ACT Research Report Series 2004-5

IRT Scale Linking Methods for Mixed-Format Tests

Seonghoon Kim

Won-Chan Lee

ACT

December 2004

For additional copies write:
ACT Research Report Series
P.O. Box 168
Iowa City, Iowa 52243-0168

IRT Scale Linking Methods for Mixed-Format Tests

Seonghoon Kim
The University of Iowa

Won-Chan Lee
ACT

Contents

Abstract

Under item response theory (IRT), obtaining a common proficiency scale is required in many applications. Four IRT linking methods, including the mean/mean, mean/sigma, Haebara, and Stocking-Lord methods, have been developed and widely used to estimate linking coefficients (slope and intercept) for a linear transformation from one scale to another. These four methods have typically been used for dichotomous IRT models but can also be extended to polytomous IRT models. This paper further extends the four linking methods to a mixture of unidimensional IRT models for mixed-format tests. The development in the present study is intended to be as general as possible so that each linking method can be applied to mixed-format tests using any mixture of the following five IRT models: the three-parameter logistic model, the graded response model, the generalized partial credit model, the nominal response model, and the multiple-choice model.

A simulation study is conducted to investigate the performance of the four linking methods extended to mixed-format tests. Overall, the Haebara and Stocking-Lord methods yield more accurate linking results than the mean/mean and mean/sigma methods. The simultaneous linking using all items with different formats is compared to the linking through items of a "dominant" item format. When the nominal response model or the multiple-choice model is used to analyze data from mixed-format tests, limitations of the mean/mean, mean/sigma, and Stocking-Lord methods are described.

IRT Scale Linking Methods for Mixed-Format Tests[1]

Introduction

A test containing a mixture of different item formats is often used in both classroom and large-scale assessments. In the present study, such a test is referred to as a *mixed-format test*. Combinations of different item formats often allow for the measurement of a broader set of skills than the use of a single format. The formats of items in a mixed-format test are usually categorized into two classes: multiple-choice (MC) and constructed-response (CR). Typically, MC items are dichotomously scored (DS) and CR items are polytomously scored (PS).

The use of item response theory (IRT) in testing applications has grown considerably over the last few decades. Some of the IRT models that have been thoroughly developed and used in practical testing programs include the three-parameter logistic (3PL) model (Birnbaum, 1968), the graded response (GR) model (Samejima, 1969, 1972), the nominal response (NR) model (Bock, 1972), the generalized partial credit (GPC) model (Muraki, 1992), and the multiple-choice (MC) model (Thissen & Steinberg, 1984). In practice, these models are applicable to various formats of items on a mixed-format test.

In IRT, the scale for measuring proficiency (as a construct) is determined up to an arbitrary linear transformation. Typically, this indeterminacy is solved in such a way that the mean and standard deviation of proficiency parameters are arbitrarily fixed to 0 and 1 ("0, 1" scale) for the group of examinees at hand. This implies that if two "0, 1" scales are obtained separately from different groups, then the two "0, 1" scales may be nonequivalent. Thus, to obtain a common scale, one scale should be linked to the other. In this case, the item and proficiency parameters on one scale are transformed to the other by a linear function relating the two scales, a process referred to here as *scale linking*. Note that the linear transformation in scale linking is made possible by the invariance property of IRT modeling (Lord, 1980). In practice,

---

[1] A previous version of the paper was presented at the Annual Meeting of the National Council on Measurement in Education, April 2004, San Diego, CA.

linking coefficients (slope and intercept) of the linear function are estimated through appropriate scale linking methods.

Several linking methods, which attempt to minimize linking error, have been developed under dichotomous IRT models. The linking methods include the mean/mean (Loyd & Hoover, 1980), mean/sigma (Marco, 1977), Haebara (Haebara, 1980), Stocking-Lord (Stocking & Lord, 1983), and minimum chi-square (Divgi, 1985) methods. These methods have also been extended to polytomous IRT models (Baker, 1992, 1993, 1997; Cohen & Kim, 1998; Kim & Cohen, 1995; Kim & Hanson, 2000, 2002). Recently, Ogasawara (2001) proposed the least squares methods for dichotomous IRT models and provided their asymptotic standard errors of linking coefficients. Kim and Song (2004) extended the least squares methods to the GR model.

Compared to other research areas, such as scoring and weighting, little research has been conducted on scale linking for mixed-format tests. Li, Lissitz, and Yang (1999) presented an extended version of the Stocking-Lord linking method for mixed-format tests consisting of DS and PS items, for which the 3PL and GPC models, respectively, were used. Tate (2000) described extended versions of the mean/sigma and Stocking-Lord linking methods for mixed-format tests with MC and CR items. In the study by Tate (2000), a dichotomous IRT model was assumed for the MC items and a modification of the GR model (see Tate, 1999) was applied to the CR items judged by raters with polytomous responses.

However, no study has been conducted on linking scales from mixed-format tests that require three or more unidimensional IRT models. (This case probably will be rare in real testing programs.) In this regard, the present paper presents a general framework for linking mixed-format tests that require more than two different unidimensional IRT models. In brief, the main purposes of the present study are: (1) to formally present four "traditional" linking methods including the mean/mean, mean/sigma, Haebara, and Stocking-Lord methods extended to mixed-format tests, and (2) to investigate the performance of each linking method under several simulation conditions.

Scale Linking Under a Mixture of IRT Models

It has been reported that some large-scale operational mixed-format tests are nearly unidimensional with respect to the constructs they measure, but others are not. Among examples of the former are the College Board's Advanced Placement computer science and chemistry examinations (Bennett, Rock, & Wang, 1991; Thissen, Wainer, & Wang, 1994) and the Wisconsin Student Assessment System mathematics and reading tests (Swygert, McLeod, & Thissen, 2001). Through a literature review for studies on the construct equivalence of MC and CR items, Traub (1993) concluded that the two types of items appear to measure different constructs for the writing domain, but *not* for the reading comprehension and quantitative domains.

