \leq k \leq K$, $1 \leq u \leq U$, $1 \leq t \leq T$, is used to specify the relationship between groups, administrations, and forms. For $1 \leq k \leq K$, $1 \leq t \leq T$, and $1 \leq u \leq U$, let m_{ktu} be 1 if $u_{kt} = u$, and let m_{ktu} be 0 otherwise. For example, in Section 1, $m_{1tt} = m_{2t(t+1)} = 0$ for $1 \leq t \leq T$ and $m_{ktu} = 0$ if u is not $t + k - 1$. In Section 1.1, $m_{ktk} = 1$ for $1 \leq k \leq K$ and $1 \leq t \leq T$ and $m_{ktu} = 0$ if $k \neq u$.

A number of restraints on the m_{ktu} necessarily exist. Only one form is administered at each

administration to each group. Thus for each Administration t and Group k,

$$\sum_{k=1}^{K} m_{ktu} = 1.$$

The number of Groups k receiving Form u at Administration t is

$$m_{+tu} = \sum_{k=1}^{K} m_{ktu}.$$

In Sections 1 and 1.1, this number is 0 or 1, but equating designs may be considered in which m_{+tu} can exceed 1. For Administration t, the number of groups is K, so that the sum

$$\sum_{u=1}^{U} m_{+tu} = K. \tag{1}$$

The sum

$$m_{++u} = \sum_{t=1}^{T} m_{+tu}$$

is the total number of groups that receive Form u in some administration. Because KT groups are present in the T administrations, the summation

$$\sum_{u=1}^{U} m_{++u} = KT. \tag{2}$$

To develop least-squares equations requires consideration of instances in which two forms appear in the same administration. Let

$$q_{uu'} = K^{-1} \sum_{t=1}^{T} m_{+tu} m_{+tu'}$$

for $1 \leq u \leq U$ and $1 \leq u' \leq U$. Note that $m_{+tu} m_{+tu'}$ is the number of pairs (k, k') of groups, $1 \leq k \leq K$ and $1 \leq k' \leq K$, such that, at Administration t, Group k receives Form u and Group k' receives Form u'. Use of (1) shows that

$$m_{++u} = \sum_{u'=1}^{U} q_{uu'}. \tag{3}$$

For Group k, $1 \leq k \leq K$, in Administration t, $1 \leq t \leq T$, let $\bar{X}_{kt}$ be the average of the examinee scores X_{ikt} for $1 \leq i \leq M$. For Administration t, let $\bar{X}_{+t}$ be the sum of the averages $\bar{X}_{kt}$ for $1 \leq k \leq K$. For Form u, $1 \leq u \leq U$, let $\bar{X}_u$ be the sum of $\bar{X}_{kt}$ for Administrations t and

Groups k such that Group k receives Form u ($u_{kt} = u$). The estimates $\hat{D}_u$, $1 \leq u \leq U$, satisfy the simultaneous equations

$$m_{++u}\hat{D}_u - \sum_{u'=1}^{U} q_{uu'}\hat{D}_{u'} = -\bar{X}_u + K^{-1}\sum_{t=1}^{T} m_{+tu}\bar{X}_{+t} \tag{4}$$

for $1 \leq u \leq U$, and $\hat{D}_1 = 0$. The $\hat{D}_u$, $1 \leq u \leq U$, have uniquely defined estimates if the m_{+tu}, $1 \leq t \leq T$, $1 \leq u \leq U$, satisfy the inseparability conditions (Goodman, 1968) that each Form u is used at least once in some Group k and Administration t ($m_{++u} > 0$ for $1 \leq u \leq U$) and no way exists to divide the U form numbers from 1 to U into two nonempty disjoint subsets A and B such that $q_{uu'} = 0$ if u is in A and u' is in B. It will be assumed that the inseparability assumption holds.

To examine the inseparability issue, first consider the equating design in Section 1. In this example, m_{++u} is 2 for $1 < u < U$ and $m_{++u} = 1$ for u equal 1 or U. Forms u and u' can only appear in the same administration if $|u - u'| \leq 1$. It follows that $q_{uu'} = 0$ if $|u - u'| > 1$, $q_{uu'} = 1/2$ if $|u - u'| = 1$, $q_{uu} = 1$ if $1 < u < U$, and $q_{11} = q_{UU} = 1/2$. Let t and t' be administration numbers for $1 \leq t \leq T$ and $1 \leq t' \leq T$, and let u and u' be form numbers for $1 \leq u \leq$U and $1 \leq u' \leq U$. If A and B are disjoint nonempty subsets of the integers from 1 to U and if each integer from 1 to U is in either A or B, then some u and u' must exist such that u is in A, u' is in B, and $|u - u'| = 1$. In such a case, $q_{uu'} > 0$. It follows that the inseparability assumption holds.

In Section 1, results are even simpler. Here $m_{++u} = T > 0$ for each Form u and $q_{uu'} = T/U > 0$ for any Forms u and u'. Because $q_{uu'}$ is always positive, the inseparability condition holds.

To study equating accuracy, variances of the estimates $\hat{D}_u$ are needed for Forms $u > 1$. For this purpose, the complete covariance matrix of the $\hat{D}_u$, $1 \leq u \leq U$, is determined. This covariance matrix is determined in stages. To begin, consider the U by U symmetric matrix $\mathbf{C}$ with the element in row u and column u', $1 \leq u \leq U$, $1 \leq u' \leq U$, equal to

$$C_{uu'} = m_{++u}\delta_{uu'} - q_{uu'} + \frac{(K-1)T}{U(U-1)},$$

where the Kronecker function $\delta_{uu'}$ is 1 if $u = u'$ and 0 otherwise. Observe that (3) implies that

$$\sum_{u'=1}^{U}(m_{++u}\delta_{uu'} - q_{uu'}) = 0$$

for $1 \leq u \leq U$. The inseparability condition implies that $\mathbf{C}$ is positive-definite and invertible.