When the constructs measured by different formats of items are claimed to be almost identical, unidimensional IRT models can be used to analyze the different item formats. For example, MC and CR items can be analyzed using the 3PL and GPC models, respectively. The item parameters for a mixed-format test can be estimated separately by format or simultaneously across formats. Separate calibration by format ("format-wise" calibration) could also be used in a multidimensional situation, where different formats of items appear to measure substantially different constructs. Simultaneous calibration not only provides IRT's answer to solving the problem of weight selection for each format in a statistically optimal way (Wainer & Thissen, 1993), but also provides a basis for calculating IRT scale scores based on patterns of summed scores (Rosa, Swygert, Nelson, & Thissen, 2001). It can be said that simultaneous calibration is more justifiable than format-wise calibration because different formats of items are not only calibrated at the same time but also placed on the same metric.

IRT Models for Subtests of a Mixed-Format Test

The probabilistic expression for each of the five IRT models (3PL, GR, GPC, NR, and MC) is briefly described for further discussion. The probability of a randomly selected examinee i with proficiency θ_i getting a score at category k of item j with K_j categories is symbolized by P_{ijk} or $P_{jk}(\theta_i)$, which is called a category response function or a category characteristic curve.

Three-parameter logistic model. Under the 3PL model (Birnbaum, 1968), the probability that a randomly selected examinee i with proficiency θ_i answers item j correctly is defined as

$$P_{ij} = P_j(\theta_i) = P(\theta_i \mid a_j, b_j, c_j) = c_j + (1 - c_j)\frac{\exp[Da_j(\theta_i - b_j)]}{1 + \exp[Da_j(\theta_i - b_j)]}, \tag{1}$$

where a_j is a discrimination parameter, b_j is a difficulty parameter, c_j is a lower asymptote, and D is a scaling constant (typically 1.7).

Graded response model. The GR model is appropriate to model PS items with ordered response categories such as in a Likert scale (Baker, 1992). Consider the logistic form in the homogeneous case of the model (Samejima, 1969, 1972), in which an item discrimination parameter is constant across all categories. Let $\tilde{P}_{ijk}$ denote the cumulative probability that a randomly selected examinee i with proficiency θ_i earns a score *at or above* category k of item j. Formally, the $\tilde{P}_{ijk}$ for category k of item j with K_j categories, given θ_i, can be expressed as

$$\tilde{P}_{ijk} = \tilde{P}_{jk}(\theta_i) = \tilde{P}(\theta_i \mid a_j, b_{jk}) = \begin{cases} 1 & k = 1 \\ \dfrac{\exp[Da_j(\theta_i - b_{jk})]}{1 + \exp[Da_j(\theta_i - b_{jk})]} & 2 \le k \le K_j, \\ 0 & k > K_j \end{cases} \tag{2}$$

where a_j is a discrimination parameter, b_{jk} are difficulty or location parameters for categories 2 through K_j, and D is a scaling constant (typically 1.7). Note that the first category does not have a difficulty parameter. Now, the category response function, P_{ijk}, is given by the difference between two adjacent cumulative probabilities as follows:

$$P_{ijk} = P_{jk}(\theta_i) = \tilde{P}_{ijk} - \tilde{P}_{ij(k+1)}. \tag{3}$$

Note in Equation 3 that if a category k is equal to 1, $P_{ij1} = 1 - \tilde{P}_{ij2}$ and if a category k is equal to K_j, $P_{ijK_j} = \tilde{P}_{ijK_j}$.

Generalized partial credit model. As for the GR model, the GPC model (Muraki, 1992) is appropriate for the analysis of responses that are successively ordered on a rating scale. The model states that the probability, P_{ijk}, is given by

$$P_{ijk} = P_{jk}(\theta_i) = P(\theta_i \mid a_j, b_{j1}, \cdots, b_{jk}, \cdots, b_{jK_j}) = \frac{\exp\left[\sum_{v=1}^{k} Da_j(\theta_i - b_{jv})\right]}{\sum_{h=1}^{K_j} \exp\left[\sum_{v=1}^{h} Da_j(\theta_i - b_{jv})\right]}, \tag{4}$$

where a_j is a discrimination parameter, b_{jk} is an item-category parameter, and D is a scaling constant (typically 1.7). Note that b_{j1} is arbitrarily defined as 0. This value is not a location constant and could be any value because the term including this parameter is canceled from the numerator and denominator of the model (Muraki, 1992).

Nominal response model. Bock's (1972) NR model can be used to model PS items whose categories are not necessarily ordered. For the NR model, the probability, P_{ijk}, is expressed as

$$P_{ijk} = P_{jk}(\theta_i) = P(\theta_i \mid a_{j1}, \cdots, a_{jK_j}, c_{j1}, \cdots, c_{jK_j}) = \frac{\exp(a_{jk}\theta_i + c_{jk})}{\sum_{h=1}^{K_j} \exp(a_{jh}\theta_i + c_{jh})}. \tag{5}$$

Since Equation 5 is invariant with respect to translation of the term $a_{jk}\theta_i + c_{jk}$ in both the numerator and denominator, sometimes two constraints are imposed for model identification:

$$\sum_{k=1}^{K_j} a_{jk} = 0 \text{ and } \sum_{k=1}^{K_j} c_{jk} = 0. \tag{6}$$

Multiple-choice model. Thissen and Steinberg's (1984) MC model can be viewed as an extended version of the NR model. The NR model has been widely used in choice/preference data. However, it might not be appropriate for multiple-choice items because as the latent proficiency approaches negative infinity, there will be one response for which the response function approaches one and the response functions associated with all other responses will approach zero. This is not consistent with the possibility that examinees with low proficiencies could choose any of the responses by guessing (Kim & Hanson, 2002). As an alternative approach to this limitation of the NR model, the probability P_{ijk} under the MC model is expressed as

$$P_{ijk} = P_{jk}(\theta_i) = P(\theta_i \mid a_{j0},\cdots,a_{jK_j},c_{j0},\cdots,c_{jK_j},d_{jk}) = \frac{\exp[a_{jk}\theta_i + c_{jk}] + d_{jk}\exp[a_{j0}\theta_i + c_{j0}]}{\sum_{h=0}^{K_j}\exp[a_{jh}\theta_i + c_{jh}]}, \quad (7)$$

where a_{j0} and c_{j0} are parameters for the "0" or "Don't Know" latent category of item j. As in the NR model, the parameters a_{jk} and c_{jk} are not identified with respect to location, so most often the following constraints are imposed on those parameters for model identification:

$$\sum_{k=0}^{K_j} a_{jk} = 0 \text{ and } \sum_{k=0}^{K_j} c_{jk} = 0. \quad (8)$$

The parameters represented by d_{jk} are proportions, representing the proportion of those who don't know that respond in each category on a multiple-choice item. Therefore, the constraint

$$\sum_{k=1}^{K_j} d_{jk} = 1$$

is required (Thissen & Steinberg, 1997).