By standard linear algebra, the matrix $\mathbf{C}$ has a decomposition into eigenvalues and eigenvectors such that

$$C_{uu'} = \sum_{v=1}^{U} \lambda_v w_{uv} w_{u'v},$$

where, for $1 \leq v \leq U$, the eigenvalue $\lambda_v > 0$, and the eigenvector $\mathbf{w}_v$ with elements w_{uv}, $1 \leq u \leq U$, satisfies the orthogonality conditions

$$\sum_{u=1}^{U} w_{uv} w_{uv'} = \begin{cases} 1, & v = v', \\ 0, & v \neq v', \end{cases}$$

for $1 \leq v' \leq v$. For $v = 1$, $\lambda_1 = (K-1)T/(U-1)$, and $w_{u1} = 1/U^{1/2}$. Standard linear algebra also implies that the inverse $\mathbf{C}^{-1}$ of $\mathbf{C}$ then has row u and column u' equal to

$$C^{uu'} = \sum_{v=1}^{U} \lambda_v^{-1} w_{uv} w_{u'v}.$$

To obtain the covariance matrix of the estimated conversion adjustments $\hat{D}_u$, $1 \leq u \leq U$, the differences

$$\tilde{D}_u = \hat{D}_u - U^{-1} \sum_{u'=1}^{U} \hat{D}_{u'}$$

between the estimate $\hat{D}_u$ and the average $U^{-1} \sum_{u'=1}^{U} \hat{D}_{u'}$ are considered for $1 \leq u \leq U$. Obviously, $\tilde{D}_u$ estimates $D_u - U^{-1} \sum_{u=1}^{Y} D_{u'}$. By the basic theory of estimable functions (Rao, 1973, pp. 224–226), $(\sigma^2/M)C^{uu'}$ is the covariance of the estimates $\tilde{D}_u$ and $\tilde{D}_{u'}$. Because $\hat{D}_u = \tilde{D}_u - \tilde{D}_1$, it follows that the variance of $\hat{D}_u$ is

$$\sigma^2(\hat{D}_u) = \frac{\sigma^2}{M}(C^{uu} - 2C^{u1} + C^{11}) = \frac{\sigma^2}{M} \sum_{v=2}^{U} \lambda_v^{-1} (w_{uv} - w_{1v})^2.$$

Obviously, $\sigma^2(\hat{D}_1) = 0$. More generally, the variance of $\hat{D}_u - \hat{D}_{u'}$, $u \neq u'$, is

$$\sigma^2(\hat{D}_u - \hat{D}_{u'}) = \frac{\sigma^2}{M}(C^{uu} - 2C^{uu'} + C^{u'u'}) = \frac{\sigma^2}{M} \sum_{v=2}^{U} \lambda_v^{-1} (w_{uv} - w_{u'v})^2. \tag{5}$$

The average variance of $\sigma^2(\hat{D}_u - \hat{D}_{u'})$ is

$$\bar{\sigma}^2 = \frac{2\sigma^2}{M(U-1)} \sum_{v=2}^{U} \lambda_v^{-1}. \tag{6}$$

This result is based on a classical relationship between mean squared differences and sample variances. For real numbers x_u, $1 \leq u \leq U$,

$$[U(U-1)]^{-2} \sum_{u=1}^{U} \sum_{u'=1}^{U} (x_u - x_{u'})^2 = 2(U-1)^{-1} \left[\sum_{u=1}^{U} x_u^2 - U^{-1} \left(\sum_{u=1}^{U} x_u \right)^2 \right]. \tag{7}$$

For $v > 1$, $\sum_{u=1}^{U} w_{uv} = 0$ and $\sum_{u=1}^{U} w_{uv}^2 = 1$. Thus (5) and (7) imply (6).

A lower bound for $\bar{\sigma}^2$ is easily constructed by use of a classical inequality for the harmonic and arithmetic means (Hardy, Littlewood, & Pòlya, 1952, pp. 26–27). For any real numbers x_v, $2 \leq v \leq U$, their harmonic mean

$$\left((U-1)^{-1} \sum_{v=2}^{U} x_v^{-1} \right)^{-1}$$

is never greater than their corresponding arithmetic mean

$$(U-1)^{-1} \sum_{v=2}^{U} x_v,$$

with equality if, and only if, the x_v are all equal. Because the trace $\sum_{u=1}^{U} C_{uu}$ of $\mathbf{C}$ is the sum of its eigenvalues (Halmos, 1958, p. 105),

$$\sum_{v=1}^{U} \lambda_v = U(K-1)T/(U-1).$$

Because $\lambda_1 = (K-1)T/(U-1)$, it follows that

$$\sum_{v=2}^{U} \lambda_v = (K-1)T,$$

so that

$$\bar{\sigma}^2 \geq \frac{2K\sigma^2(U-1)}{N(K-1)},$$

with equality if, and only if, λ_v is constant for $v > 1$. The condition that λ_v is constant for $v > 1$ holds if, and only if, $\lambda_v = (K-1)T/(U-1)$ for $1 \leq v \leq U$ and $\mathbf{C} = [(K-1)T/(U-1)]\mathbf{I}$, where $\mathbf{I}$ is the U by U identity matrix. When the lower bound on $\bar{\sigma}^2$ is achieved,

$$\sigma^2(\hat{D}_u - \hat{D}_{u'}) = \bar{\sigma}^2$$

for Forms u and $u' \neq u$.

Observe that, given a fixed number N of examinees, a fixed score variance σ^2, and a fixed number K of forms per administration, the lower bound on $\bar{\sigma}^2$ is proportional to $U-1$, the

number of forms minus 1. The square root of $\bar{\sigma}^2$, a measure of root mean squared error, is then proportional to $(U-1)^{1/2}$. This result does imply that more forms inevitably results in less accuracy, but the rate of increase does not directly involve the number of administrations. The key issue is that no constraint has been introduced on form reuse.

The lower bound on $\bar{\sigma}^2$ is achievable. It applies under the conditions of section 1.1, for $m_{++u} = T$ and $q_{uu'} = T/K$ for Forms u and u', where u and u' are positive integers no greater than $U = K$. Thus $\mathbf{C} = T\mathbf{I}$, and

$$\bar{\sigma}^2 = 2\sigma^2 U/N.$$

The lower bound on $\bar{\sigma}^2$ is also achieved for the balanced incomplete block case with $m_{++u} = KT/U$, $q_{uu} = T/U$, and $q_{uu'} = (K-1)T/[U(U-1)]$ for $u \neq u'$ (Cochran & Cox, 1957, ch. 11). In this case,

$$\bar{\sigma}^2 = \frac{2K\sigma^2(U-1)}{N(K-1)}.$$

Unfortunately, the practical constraints on the equating design of section 1.1 normally also apply to balanced incomplete blocks. In an equating design with balanced incomplete blocks, it is necessary that KT/U and $(K-1)KT/[U(U-1)]$ must both be integers. For a number T of administrations sufficiently large, this condition cannot hold if the security constraint is imposed that, for some integer $Q \geq 2$, the total number m_{++u} of times Form u is used in some group for some administration satisfies the constraint $m_{++u} \leq Q$ for $1 \leq u \leq U$. Thus KT/U cannot exceed Q and U must be at least KT/Q. Thus the lower bound on $\bar{\sigma}^2$ is then at least $2K(KT-Q)\sigma^2/[Q(K-1)N]$, so that more administrations T leads to a higher average variance $\bar{\sigma}^2$.

Lower bounds can also be considered for variances of linear contrasts of the estimated adjustments $\hat{D}_u$ for Forms u from 1 to U. These bounds can provide insight into commonly observed increases in variances of the $\hat{D}_u$ as the form number u increases. Let $\mathbf{b}$ be the U-dimensional vector with elements b_u, $1 \leq u \leq u$, where the sum of the b_u is 0. Consider the variance of the estimate

$$\hat{g} = \sum_{u=1}^{U} b_u \hat{D}_u = \sum_{u=1}^{U} b_u \tilde{D}_u$$

of

$$g = \sum_{u=1}^{U} b_u D_u.$$

Let

$$\mathbf{x}'\mathbf{y} = \sum_{u=1}^{U} x_u y_u$$

for any U-dimensional vector $\mathbf{x}$ with elements x_u for $1 \le u \le U$ and any U-dimensional vector $\mathbf{y}$ with elements y_u, $1 \le u \le U$. Then

$$\mathbf{y}'\mathbf{C}^{-1}\mathbf{y}\mathbf{x}'\mathbf{C}\mathbf{x} \ge (\mathbf{x}'\mathbf{y})^2$$

(Rao, 1973, p. 54), with equality if

$$\mathbf{y} = c\mathbf{C}\mathbf{x} \tag{8}$$

for some real c. Note that (8) implies that

$$\mathbf{y}'\mathbf{C}^{-1}\mathbf{y} = c^2\mathbf{x}'\mathbf{C}\mathbf{x} = c\mathbf{y}'\mathbf{x}.$$

The case of $\mathbf{x} = \mathbf{y} = \mathbf{b}$ shows that

$$\sigma^2(g) \ge \frac{\sigma^2(\mathbf{b}'\mathbf{b})^2}{M\mathbf{b}'\mathbf{C}\mathbf{b}},$$

where

$$\mathbf{b}'\mathbf{C}\mathbf{b} = \sum_{u=1}^{U} m_{++u} b_u^2 - \sum_{u=1}^{U}\sum_{u'=1}^{U} b_u b_{u'} q_{uu'}.$$

By (3), it follows that

$$\mathbf{b}'\mathbf{C}\mathbf{b} = 2^{-1}\sum_{u=1}^{U}\sum_{u'=1}^{U} (b_u - b_{u'})^2 q_{uu'}.$$

Equality holds if, and only if, for some real c,

$$b_u = c\sum_{u'=1}^{U} q_{uu'}(b_u - b_{u'})$$

for $1 \le u \le U$, so that

$$\sigma^2(g) = \sigma^2 c\mathbf{b}'\mathbf{b}/M.$$

For example, if v and v' are distinct positive integers no greater than U, $b_v = 1$, $b_{v'} = -1$, and $b_u = 0$ for u neither equal to v nor v', then

$$\sigma^2(\hat{D}_v - \hat{D}_{v'}) \ge \frac{4K\sigma^2}{M[(K-1)(m_{++v} + m_{++v'}) + 2Kq_{vv'}]}.$$

Equality holds only if $q_{uv} = q_{uv'}$ for u neither v nor v' and $m_{++v} = m_{++v'}$. For example, in section 1.1, $U = K$ and $q_{uu'} = T/U$ for Forms u and u'. Thus $\mathbf{C}$ is $(T/U)\mathbf{I}$, and

$\sigma^2(\hat{D}_v - \hat{D}_{v'}) = 2\sigma^2 U/N$. In general, if no form can be used more than Q times, so that $m_{++u} \leq Q$ for $1 \leq u \leq U$, then $\sigma^2(\hat{D}_v - \hat{D}_{v'})$ is at least $2KT\sigma^2/(NQ)$.

For a more complex example often relevant to equating situations in which very old forms are not directly compared with new forms, consider the case of $\mathbf{y} = \mathbf{b}$ for $b_v = 1$, $b_{v'} = -1$, and $b_u = 0$ for u neither v nor v'. For simplicity, let $v' < v$. Let $\mathbf{x}$ be defined so that $x_u = u - (v + v')/2$ for $1 \leq u \leq U$. Then

$$\sigma^2(\hat{D}_v - \hat{D}_{v'}) \geq \frac{(v - v')^2\sigma^2}{M\mathbf{x}'\mathbf{C}\mathbf{x}},$$

where

$$\begin{aligned} \mathbf{x}'\mathbf{C}\mathbf{x} &= \sum_{u=1}^{U}[u - (U+1)/2]^2 m_{++u} - \sum_{u=1}^{U}\sum_{u'=1}^{U}[u - (U+1)/2][u' - (U+1)/2]q_{uu'} \\ &= 2^{-1}\sum_{u=1}^{U}\sum_{u'=1}^{U}(u - u')^2 q_{uu'}. \end{aligned}$$

Equality holds only if, for some real c,

$$1 = c\sum_{u'=1}^{U}(v - u')q_{vu'},$$

$$0 = c\sum_{u'=1}^{U}(u - u')q_{uu'}$$

for u neither v nor v', and

$$-1 = c\sum_{u'=1}^{U}(v' - u')q_{v'u'}.$$

When equality holds, $\sigma^2(\hat{D}_v - \hat{D}_{v'}) = c(v - v')\sigma^2/M$. In section 1, $K = 2$, $m_{++1} = m_{++U} = 1$, $m_{++u} = 2$ for $1 < u < U$, $T = U - 1$, $q_{uu'} = 1/2$ for $|u - u'| = 1$ and $q_{uu'} = 0$ for $|u - u'| > 1$. If $v = U$ and $v' = 1$, then equality holds with $c = 2$, so that

$$\sigma^2(\hat{D}_U) = \sigma^2(\hat{D}_U - \hat{D}_1) = 2(U-1)\sigma^2/M = 4T^2\sigma^2/N,$$

as expected from section 1 if one recalls that $M = N/(2T)$ in this case.