Dichotomous IRT Models Revisited

It is helpful to view dichotomous IRT models—e.g., the 3PL model—from the perspective of polytomous IRT models to deal with both types of IRT models consistently in mixed-format tests. Dichotomous IRT models actually have two response categories—*incorrect* and *correct*. With the notation used in Equation 1, the probability for the incorrect response category is symbolized as $1 - P_{ij}$. However, in this study for the purpose of generalization in line with polytomous IRT models, the incorrect response category is regarded as the *first* category and the correct response category as the second category. Therefore, for item j, the probability for the incorrect response category can be symbolized as P_{ij1} and the probability for the correct response category as P_{ij2}. Note that the first category does not have a difficulty parameter (as with the GR model) and the difficulty parameter for the correct response category of item j can be symbolized as b_{j2}.

The Nature of Scale Linking Through Common Items

Consider a situation in which item parameters for a given set of items from a mixed-format test, all intended to measure a single proficiency, are independently estimated using item response data obtained from two groups of examinees, old and new. Suppose that the two separate calibrations use their respective "0, 1" scales to remove scale indeterminacy. The two "0, 1" scales are group dependent and are not expected to be equivalent unless the proficiency distributions for the two groups have the same mean and standard deviation. Denote the two "0, 1" scales from the old and new groups as θ_O (old scale) and θ_N (new scale), respectively. For the two "0, 1" scales to be compared and used interchangeably, they should be placed on a common scale. Although the common scale can be arbitrarily defined, usually one of the two "0, 1" scales is used as the common scale; in this study, the *old* scale is assumed to be the common scale.

Although the two "0, 1" scales, θ_O and θ_N, are group dependent, they should be linearly related because of the invariance property of IRT modeling (Lord, 1980)—as long as the model and assumptions hold—in such a way that

$$\theta_O = A\theta_N + B. \tag{9}$$

The slope A and intercept B of the linear function (or, linear transformation) are often referred to as (scale) linking coefficients. Given the relation $\theta_O = A\theta_N + B$, the item parameters from separate calibrations should be also linearly related as follows. Under the 3PL, GR, and GPC models (see, e.g., Baker, 1992; Lord, 1980),

$$a_{jO} = a_{jN} / A \tag{10}$$

and

$$b_{jkO} = A b_{jkN} + B, \tag{11}$$

where a_{jO} and b_{jkO} are item parameters for category k of item j expressed on the old scale θ_O, and a_{jN} and b_{jkN} are those on the new scale θ_N. Under the NR and MC models (see, e.g., Baker, 1993; Kim & Hanson, 2000, 2002),

$$a_{jkO} = a_{jkN} / A \tag{12}$$

and

$$c_{jkO} = c_{jkN} - (B / A)\, c_{jkN}. \tag{13}$$

The parameters, c_j under the 3PL model and d_{jk} under the MC model, do not depend on the latent proficiency and thus are not affected by the linear transformation. Thus,

$$c_{jO} = c_{jN} \tag{14}$$

and

$$d_{jkO} = d_{jkN}. \tag{15}$$

However, Equations 10 through 15 typically do not hold for estimated item parameters because of sampling error and possible model misfit. This implies that with sample data, linking coefficients A and B should be properly estimated so as to minimize linking error. Scale linking methods provide solutions to this kind of estimation.

This paper focuses on the four linking methods applicable to mixed-format tests: mean/mean, mean/sigma, Haebara, and Stocking-Lord methods. The mean/mean and mean/sigma methods are often called the *moment* methods, and the Haebara and Stocking-Lord methods are referred to as the *characteristic curve* methods (see Kolen & Brennan, 1995). The four methods have been widely used due to their simplicity (for the moment methods) or superiority (for the characteristic curve methods). It has been reported that the characteristic curve methods produce more stable results than the moment methods (Baker & Al-Karni, 1991; Hanson & Béguin, 2002; Kim & Cohen, 1992; Ogasawara, 2001).

To discuss the scale linking for mixed-format tests, the next sections assume that the constructs measured by different formats of items are similar enough that some or all of the five unidimensional IRT models just described can be used together to analyze response data from mixed-format tests. For ease of further discussion, it is also assumed that M different unidimensional IRT models are used to analyze item response data from a mixed-format test. The m^{th} ($m = 1, 2, \cdots, M$) single-format set of items (or, model) has J_m items, and the mixed-format test has a total of n items. That is, $n = \sum_{m=1}^{M} J_m$.

Moment Methods: Mean/Mean and Mean/Sigma Methods

The moment methods attempt to find the appropriate slope and intercept of the new-to-old transformation by expressing Equations 10 through 13 in terms of a group of items. A problem in doing so is that Equation 11 under the 3PL, GR, and GPC models does not agree in form with Equation 13 under the NR and MC models. The following reparameterization, as was done in Kim and Hanson (2000), is adopted to solve the problem:

$$a_{jk}\theta_i + c_{jk} = a_{jk}(\theta_i - b_{jk}), \tag{16}$$

where $b_{jk} = -c_{jk} / a_{jk}$. Equation 13, then, can be re-expressed in terms of b_{jk} as

$$b_{jkO} = Ab_{jkN} + B. \tag{17}$$

Notice that Equation 17 is identical in form to Equation 11.

Let $M(\cdot)$ and $SD(\cdot)$ be the operators for the two descriptive statistics, mean and standard deviation. Taking the mean over a-parameters based on Equations 10 and 12 and then expressing the resulting relationship with respect to A,

$$A = M(a_N) / M(a_O), \tag{18}$$

where a_N represents all the discrimination parameters, $a_{jN(m)}$ or $a_{jkN(m)}$ [$j = 1, 2, \cdots, J_m$; $k = 1, 2, \cdots, K_{j(m)}$; $m = 1, 2, \cdots, M$] on the new scale; and a_O is the counterpart on the old scale. Note that the notation (m) in the subscripts is used to indicate that item j is nested in the m^{th}

model. Taking the mean and standard deviation over b-parameters based on Equations 11 and 17 and then expressing the resulting relationships with respect to A and B,

$$A = SD(b_O) / SD(b_N) \tag{19}$$

and

$$B = M(b_O) - A \cdot M(b_N), \tag{20}$$

where b_N represents all the b-parameters (including both original and reparameterized ones) on the new scale, *except* nonexistent and arbitrary ones such as b_{j1N} under the 3PL, GR, and GPC models. A similar representation is applied to b_O, based on the old scale. Note that Equation 20 assumes that A has been evaluated.

Mean/mean method. From Equations 18 and 20, with sample data, linking coefficient estimates in the mean/mean method are obtained by

$$\hat{A}_{MM} = M(\hat{a}_N) / M(\hat{a}_O) \tag{21}$$

and

$$\hat{B}_{MM} = M(\hat{b}_O) - \hat{A}_{MM} M(\hat{b}_N). \tag{22}$$

Note that item parameters are replaced with their estimates.

Mean/sigma method. From Equations 19 and 20, the mean/sigma method estimates linking coefficients by

$$\hat{A}_{MS} = SD(\hat{b}_O) / SD(\hat{b}_N) \tag{23}$$

and

$$\hat{B}_{MS} = M(\hat{b}_O) - \hat{A}_{MS} M(\hat{b}_N). \tag{24}$$

Several issues need to be discussed in regard to the moment methods extended to mixed-format tests. First, the moment methods presented above do not specify whether the scaling constant, D, should be used consistently across the models; nor do they specify whether any weight should be given to the estimates of a_{jk} or b_{jk}. The reparameterization, $a_{jk}(\theta_i - b_{jk})$, for the NR and MC models could possibly be replaced with $Da^{**}_{jk}(\theta_i - b_{jk})$, where $a^{**}_{jk} = a_{jk} / D$, and thus a_{jk} could be replaced with a^{**}_{jk}. Furthermore, compared to the estimates of a_{jk} for the NR and MC models, the estimate of a_j for item j with K_j categories from, for example, the GR model, could be given some weight, for example, $K_j - 1$. Although these considerations are reasonable, they are not reflected in developing the moment methods.

Second, consider a situation in which the NR and/or MC models are employed for mixed-format tests and the constraints shown in Equations 6 and 8 are imposed for model identification. In this case, the sum of a_N parameter estimates from the new group and the counterpart from the old group both equal zero and thus do not contribute to the calculation of $M(\hat{a}_N)$ and $M(\hat{a}_O)$ in the mean/mean method.

Third, in the above situation, estimated values of the b_{jk} $(= -c_{jk} / a_{jk})$ reparameterized for the NR and MC models might be unstable though the reparameterization is theoretically legitimate. The instability can increase when both $\hat{a}_{jk}$ and $\hat{c}_{jk}$ are near zero and thus the resulting $\hat{b}_{jk}$ can be numerically unstable in magnitude and sign. This instability could negatively affect the estimation of the linking coefficients in both the mean/mean and mean/sigma methods. Indeed, Kim and Hanson (2002) pointed out that, as a result of such instability, the mean/mean and mean/sigma methods presented by Kim and Hanson (2000) are not feasible for the MC model.

Characteristic Curve Methods: Haebara and Stocking-Lord Methods

To develop the characteristic curve methods, each of item category characteristic curves or a test characteristic curve on the new scale is intended to be transformed and matched with the counterpart on the old scale, and vice versa. This transformation requires item parameter estimates on one scale to be expressed as those transformed to the other, while two ways of transformation—new-to-old and old-to-new—are conducted. The new-to-old and old-to-new transformations vary by IRT model. Under the 3PL, GR, and GPC models, from Equations 10 and 11, item parameter estimates on the new scale are transformed to the old scale by

$$\hat{a}^*_{jN} = \hat{a}_{jN} / A \text{ and } \hat{b}^*_{jkN} = A\hat{b}_{jkN} + B,$$

and item parameter estimates on the old scale are transformed to the new scale by

$$\hat{a}^\#_{jO} = A\hat{a}_{jO} \text{ and } \hat{b}^\#_{jkO} = (\hat{b}_{jkO} - B)/A.$$

Under the NR and MC models, from Equations 12 and 13, the new-to-old transformation is conducted by

$$\hat{a}^*_{jkN} = \hat{a}_{jkN} / A \text{ and } \hat{c}^*_{jkN} = \hat{c}_{jkN} - (B/A)\,\hat{a}_{jkN},$$

and the old-to-new transformation is conducted by

$$\hat{a}^\#_{jkO} = A\hat{a}_{jkO} \text{ and } \hat{c}^\#_{jkO} = \hat{c}_{jkO} + B\hat{a}_{jkO}.$$

The estimates of the parameters c_j (under the 3PL model) and d_{jk} (under the MC model) are not converted and remain the same on both scales. For example, $\hat{c}^*_{jN} = \hat{c}_{jN}$ and $\hat{c}^\#_{jO} = \hat{c}_{jO}$ in the case of the 3PL model. Here, $\hat{c}_{jN}$ is not necessarily equal to $\hat{c}_{jO}$ because of sampling error and possible model misfit.