One may interpret $\mathbf{x}'\mathbf{C}\mathbf{x}$ for $x_u = u - (U+1)/2$ in terms of a variance. Consider random variables Z_1 and Z_2 with integer values from 1 to U. Let

$$W = KT - \sum_{u=1}^{U} q_{uu}.$$

Let Z_1 never equal Z_2, and let the joint probability that $Z_1 = u$ and $Z_2 = u'$ be $q_{uu'}/W$ for $u \neq u'$. Then

$$\mathbf{x}'\mathbf{C}\mathbf{x} = W\sigma^2(Z_1 - Z_2)/2,$$

so that

$$\sigma^2(\hat{D}_v - \hat{D}_{v'}) \geq \frac{2(v - v')^2\sigma^2}{WM\sigma^2(Z_1 - Z_2)}. \tag{9}$$

If the restriction is imposed that, for some positive integer r, Forms u and u' never appear in the same administration if $|u - u'| > r$, then $|Z_1 - Z_2| \leq r$ with probability 1, so that

$$\sigma^2(Z_1 - Z_2) \leq r^2,$$

with equality if, and only if, $r = 1$ (Haberman, 1996, p. 272), so that

$$\sigma^2(\hat{D}_v - \hat{D}_{v'}) \geq \frac{2(v - v')^2\sigma^2}{WMr^2}. \tag{10}$$

In particular,

$$\sigma^2(\hat{D}_U) \geq \frac{2(U - 1)^2\sigma^2}{WMr^2}, \tag{11}$$

In (11), equality holds for $r = 1$ under the conditions in section 1. In addition, (7) and the standard formula

$$U^{-1}\sum_{u=1}^{U}[u - (U + 1)/2]^2 = (U + 1)(U - 1)/12$$

(Stuart, 1950) leads to the inequality

$$\bar{\sigma}^2 \geq \frac{(U + 1)U\sigma^2}{3WMr^2}. \tag{12}$$

Observe that $W \leq KT$, so that WM is no greater than N. If no form is used with more than a single group in an administration, then $WM = (K - 1)N/K$. The practical implication of (10), (11), and (12) is that, for fixed sample size N, forms K per administration, and positive integer r, the variance of equating adjustments increases very rapidly when the number of forms is large.

To illustrate results, consider a case with 11 administrations and four forms per administration. Consider a total N of 110,00 examinees, and let the standard deviation σ be 100. At Administration t, let Forms t to $t + 3$ be used. The standard deviations of the $\hat{D}_u$ are then summarized in Table 1. For comparison, results are supplied for the approach of section 1 with the same number of examinees and the same number of administrations. The only case in which a form has a lower standard error of equating for two rather than four forms per administration is

Table 1
Standard Error of Equating Adjustment for Mean Equating

Form number	Standard error 2 forms	4 forms
2	2.00	2.57
3	2.83	2.51
4	3.46	2.50
5	4.00	2.72
6	4.47	2.85
7	4.90	2.99
8	5.29	3.12
9	5.66	3.25
10	6.00	3.37
11	6.32	3.49
12	6.63	3.62
13		3.80
14		4.18

for Form 2, and Form 2 is used by 10,000 examinees if there are two forms per administration and by 5,000 examinees if there are four forms per administration. The gains are particularly dramatic in the case of four forms per administration for higher form numbers. Nonetheless, it should be emphasized that the standard errors of the $\hat{D}_u$ still increase as the form number u increases. In the case of four forms, the lower bound for $\sigma(\hat{D}_U)$ from (11) is 2.13, a figure considerably lower than the actual value. The lower bound on $\sigma(\hat{D}_U)$ based on (9) is 3.51. In the case of two forms per administration, the lower bound on $\sigma(\hat{D}_U)$ based on (11) is equal to the observed value.

Several caveats are needed concerning results in this section. In practice, due to the need to report test scores in a timely manner, scores D_u typically must be estimated at Administration t by use of the means $\bar{X}_{kt'}$ for Group k used in Administration t' for $t' \leq t$. Bounds here permit use of all means $\bar{X}_{kt}$ for Group k in an Administration t, $t \leq T$.

Results apply most effectively in an ideal situation in which the score distribution for the X_{kt} is normal with mean μ_{kt} and variance σ^2 common to all forms and administrations. This assumption obviously does not apply exactly in commonly used educational tests.

Interactions between group and administration can greatly increase variability. Consider the following model change. For Group k at Administration t, let e_{kt} be a random variable with mean 0 and variance σ_e^2 that represents a random interaction of group and administration. In

many typical cases, each group receives a different form, and e_{kt} is really an interaction of form and administration. Let the e_{kt} be independent, and let the $X_{ikt} - e_{kt}$ be independent random variables with mean $\mu_{kt} = \alpha_t - D_{u_{kt}}$ and variance σ^2. Then variances of the $\hat{D}_u$ and $\hat{D}_u - \hat{D}_{u'}$, $u \neq u'$, are multiplied by $1 + M\sigma_e^2/\sigma^2$. The practical effect of interaction is very large, for it becomes an increasingly large fraction of the variability in $\hat{D}_u$ as the sample size M for a form used at an administration becomes increasingly large.

Rather remarkably, the analysis for mean equating for a series of administrations with equivalent groups provides a general basis for discussion of equating errors. The following sections consider some of the many applications to other equating designs and other equating methods.

2 Multiple Tests per Examinee

In many equating designs, operational tests at different administrations are compared through internal or external anchor tests. Such equating designs can be described in terms of examinees who receive multiple tests. As in Section 1.2, consider a case with $T \geq 1$ administrations, K groups per administration, N examinees, and $M = N/(KT)$ examinees per group and administration. Let each examinee receive $H \geq 1$ different tests h, $1 \leq h \leq H$. For example, one might have $H = 2$ and have Test 1 be an operational test and Test 2 be an external anchor test. Other alternatives are possible. Test 2 might be an internal anchor test rather than an external anchor test. One might also have an operational test with H sections, with a score provided for each section.

Whatever the interpretation of the tests, different forms are associated with different tests. For this purpose, test forms will be described by pairs of integers. Thus Test h can use Form (u, h) for $1 \leq u \leq U_h$, where U_h is a positive integer. At Administration t, for Test h, Group k receives Form (u_{kth}, h), $1 \leq u_{kth} \leq U_h$. For simplicity, let $u_{11h} = 1$ for $1 \leq h \leq H$. The base form for Test h will be Form $(1, h)$.

For a relatively simple example, consider a testing program in which Test 1 is an operational test and Test 2 is an external anchor test. Let all examinees at Administration t receive the same operational Form $(t, 1)$, so that $U_1 = T$. On the other hand, as in Section 1, let examinees in Administration t be divided randomly into $K = 2$ groups of equal size $M = N/(2T)$. Let Group k, $1 \leq k \leq 2$, receive Form $(t + k - 1, 2)$. In this case, $u_{kt1} = t$ and $u_{kt2} = t + k - 1$.

Let Examinee i, $1 \leq i \leq M$, in Group k of Administration t have score X_{itkh} on Form (u_{kth}, h). Let the score vectors $\mathbf{X}_{ikt}$ with elements X_{ikth}, $1 \leq h \leq H$, be independent. Let $\mathbf{X}_{ikt}$ have mean

$\boldsymbol{\mu}_{kt}$ with elements μ_{kth}, $1 \leq h \leq H$, and positive-definite covariance matrix $\boldsymbol{\Gamma}$ with elements $\gamma_{hh'}$, $1 \leq h \leq H$, $1 \leq h \leq H$. For simplicity, let the covariance matrix $\boldsymbol{\Gamma}$ be known. As in Section 1.2, an additive model for the means μ_{kth} is employed. For each Test h, for some form parameters D_{uh}, $1 \leq u \leq U_h$ and administration parameters α_{th}, $1 \leq t \leq T$, it is assumed that

$$\mu_{kth} = \alpha_{th} - D_{u_{kth}h} \tag{13}$$

for $1 \leq k \leq K$ and $1 \leq t \leq T$. To identify parameters, it is assumed that $D_{1h} = 0$ for $1 \leq h \leq H$, so that the adminstration parameter for Administration 1 and Test h is $\alpha_{1h} = \mu_{11h}$. The difference $\beta_{th} = \alpha_{th} - \alpha_{1h}$ provides a measure of the proficiency on Test h of examinees at Administration t relative to examinees at Administration 1, while D_{uh} measures the difficulty of Form (u, h) relative to Form $(1, h)$. If the D_{uh} are known, then conversions are easily accomplished. A score x on Test h on Form (u, h) is converted to a score $x + D_{uh}$ on Form $(1, h)$. The simplifying assumption is made that the differences β_{th} are proportional in the sense that

$$\beta_{th} = \nu_h \beta_{1h}, \tag{14}$$

where the ν_h are known constants and $\nu_1 = 1$. The vector $\boldsymbol{\nu}$ has elements ν_h for $1 \leq h \leq H$. It is often the case that $\nu_h = (\gamma_{hh}/\gamma_{11})^{1/2}$, so that the differences β_{th} are proportional to the standard deviations of the X_{ikth}.

Estimation of parameters is typically somewhat more complex than in Section 1.2, although many equating designs lead to simplified computations. Under the assumption that the covariance matrix $\boldsymbol{\Gamma}$ is known, all remaining model parameters can be estimated by weighted least squares, but analysis is a bit more complicated in general than in Section 1.2. Linkage involves both forms administered to the same examinee and forms administered to different examinees in the same or different administrations. For example, in the example with an operational test and an external anchor in which the operational test is different for each administration, the operational tests are linked only through the external anchors.

To describe the general use of weighted least squares, for Group k, Test h, Form (u, h), and Administration t, let m_{ktuh} be 1 if $u_{kth} = u$ and 0 otherwise, and let $\mathbf{m}_{ktuh}$ be the vector with elements $m_{ktuh'}\delta_{hh'}$ for $1 \leq h' \leq H$. Let $m_{+tuh} = \sum_{k=1}^{K} m_{ktuh}$, let $\psi = \boldsymbol{\nu}'\boldsymbol{\Gamma}^{-1}\boldsymbol{\nu}$, and let

$$B_{uhu'h'} = \mathbf{m}'_{ktuh}\boldsymbol{\Gamma}^{-1}\mathbf{m}_{ktu'h'} - \psi^{-1}(\mathbf{m}'_{ktuh}\boldsymbol{\Gamma}^{-1}\boldsymbol{\nu})(\mathbf{m}'_{ktu'h'}\boldsymbol{\Gamma}^{-1}\boldsymbol{\nu})$$

and

$$C_{uhu'h'} = B_{uhu'h'} + \frac{(K-1)T}{U_h(U_h-1)}\delta_{hh'}.$$

Assume that the identifiability condition holds that

$$\sum_{h'=1}^{H}\sum_{u'=1}^{U_{h'}} C_{uhu'h'}x_{u'h'} = 0$$

for $1 \leq u \leq U_h$ and $1 \leq h \leq H$ only if $x_{uh} = 0$ for $1 \leq u \leq U_h$ and $1 \leq h \leq H$. Let $\bar{\mathbf{X}}_{kt}$ be the average of the score vectors $\mathbf{X}_{ikt}$ for $1 \leq i \leq M.$ and let $\bar{\mathbf{X}}_{+t}$ be the sum of the $\bar{\mathbf{X}}_{kt}$ for $1 \leq k \leq K$. One then minimizes the weighted sum of squares

$$\sum_{t=1}^{T}\sum_{k=1}^{K}(\bar{\mathbf{X}}_{kt} - \boldsymbol{\mu}_{kt})'\boldsymbol{\Gamma}^{-1}(\bar{\mathbf{X}}_{kt} - \boldsymbol{\mu}_{kt})$$

under the constraint that (13) and (14) hold and $D_{1h} = 0$ for $1 \leq h \leq H$. A similar argument to that in Section 1.2 shows that the weighted least squares estimates $\hat{D}_{uh}$ of D_{uh} satisfy the equations

$$\sum_{h'=1}^{H}\sum_{u'=1}^{U_{h'}} B_{uhu'h'}\hat{D}_{u'h'} = -\sum_{t=1}^{T}\sum_{k=1}^{K}\mathbf{m}_{ktuh}\boldsymbol{\Gamma}^{-1}[\bar{\mathbf{X}}_{+kt} - \psi^{-1}(\bar{\mathbf{X}}'_{+kt}\boldsymbol{\Gamma}^{-1}\boldsymbol{\nu})\boldsymbol{\nu}]$$

for $1 \leq u \leq U_h$ and $1 \leq h \leq H$, where $\hat{D}_{1h} = 0$ for $1 \leq h \leq H$. A score x on Form (u, h) is then converted to score $x + \hat{D}_{uh}$ on Form $(1, h)$. Variances can be computed as in weighted linear regression.

A covariance matrix for the $\hat{D}_{uh}$ may be computed as in section 1.2, although results are a bit more complex. To facilitate use of matrices, consider the index variables $\pi(u, 1) = u$ for $1 \leq u \leq U_1$ and $\pi(u, h) = \pi(U_h, h-1) + u$ for $1 \leq u \leq U_h$ and $1 < h \leq H$. Let $U = \pi(U_H, H)$ be the sum of the U_h for $1 \leq j \leq H$. If

$$\tilde{D}_{uh} = \hat{D}_{uh} - U_h^{-1}\sum_{u'=1}^{U_h}\hat{D}_{u'h},$$

then the covariance $C^{uhu'h'}$ of $\tilde{D}_{uh}$ and $\tilde{D}_{u'h'}$, $1 \leq u \leq U_h$, $1 \leq h \leq H$, $1 \leq u' \leq U_{h'}$, $1 \leq h' \leq H$, is row $\pi(u, h)$ and column $\pi(u', h')$ of the inverse of the U by U matrix $\mathbf{C}$ with row $\pi(u, h)$ and column $\pi(u', h')$ equal to $C_{uhu'h'}$.