For ease of further discussion, the estimated and transformed category characteristic functions for category k of item j are symbolized as follows: $\hat{P}_{jkO}(\theta_O)$ and $\hat{P}^*_{jkN}(\theta_O)$ for the old scale θ_O, and $\hat{P}_{jkN}(\theta_N)$ and $\hat{P}^\#_{jkO}(\theta_N)$ for the new scale θ_N. Specifically, for example, under the GR model the four functions (i.e., probabilities) are expressed as

$$\hat{P}_{jkO}(\theta_O) = P(\theta_O \mid \hat{a}_{jO}, \hat{b}_{jkO}, \hat{b}_{j(k+1)O}),\ \hat{P}^*_{jkN}(\theta_O) = P(\theta_O \mid \hat{a}^*_{jN}, \hat{b}^*_{jkN}, \hat{b}^*_{j(k+1)N}),$$
$$\hat{P}_{jkN}(\theta_N) = P(\theta_N \mid \hat{a}_{jN}, \hat{b}_{jkN}, \hat{b}_{j(k+1)N}), \text{ and } \hat{P}^\#_{jkO}(\theta_N) = P(\theta_N \mid \hat{a}^\#_{jO}, \hat{b}^\#_{jkO}, \hat{b}^\#_{j(k+1)O}).$$

Haebara method. For the Haebara method, the linking coefficients are obtained by minimizing the following criterion function, Q,

$$Q = Q_1 + Q_2, \tag{25}$$

where

$$Q_1 = \frac{1}{L} \int_{-\infty}^{+\infty} \sum_{m=1}^{M} \sum_{j=1}^{J_m} \sum_{k=1}^{K_{j(m)}} \left[\hat{P}_{jkO(m)}(\theta_O) - \hat{P}^{*}_{jkN(m)}(\theta_O) \right]^2 \psi_1(\theta_O) d\theta_O \tag{25a}$$

and

$$Q_2 = \frac{1}{L} \int_{-\infty}^{+\infty} \sum_{m=1}^{M} \sum_{j=1}^{J_m} \sum_{k=1}^{K_{j(m)}} \left[\hat{P}_{jkN(m)}(\theta_N) - \hat{P}^{\#}_{jkO(m)}(\theta_N) \right]^2 \psi_2(\theta_N) d\theta_N \,, \tag{25b}$$

where

$\psi_1(\theta_O)$ is a continuous distribution of θ_O,

$\psi_2(\theta_N)$ is a continuous distribution of θ_N,

M is the number of models,

J_m is the number of items nested in the m^{th} model,

$K_{j(m)}$ is the number of categories of item j under the m^{th} model,

$P_{jk(m)}(\theta)$ is a more specific expression of $P_{jk}(\theta)$ to indicate that the probability is defined under the m^{th} model, and

$$L = \sum_{m=1} \sum_{j=1} K_{j(m)} \,.$$

Note that parentheses around the subscript m are used to indicate that item j is nested in the m^{th} model. The quantity L is a factor to standardize the criterion function. Since it is not supposed to affect the solutions of the slope and intercept, L can be ignored.

Note in Equation 25 that a dichotomous IRT model is characterized as a special case of a polytomous IRT model having two categories. Thus, the incorrect response categories for the dichotomous IRT model also are used in defining the criterion function. If only the dichotomous IRT model is involved in calculating the criterion function, Q, it can be shown that the criterion function simplifies to the function suggested by Haebara (1980), in which only the correct response categories are taken into account.

To implement the Haebara method in practice requires a procedure to perform the integration in Equation 25. One possibility is a form of numerical integration that approximates Q. A practical form of Q, thus, is

$$Q \cong Q^* = Q_1^* + Q_2^*, \tag{26}$$

where

$$Q_1^* = \frac{1}{L_1}\sum_{i=1}^{G_O}\sum_{m=1}^{M}\sum_{j=1}^{J_m}\sum_{k=1}^{K_{j(m)}}\left[\hat{P}_{jkO(m)}(\theta_{iO}) - P^*_{jkN(m)}(\theta_{iO})\right]^2 W_1(\theta_{iO}) \tag{26a}$$

and

$$Q_2^* = \frac{1}{L_2}\sum_{i=1}^{G_N}\sum_{m=1}^{M}\sum_{j=1}^{J_m}\sum_{k=1}^{K_{j(m)}}\left[\hat{P}_{jkN(m)}(\theta_{iN}) - \hat{P}^{\#}_{jkO(m)}(\theta_{iN})\right]^2 W_2(\theta_{iN}), \tag{26b}$$

where θ_{iO} $(i = 1,2,\cdots,G_O)$ and $W_1(\theta_{iO})$ are proficiency points and weights intended to reflect the distribution of θ_O; θ_{iN} $(i = 1,2,\cdots,G_N)$ and $W_2(\theta_{iN})$ are proficiency points and weights for θ_N;

$$L_1 = \left(\sum_{i=1}^{G_O} W_1(\theta_{iO})\right)\times\left(\sum_{m=1}^{M}\sum_{j=1}^{J_m} K_{j(m)}\right); \text{ and } L_2 = \left(\sum_{i=1}^{G_N} W_2(\theta_{iN})\right)\times\left(\sum_{m=1}^{M}\sum_{j=1}^{J_m} K_{j(m)}\right).$$

Notice that the two factors of standardization, L_1 and L_2, are more generally defined in Equations 26a and 26b. Because the criterion function Q^* is non-linear with respect to A and B, a computationally intensive multivariate search technique (e.g., Dennis & Schnabel, 1996) is required to solve for A and B minimizing Q^*.