In typical applications which involve anchor tests, actual computations are much simpler than for the general case. Consider the following situation. There are two tests per examinee, so

that $H = 2$. Test 1 is an operational test and Test 2 is an anchor test. The m_{+tu2}, $1 \le t \le T$, $1 \le u \le U_2$, for Test 2 satisfy the inseparability requirement for a single test. In the case of Test 1, only one form is used in an administration, and all examinees for the examination use that form. Thus $m_{ktt1} = 1$ for $1 \le k \le K$, $1 \le t \le T$, $U_1 = T$, and $m_{ktu1} = 0$ if $1 \le k \le K$, $1 \le t \le T$, $1 \le u \le U$, and $t \ne u$. Thus the m_{ktu1} do not satisfy the inseparability conditions. Nonetheless, for each Group k and Administration t, $\bar{X}_{kt1}$ and $\bar{X}_{kt2}$ have correlation $\gamma_{12}/(\gamma_{11}\gamma_{22})$, so that estimation of D_{u2} for $1 \le u \le U_2$ is affected to some extent by the operational test results $\bar{X}_{kt1}$. The estimates $\hat{D}_{u2}$ may be obtained as in section 1.2 from the observed differences $X_{ikt2} - (\gamma_{12}/\gamma_{11})X_{ikt1}$, $1 \le i \le M$, $1 \le k \le K$, $1 \le t \le T$. In the computations leading to (4), U is replaced by U_2, m_{ktu} is replaced by m_{ktu2}, $\hat{D}_u$ is replaced by $\hat{D}_{u2}$, and $\bar{X}_{kt}$ is replaced by $\bar{X}_{kt2} - (\gamma_{12}/\gamma_{11})\bar{X}_{kt1}$. After some algebraic manipulation, one finds that the estimate $\hat{D}_{t1}$ of D_{t1} is

$$\hat{D}_{t1} = K^{-1}\left[\nu_2^{-1}(\bar{X}_{+t2} - \bar{X}_{+12}) - (\bar{X}_{+t1} - \bar{X}_{+112}) + \sum_{u=1}^{U_2}(m_{+tu2} - m_{+1u2})\hat{D}_{u2}\right].$$

The expectation of the equating adjustment $\hat{D}_{t1}$ for Test 1 at Administration t (Form $(t,1)$) is D_{t1}. To find the variance of $\hat{D}_{t1}$, let $\mathbf{C}_2$ be the U_2 by U_2 matrix with row u and column u' equal to

$$C_{uu'2} = m_{++u2}\delta_{uu'} - q_{uu'2} + \frac{(K-1)T}{U_2(U_2-1)},$$

where m_{++u2} is the sum of the m_{+th2} for $1 \le t \le T$ and $q_{uu'2} = K^{-1}\sum_{t=1}^{T} m_{+tu2}m_{+tu'2}$. Let the inverse $\mathbf{C}_2^{-1}$ of $\mathbf{C}_2$ have row u and column u' equal to $C_2^{uu'}$. Then

$$\begin{aligned}\sigma^2(\hat{D}_{t1}) &= \frac{2T\gamma_{11}(\nu_2 - \gamma_{12}/\gamma_{11})^2}{N\nu_2^2} \\ &\quad + \frac{T(\gamma_{22} - \gamma_{12}^2/\gamma_{11})}{N\nu_2^2}\left[2 + K^{-1}\sum_{u=1}^{U_2}\sum_{u'=1}^{U_2}(m_{+tu2} - m_{+1u2})(m_{+tu'2} - m_{+1u'2})C_2^{uu'}\right].\end{aligned}$$

For comparison of Administrations t and t', $t \ne t'$, note that $\hat{D}_{t1} - \hat{D}_{t'1}$ has mean $D_{t1} - D_{t'1}$ and variance

$$\begin{aligned}\sigma^2(\hat{D}_{t1} - \hat{D}_{t'1}) &= \frac{2T\gamma_{11}(\nu_2 - \gamma_{12}/\gamma_{11})^2}{N\nu_2^2} \\ &\quad + \frac{T(\gamma_{22} - \gamma_{12}^2/\gamma_{11})}{N\nu_2^2}\left[2 + K^{-1}\sum_{u=1}^{U_2}\sum_{u'=1}^{U_2}(m_{+tu2} - m_{+t'u2})(m_{+tu'2} - m_{+t'u'2})C_2^{uu'}\right].\end{aligned}$$

For fixed sample size N, an increase in the number T of administrations obviously leads to increased variance; however, in typical situations with a large number U_2 of anchor forms, the most

serious problem involves the contribution to variance due to the large number of anchor forms rather than the large number of administrations. For example, as in Table 1, consider $T = 11$ administrations and $U_2 = 12$ anchor forms, where $K = 2$, $N = 110,000$, and Form $(t, 1)$ and Form $(t + 1, 1)$ are used in Administration t. Observe that $M = 5,000$. Let $\gamma_{11} = \gamma_{22} = 10,000$, let $\gamma_{12} = 7,500$, and let $\nu_2 = 1$. In this example,

$$\hat{D}_{u2} = \sum_{t=1}^{u-1}[(\bar{X}_{1t2} - \bar{X}_{2t2}) - 0.75(\bar{X}_{1t1} - \bar{X}_{2t1})]$$

if $2 \leq u \leq T + 1$ and

$$\hat{D}_{t1} = 2^{-1}[\bar{X}_{+t2} - \bar{X}_{+12} - \bar{X}_{+t1} + \bar{X}_{+12} + (\hat{D}_{t2} + \hat{D}_{(t+1)2} - \hat{D}_{22})].$$

It follows after some calculation that

$$\sigma^2(\hat{D}_{t1}) = 1 + 7[1 + 4(t - 2)]/32.$$

For example, $\sigma(\hat{D}_{T1}) = 3.02$ is much larger than $\sigma(\hat{D}_{21}) = 1.10$.

3 Linear Equating

In linear equating, both means and standard deviations are employed. Linear equating is most appropriate for observed scores with normal distributions. Consider the following variation on the model in Section 1.2. At each Administration t, $1 \leq t \leq T$, examinees are divided into $K \geq 2$ groups of M examinees, so that there are a total of $N = KTM$ examinees in the T administrations. Forms 1 to U are to be linked, where $U \geq 2$, and Group k receives Form u_{kt} at Administration t. There are $K \geq 2$ distinct forms used. The raw score X_{ikt} of Examinee i from Group k at Administration t is a random variable with mean μ_{kt} and variance σ^2_{kt}, and the reliability coefficient is ρ^2 for Form u_{kt} and Administration t. The X_{ikt} are assumed to be mutually independent. If $u_{kt} = u$, $1 \leq u \leq U$, then $m_{ktu} = 1$. Otherwise, $m_{ktu} = 0$. The definitions of the sums m_{+kt} and m_{++u} are then as in Section 1.2. For some real α_t, $1 \leq t \leq T$, D_u, $1 \leq u \leq U$, $\tau_t > 0$, $1 \leq t \leq T$, and $\zeta_u > 0$, $1 \leq u \leq U$, it is assumed that