Stocking-Lord method. By convention, the criterion function for the Stocking-Lord method has been defined to be non-symmetric so that only the target scale (i.e., θ_O in the case of the new-to-old transformation) is taken into account. However, it may be desirable that the transformed scale (i.e., θ_N in the case of the new-to-old transformation) also be taken into account to define the criterion function. A general version of the Stocking-Lord method chooses A and B to minimize the following criterion function, F,

$$F = F_1 + F_2, \tag{27}$$

where

$$F_1 = \int_{-\infty}^{+\infty}\left[\hat{T}(\theta_O) - \hat{T}^*(\theta_O)\right]^2 \psi_1(\theta_O)d\theta_O \tag{27a}$$

and

$$F_2 = \int_{-\infty}^{+\infty} \left[\hat{T}(\theta_N) - \hat{T}^{\#}(\theta_N) \right]^2 \psi_2(\theta_N) d\theta_N . \tag{27b}$$

In Equations 27a and 27b, the estimated test characteristic functions are defined as

$$\hat{T}(\theta_O) = \sum_{m=1}^{M} \sum_{j=1}^{J_m} \sum_{k=1}^{K_{j(m)}} U_{jk(m)} \hat{P}_{jkO(m)}(\theta_O), \ \hat{T}^{*}(\theta_O) = \sum_{m=1}^{M} \sum_{j=1}^{J_m} \sum_{k=1}^{K_{j(m)}} U_{jk(m)} \hat{P}^{*}_{jkN(m)}(\theta_O),$$

$$\hat{T}(\theta_N) = \sum_{m=1}^{M} \sum_{j=1}^{J_m} \sum_{k=1}^{K_{j(m)}} U_{jk(m)} \hat{P}_{jkN(m)}(\theta_N), \text{ and } \hat{T}^{\#}(\theta_N) = \sum_{m=1}^{M} \sum_{j=1}^{J_m} \sum_{k=1}^{K_{j(m)}} U_{jk(m)} \hat{P}^{\#}_{jkO(m)}(\theta_N),$$

where $U_{jk(m)}$ is a weight allocated to the response category k of item j under the m^{th} model. The weight U_{jk} is called a scoring function and it is usually defined in such a way that $U_{jk} = k-1$ or $U_{jk} = k$. Note that the models involved in the criterion function, F, should be confined to the models under which the test characteristic curve is properly defined. Under the NR and MC models, the test characteristic function typically is not defined due to the lack of the definition of a true score.

A practical approximation of F is

$$F \cong F^{*} = F_1^{*} + F_2^{*}, \tag{28}$$

where

$$F_1^{*} = \frac{1}{L_1^{*}} \sum_{i=1}^{G_O} \left[\hat{T}(\theta_{iO}) - \hat{T}^{*}(\theta_{iO}) \right]^2 W_1(\theta_{iO}) \tag{28a}$$

and

$$F_2^{*} = \frac{1}{L_2^{*}} \sum_{i=1}^{G_N} \left[\hat{T}(\theta_{iN}) - \hat{T}^{\#}(\theta_{iN}) \right]^2 W_2(\theta_{iN}), \tag{28b}$$

where $L_1^{*} = \sum_{i=1} W_1(\theta_{iO})$, and $L_2^{*} = \sum_{i=1} W_2(\theta_{iN})$. The criterion function F^{*} is also non-linear with respect to A and B, and thus a multivariate search technique is required to solve for A and B minimizing F^{*}.

Simulation Study

A simulation study was conducted to compare the four linking methods for mixed-format tests. The mixed-format test selected was assumed to consist of DS items for which the 3PL model was appropriate for item analysis and PS items for which the GPC model was appropriate for item analysis. The types of items for the GPC model may include essay, passage-based, and rating-scale types. The present study, for generality, does not specify the type of PS items but assumes that they have characteristics of ordered or, possibly, semi-ordered responses. Although separate calibration by format (i.e., format-wise calibration) could have been used for test calibration, simultaneous calibration across formats was employed to estimate item parameters on the mixed-format test.

The linking scenario was that only one mixed-format test form was administered independently to two different groups, old and new. This means that all items in the test form were used as common items to link two "0, 1" scales from the two groups. The new-to-old transformation was considered, and item parameter estimates on the new scale were transformed to the old scale. The test form administered to the old group is called the "old" form and the same form administered to the new group is called the "new" form.

The criterion functions for the characteristic curve methods were defined as follows. The Haebara and Stocking-Lord methods used "symmetric" criterion functions that consider both of the old and new scales. Specifically, both Q_1^* and Q_2^* in Equation 26 were used to define the Haebara criterion function, and both F_1^* and F_2^* in Equation 28 were used to define the Stocking-Lord criterion function. For the summation scheme for the symmetric criterion functions, the proficiency distributions of the old and new groups were taken to be normal, although different distributions could have been used for the two groups. For the two distributions, 100 equally spaced proficiency points were chosen with the range of −4 to 4. At each point, the density of the standard normal distribution was found since the "0, 1" scale was used for each of the old and new forms. Each density was then divided by the sum of all the densities to standardize the densities, and the standardized densities served as the 100 weights.

Factors Investigated

Five factors were considered for the simulation study. The combination of the five factors led to 72 conditions [2 (levels of nonequivalence) × 2 (sample sizes) × 3 (types of mixed-format test) × 3 (types of linking) × 2 (computer programs)], under each of which the four linking methods for mixed-format tests were compared.

Equivalent versus nonequivalent groups linking. The need for obtaining a common scale typically occurs when the two distributions of proficiency for the old and new groups differ. Two linking situations were considered: (1) linking with equivalent groups and (2) linking with nonequivalent groups. The difference in proficiency between the new and old groups can be expressed as a new-to-old linear transformation, $\theta_N^* = A\theta_N + B$, where θ_N^* is equivalent to θ_O. If A = 1 and B = 0, the two groups are equivalent. For the nonequivalent groups linking, A = 1 and B = 1 were chosen, as was done in Hanson and Béguin (2002). In both equivalent and nonequivalent groups linking situations, the four linking methods were conducted. In the equivalent groups linking, no scale linking (hereafter, "no scaling" method) was also considered in addition to the four linking methods, since when the two groups are assumed to be equivalent one may assume A = 1 and B = 0 without performing any scale transformation.

Let $N(\mu, \sigma)$ be a normal distribution with mean μ and standard deviation σ. To simulate the conditions of equivalent and nonequivalent groups linking, examinees for the old group were generated by randomly sampling proficiency values from a N (0, 1). Then, for the equivalent groups linking, examinees for the new group were generated by randomly sampling proficiency values from a N (0, 1). For the nonequivalent groups linking, proficiency values were generated from a N (1, 1) for the new group.