$$\mu_{kt} = (\alpha_t - D_{u_{kt}})/\zeta_{u_{kt}}$$

and

$$\sigma_{kt} = \tau_t/\zeta_{u_{kt}}.$$

To identify parameters, it is assumed that $\zeta_1 = 1$ and $D_1 = 0$. For convenience, it is assumed that $u_{11} = 1$, so that α_1 is the mean score at Administration 1 on Form 1 and $\sigma_{11} = \tau_1$ is the corresponding standard deviation. Observe that for Forms u and u', in any Administration t such that, for Groups k and k', $u_{kt} = u$ and $u_{k't} = u'$, then

$$\sigma_{kt}/\sigma_{k't} = \zeta'_u/\zeta_u \tag{15}$$

and

$$\zeta_u\mu_{kt} + D_u = \zeta_{u'}\mu_{k't} + D_{u'}. \tag{16}$$

In linear equating, a score of x on Form u is converted to a score of $e_u(x) = \zeta_u x + D_u$ on Form 1. Thus linear equating reduces to mean equating if all ζ_u are equal to 1. More generally, this conversion rule implies that a score of x on Form u corresponds to a score of $\zeta_{u'}^{-1}(\zeta_u x + D_u - D_{u'})$ on Form u'. This conversion is consistent with (15) and (16). These equation correspond with customary requirements for chained equating.

If the X_{ikt} are normally distributed and if the inseparability requirement of Section 1.2 is satisfied, then the α_t, ζ_u, D_u, and τ_t may be estimated by use of maximum likelihood. Let hats be used to denote maximum-likelihood estimates, so that $\hat{\alpha}_t$ is the maximum-likelihood estimate of α_t, $\hat{\zeta}_u$ is the maximum-likelihood estimate of ζ_u, $\hat{D}_u$ is the maximum-likelihood estimate of D_u, $\hat{\tau}_t$ is the maximum-likelihood estimate of τ_t, and $\hat{e}_u(x)$ is the maximum-likelihood estimate of $e_u(x)$. Standard large-sample approximations for maximum-likelihood estimates can be applied with little complication to provide normal approximations for all maximum-likelihood estimates of interest under the condition that M becomes large. Although results simplify somewhat because $X_{ikt} - \mu_{ikt}$ is uncorrelated with $(X_{ikt} - \mu_{ikt})^2$ under the normality assumption, the asymptotic variances and covariances of parameter estimates are somewhat more complex than in mean equating except in special cases. Normal approximations can be expressed in terms of a regression model. Let $\chi_t = \log \tau_t$, $\hat{\chi}_t = \log \hat{\tau}_t$, $\omega_u = \log \zeta_u$, and $\hat{\omega}_u = \log \hat{\zeta}_u$. Let $\psi_{kt} = (\alpha_t - D_{u_{kt}})/\tau_t$ for $1 \le k \le K$ and $1 \le t \le T$. Consider a hypothetical linear regression model in which

$$Y_{kt} - 2^{-1/2}(\chi_t - \omega_{u_{kt}})$$

and

$$Z_{kt} - \tau_t^{-1}(\alpha_t - D_{u_{kt}}) - \psi_{kt}\omega_{u_{kt}}$$

are independent normal random variables with common mean 0 and variance M^{-1} for $1 \leq k \leq K$ and $1 \leq t \leq T$. In this model, χ_t, ω_u, α_t, and D_u are treated as unknown parameters to be estimated, while τ_t and ψ_{kt} are treated as known. (The relationship of τ_t to χ_t and the relationship of ψ_{kt} to χ_t, α_t, and $D_{u_{kt}}$ is ignored.) The restrictions are imposed that $D_1 = \omega_1 = 0$. Under the inseparability assumption, the least-squares estimates χ_t^* of χ_t, ω_u^* of ω_u, α_t^* of α_t, and D_u^* of D_u are uniquely defined, unbiased, and normal distributed with variances and covariances readily found as in standard regression analysis. The joint distribution of the estimates $\hat{\alpha}_t$, $1 \leq t \leq t$, $\hat{D}_u$, $1 \leq u \leq U$, $\hat{\chi}_t$, $1 \leq t \leq T$, and $\hat{\omega}_u$, $1 \leq u \leq U$, is approximately the same as the joint distribution of the hypothetical estimates α_t^*, $1 \leq t \leq t$, D_u^*, $1 \leq u \leq U$, χ_t^*, $1 \leq t \leq T$, and ω_u^*, $1 \leq u \leq U$. For fixed number K of groups per administration and fixed number T of administrations, the approximation is increasingly accurate as the sample size M per group within administration becomes increasingly large. The estimate $\hat{\zeta}_u$ is approximately distributed as $\zeta_u(1 + \omega_u^*)$, so that $\hat{e}_u(x)$ is approximately distributed as $\zeta_u(1 + \omega_u^*) + D_u^*$.

If the model for mean equating holds, then $\tau_t = \tau_1$, $\chi_t = \log \tau_1$, $\omega_u = 0$, $\psi_{kt} = (\alpha_t - D_{u_{kt}})/\tau_1$, and $\zeta_u = 1$, so that $e_u(x) = x + D_u$. In this case, linear equating leads to less satisfactory results than does mean equating. The basic argument involves a general observation concerning regression analysis. Consider a linear regression model of the form $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}$, where $\mathbf{Y}$ is a random vector with n elements, $\mathbf{X}$ is a fixed n by p matrix of rank p for some positive integer p, $\boldsymbol{\beta}$ is an unknown fixed vector with p elements, and $\mathbf{e}$ is a random vector with n independent elements, each of which has mean 0 and variance $\sigma^2 > 0$. As is well known, $\boldsymbol{\beta}$ has least-squares estimate $\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$ with mean $\boldsymbol{\beta}$ and covariance matrix $\sigma^2(\mathbf{X}'\mathbf{X})^{-1}$. On the other hand, if for some positive integer $q < p$, $\beta_j = 0$ for $q < j \leq p$, then one can consider the use of least squares subject to the restriction that $\beta_j = 0$ for $q < j \leq p$. In this case, a new least-squares estimate $\mathbf{b}^*$ is obtained. The elements $b_j^* = 0$ of $\mathbf{b}$ are 0 for $q < j \leq p$. If $\mathbf{Z}$ is the n by q matrix formed from the first q columns of $\mathbf{X}$, and if $\mathbf{b}^-$ is the q-dimensional vector with elements b_j^* for $1 \leq j \leq q$, then $\mathbf{b}^- = (\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{Y}$. If $\mathbf{x}$ is a p-dimensional vector with elements x_j for $1 \leq j \leq p$, some x_j is not 0, and $\mathbf{z}$ is a q-dimensional vector with elements x_j for $1 \leq j \leq q$, then $\mathbf{x}'\mathbf{b}$ has variance $\sigma^2\mathbf{x}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}$, while $\mathbf{x}'\mathbf{b}^* = \mathbf{z}'\mathbf{b}^-$ has variance $\sigma^2\mathbf{z}'(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{z}$. By the Gauss-Markov theorem (Rao, 1973, ch. 4),

$$\mathbf{x}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x} > \mathbf{z}'(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{z}.$$

Both $\mathbf{x}'\mathbf{b}$ and $\mathbf{x}'\mathbf{b}^*$ have mean $\mathbf{x}'\boldsymbol{\beta}$. These results apply to linear equating by consideration of

the case in which ω_u is assumed 0 and χ_t is assumed constant. It follows that the approximate variance $\sigma^2(\hat{e}_u(x))$ exceeds the variance of $\hat{D}_u$ from linear equating for each Form $u > 1$. A practical implication of the result is that cautions concerning equating of many forms that were developed under mean equating must also apply in the case of linear equating.