Sample size. Two sample sizes were considered: (1) 500 examinees per form and (2) 3,000 examinees per form. For each set of the simulated item responses, 500 and 3,000 proficiency values were generated with the random seed being changed.

Types of mixed-format test. Three types of mixed-format test were used in terms of the number of DS and PS items: (1) 10 DS items and 10 PS items, (2) 20 DS items and 5 PS items, and (3) 30 DS items and 2 PS items. (All PS items were assumed to have five response categories.) The three test types are referred to, respectively, as 10/10, 20/5, and 30/2. The item parameters for the three types that were used for simulations are presented in Table 1. Note that the first 10 and 20 DS items on the 30/2 type were used for the 10/10 and 20/5 types, respectively. Similarly, the first 2 and 5 PS items on the 10/10 type were used for the 30/2 and 20/5 types, respectively. Therefore, the three types had 10 DS items and 2 PS items in common.

When a set of anchor items comprises several item formats in a mixed-format test, it can be useful to know which item type dominates the available formats because scale linking is often accomplished with only a single format (e.g., with DS items only) when the test is unidimensional. In the context of scale linking, a dominant item type is defined in this paper as the one that has more information to estimate the slope and intercept than any other. The amount of information available for an item type depends on the number of response categories. For example, the PS items are considered dominant for the 10/10 type, while the DS items are dominant for the 30/2 type.

For the 20/5 type, the decision depends on the linking method being used. Considering the slope and intercept separately, the DS items are dominant for the mean/mean method, and either the DS or PS items can be a dominant type for the mean/sigma method. Considering the number of response categories over items, on the other hand, the PS items can be viewed as the dominant type for the characteristic curve methods because 5 five-category PS items have a total of 25 response categories, whereas 20 DS items have a total of 20.

The decision on the dominant type, however, also should take into account the overall stability of item parameter estimates for the type. Therefore, a more sophisticated definition of dominant item type should include both the number of response categories and the stability of item parameter estimates.

Types of linking through different anchor item sets. To estimate linking coefficients of the new-to-old transformation, all or some of the common items can be used as actual anchor items. Although anchor items can be selected in various ways, the following three sets of anchor items were considered for each type of mixed-format test: (1) both DS and PS items, (2) DS items only, and (3) PS items only. In each set, all items of a single type were used as anchor items. The three sets of anchor items led to the following three types of linking: (1) simultaneous linking, (2) linking through DS items only, and (3) linking through PS items only. When two alternate mixed-format test forms have both DS and PS items in common, simultaneous linking can be viewed as ideal because both DS and PS items better represent their total test in content and characteristics. However, linking through DS items only is often chosen for practical reasons such as concerns about PS items in terms of reliability, security, and rater drift. Linking through PS items only was considered for comparative purposes, although its practical use would be rare. Linking coefficients from each of the three types of linking were used to transform the item parameter estimates for both DS and PS items from the new scale to the old scale.

Computer programs for calibration (MULTILOG versus PARSCALE). Two computer programs—MULTILOG (Thissen, 1991) and PARSCALE (Muraki & Bock, 1997)—were used for test calibration. By default, both computer programs determine the proficiency scale by setting the mean and standard deviation of the proficiency distribution used in the marginal maximum likelihood estimation at 0 and 1, respectively. MULTILOG handles the GPC model as a constrained version of the NR model by defining transformation matrices (specifically, triangle **T**-matrices) to fit the GPC model (see Appendix A). In addition, the item parameter estimates from MULTILOG must be transformed so they can be treated as parameter estimates from the GPC model. Specifically, for item j, the first-order contrast coefficient for $A(K)$ and the contrast coefficients for $C(K)$ from MULTILOG need to be transformed into the corresponding item parameter estimates (a and b_k) from PARSCALE according to the following relationships:

$$a = \frac{A(K)}{D} \text{ and } b_k = \frac{C(K)}{A(K)},$$

where D is a scaling constant. These transformations are practical. Childs and Chen (1999) described more technically the relationships between the estimates from MULTILOG and PARSCALE on the basis of the **T** matrices in MULTILOG.

Data Generation

True item parameters by model. For each of the three test types, 10/10, 20/5, and 30/2, the population item parameters for the 3PL model associated with the DS items and the GPC model associated with the PS items were generated as follows. The slope, difficulty, and lower asymptote parameters, a, b, and c, for the 3PL model were generated such that $a \sim LN(0, 0.2)$, $b \sim N(0, 1)$, and $c \sim BETA(8,32)$, where $LN(\mu, \sigma)$ designates a log normal distribution with mean μ and standard deviation σ, and $BETA(\alpha, \beta)$ a beta distribution with two parameters α and β. The item discrimination parameters, a, for the GPC model were generated from the same log normal distribution as for the 3PL model. Since all PS items had 5 response categories, four item-category parameters for each PS item were sampled from $N(-1.5, 0.2)$, $N(-0.5, 0.2)$, $N(0.5, 0.2)$, and $N(1.5, 0.2)$, and assigned to categories 2 through 5 in order. Because of the locations and scales for the four normal distributions, the four item-category parameters generated were sequentially ordered, although this is not required for the GPC model.

Dichotomous and graded response data. A dichotomous item response, U_{ij}, for a DS item associated with a sampled examinee was generated by comparing a value of the uniform random number, R, in the range (0, 1) to the population value of the correct response probability, P_{ij}, by the following rule: if $R \le P_{ij}$, then U_{ij} = 1; otherwise U_{ij} = 0. Similarly, a polytomous item response, U_{ijk}, for a PS item was generated by the following rule: for $k = 1, 2, \cdots, 5$, if $\breve{P}_{ij(k-1)} < R \le \breve{P}_{ijk}$, then $U_{ijk} = k-1$, where $\breve{P}_{ij0} = 0$ and $\breve{P}_{ijG} = \sum_{k=1}^{G} P_{ijk}$.

Computer Software for Scale Linking Methods

A computer program was written to implement the four scale linking methods for mixed-format tests. For the characteristic curve methods, the computer program employed the routines described in Dennis and Schnabel (1996) to find values of the slope and intercept that minimize the criterion functions for each of the Haebara and Stocking-Lord methods. The item parameter estimates from MULTILOG and PARSCALE served as input data to the computer program. (The computer program is available from the authors.)