An added and more general lower bound for variances for normal approximations associated with linear equating can be obtained by consideration of the case of ω_u known. In this case, similar arguments to those for ω_u equal 0 show that variances are very similar to those associated with mean equating. The normal approximation for $\hat{e}_u(x)$ has a variance at least as large as the variance obtained for $\hat{D}_u$ in section 1.2 for σ^2 equal to the smallest value of τ_t^2, $1 \leq t \leq T$.

The linear equating arguments used here have a simple application to item response theory. Suppose that the X_{ikt} in the model for linear equating are latent variables with normal distributions, so that they correspond to conventional θ-parameters. Let each Form u have r dichotomous items, and let the observed response on Item j for Examinee i from Group k at Administration t be Y_{jikt} equal to 0 or 1. Let the Y_{jikt}, $1 \leq j \leq r$, $r \geq 3$, be conditionally independent given the X_{ikt}. Let the conditional probability that $Y_{jikt} = 1$ given $X_{ikt} = x$ be

$$\frac{\exp(\gamma_{jkt}x - \beta_{jkt})}{1 + \exp(\gamma_{jkt}x - \beta_{jkt})}$$

for some unknown constants $\gamma_{jkt} > 0$ and β_{jkt}. If the added restriction is imposed that $\tau_1 = 1$ and $\alpha_1 = 0$, then all γ_{jkt} and β_{jkt} can be estimated by marginal maximum likelihood, together with α_t, D_u, τ_t, and ζ_u. Normal approximations for maximum-likelihood estimates are readily derived, but results are relatively complicated. Nonetheless, a rather trivial lower bound can be obtained for the variances of normal approximations for the maximum-likelihood estimate $\hat{e}_u(x)$ of $e_u(x)$. The variance of the normal approximation for $\hat{e}_u(x)$ for the item response model is at least as great as the variance of the normal approximation for $\hat{e}_u(x)$ which is obtained under the ordinary case of linear equating in which the X_{ikt} are directly observed (Sundberg, 1974).

The arguments just used also apply if each Group k at Administration t has a distinct form but nonempty subsets V_{kt} of the integers 1 to r exist for each Group k and Administration t such that, if $u_{kt} = u_{k't'}$, then $V_{kt} = V_{k't'}$ and, for j in V_{kt}, $\gamma_{jkt} = \gamma_{jk't'}$ and $\gamma_{jkt} = \gamma_{jk't'}$. Assume that X_{ikt} has mean $\alpha_t - D_{u_{kt}}$ and standard deviation $\tau_t/\zeta_{u_{kt}}$, and retain the assumption that $\tau_1 = 1$ and $\alpha_1 = 0$. Then a value x for X_{ikt} for Group k and Administration t is adjusted to $e_u(x) = \zeta_u x + D_u$ for Group 1 at Administration 1 if $u_{kt} = u$. Lower bounds of variances for

normal approximations for maximum-likelihood estimate of $e_u(x)$ are found for the case of linear equating in which the X_{ikt} are known.

4 Conclusion

The analysis provided has some strong implications in practice. Accuracy of equating methods is limited by sample size. For a given total sample N distributed over T administrations, limits on accuracy involve the number U of distinct forms employed. As the number U increases, the accuracy of equating results decreases. The decrease is especially severe if limits are placed on how long an older form can remain in use. The implications are important for programs in which a very large number of forms is used due to a very high frequency of administration and due to security concerns that limit reuse of forms. Under the assumption that the number of examinees per year is not materially affected by the frequency of administration, it is reasonable to expect that accuracy of equating will be much lower than in programs with comparable yearly volume in which few test forms are administered in a given year. As a consequence, the comparability of scores on different examinations may be compromised. Such an outcome can arise even if equating procedures perform perfectly and the only complication is sampling error. In the real world, in which equating procedures are not perfect, results can be substantially less satisfactory.

Mitigation of the problems of equating error involves careful data collection; however, even the most careful data collection will have limitations if form reuse is severely restricted and the number of forms is very large. It is important to consider the number of forms which is sufficient so that inappropriate study of past forms has no realistic possibility of affecting an examinee score due to the limitations of human memory and due to the labor involved in such study. If the number of forms produced is sufficiently limited, then so is the problem of equating error.

If reuse is not an option, then it may be necessary periodically to restart equating procedures with newer base forms. Such a procedure may be tolerable in cases in which test results can only be used for a limited period, say two years.

Because of computer-based testing, the problem of frequent administration is likely to be a continuing issue. It is certainly advisable that new testing programs consider the implication of linking large numbers of forms prior to their first administration rather than afterwards.

References

Cochran, W. G., & Cox, G. M. (1957). *Experimental designs* (2nd ed.). New York, NY: John Wiley.

Goodman, L. A. (1968). The analysis of cross-classified data: independence, quasi-independence, and interactions in contingency tables with or without missing entries. *Journal of the American Statistical Association*, *63*, 1091–1131.

Haberman, S. J. (1996). *Advanced statistics. Volume I: Description of populations*. New York, NY: Springer.

Halmos, P. R. (1958). *Finite-dimensional vector spaces* (2nd ed.). Princeton, NJ: Van Nostrand.

Hardy, G., Littlewood, J. E., & Pòlya, G. (1952). *Inequalities* (2nd ed.). Cambridge, Enlgand: Cambridge University Press.

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

Rao, C. R. (1973). *Linear statistical inference and its applications* (2nd ed.). New York, NY: John Wiley.

Scheffé, H. (1959). *The analysis of variance*. New York, NY: John Wiley.

Stuart, A. (1950). The cumulants of the first n natural numbers. *Biometrika*, *37*, 446.

Sundberg, R. (1974). Maximum likelihood theory for incomplete data from an exponential family. *Scandinavian Journal of Statistics*, *1*, 49–58.

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