Evaluation Criteria

In each condition, there were 100 sets of item parameter estimates for each of the old and new forms (i.e., 100 replications). It is expected that in each of the 100 replications the item parameter estimates for the new form should be on the same scale as the population item parameters after transformation. To evaluate the performance of the four linking methods, two evaluation criteria—scale linking coefficient (SLC) and category response curve (CRC)—were used.

The SLC criterion is based on the difference between the estimated and true linking coefficients. Thus, the estimates of A and B, denoted $\hat{A}$ and $\hat{B}$, are evaluated against their respective true values in each condition. Recall that $A = 1$ and $B = 0$ for the equivalent groups linking and $A = 1$ and $B = 1$ for the nonequivalent groups linking. The difference between the estimated and true values for each of A and B is quantified using the mean squared error (MSE), which can be decomposed into the squared bias and variance over 100 replications, as follows:

$$\frac{1}{100}\sum_{r=1}^{100}(A-\hat{A}_r)^2 = (A-M_{\hat{A}})^2 + \frac{1}{100}\sum_{r=1}^{100}(\hat{A}_r - M_{\hat{A}})^2 \tag{29}$$

and

$$\frac{1}{100}\sum_{r=1}^{100}(B-\hat{B}_r)^2 = (B-M_{\hat{B}})^2 + \frac{1}{100}\sum_{r=1}^{100}(\hat{B}_r - M_{\hat{B}})^2, \tag{30}$$

where

$$M_{\hat{A}} = \frac{1}{100}\sum_{r=1}^{100}\hat{A}_r \text{ and } M_{\hat{B}} = \frac{1}{100}\sum_{r=1}^{100}\hat{B}_r .$$

The CRC criterion is based on the difference between the estimated and true category response curves. Since the focus of this paper is to evaluate how well the new form item parameter estimates are put on the old scale, only the new form items are used in this criterion. The CRC criterion for item j with K_j categories (either a DS item or a PS item) is

$$\int_{-\infty}^{+\infty}\left\{\frac{1}{K_j}\sum_{k=1}^{K_j}\frac{1}{100}\sum_{r=1}^{100}[P_{jk}(\theta_O)-\hat{P}_{jkr}(\theta_O)]^2\right\}\psi(\theta_O)d\theta_O , \tag{31}$$

where $P_{jk}(\theta_O)$ is the category response function calculated with the population item parameters for category k of item j expressed on the old scale, $\hat{P}_{jkr}(\theta_O)$ is calculated with the transformed item parameter estimates for the item from replication r, and $\psi(\theta_O)$ is the density of a standard normal distribution. Equation 31 can be viewed as an adapted version of the criterion employed in Hanson and Béguin (2002) and expresses the average of the MSEs over categories of the difference between the estimated and true category response curves for item j. The quantity in Equation 31 is simply called the MSE below unless otherwise noted. The MSE in Equation 31 can be written as

$$\int_{-\infty}^{+\infty}\left\{\frac{1}{K_j}\sum_{k=1}^{K_j}\frac{1}{100}\sum_{r=1}^{100}[P_{jk}(\theta_O)-\hat{P}_{jkr}(\theta_O)]^2\right\}\psi(\theta_O)d\theta_O$$

$$= \int_{-\infty}^{+\infty}\left\{\frac{1}{K_j}\sum_{k=1}^{K_j}[P_{jk}(\theta_O)-m_{jk}(\theta_O)]^2\right\}\psi(\theta_O)d\theta_O$$

$$+ \int_{-\infty}^{+\infty}\left\{\frac{1}{K_j}\sum_{k=1}^{K_j}\frac{1}{100}\sum_{r=1}^{100}[\hat{P}_{jkr}(\theta_O)- \; m_{jk}(\theta_O)]^2\right\}\psi(\theta_O)d\theta_O , \tag{32}$$

where

$$m_{jk}(\theta_O) = \frac{1}{100}\sum_{r=1}^{100} \hat{P}_{jkr}(\theta_O).$$

The first term on the right side of Equation 32 is the squared bias, and the second term is the variance. To numerically calculate the integral involved in Equation 32, the Gauss-Hermite quadrature was used because the integrand incorporates a normal density. To obtain results with a high precision, 100 quadrature points and their weights were used. Then, average values of the squared bias, variance, and MSE over all new form items were computed in each condition.

Results

Appendices A and B present the command files used for simultaneous calibration across formats with MULTILOG and PARSCALE, respectively. Using 200 for the maximum number of EM cycles, the two programs converged in all simulation conditions.

Values of the squared bias and MSE for the SLC criterion are presented in Figures 1 through 4. The original values were too small and thus multiplied by 1,000 for ease of presentation. In each figure there are eight plots arranged in two columns. The plots in the two columns give the results for sample sizes of 3,000 and 500, respectively. In each plot, there are three line charts for the three test types (10/10, 20/5, and 30/2), and in each line chart results are compared among the three linking types (labeled SI, DS, and PS). Note that to make a clearer comparison among conditions, each plot employs its unique scale for the vertical axis.

Figures 5 through 8 present values of the average squared bias and MSE for the CRC criterion. The values were again multiplied by 1,000, and in each figure four plots are arranged in two columns. Note again that each plot employs its unique scale for the vertical axis. In Figures 5 and 6 for the linking with equivalent groups, the CRC criterion results without performing any scale linking are plotted as the "No Scaling" line.

In what follows, the results shown in the figures are described in detail. This paragraph summarizes some pronounced results from comparing among the scale linking methods regardless of the factors investigated, and the following sections describe results by factor investigated. With a few exceptions, for both the SLC and CRC criteria, the characteristic curve methods had lower MSE than the moment methods, and the Haebara method usually had the lowest MSE among the scale linking methods. The lower MSE for the Haebara and Stocking-Lord methods was due primarily to the lower variance, as found in Hanson and Béguin (2002). In the equivalent groups linking condition, in most cases the no scaling method led to less MSE than the scale linking methods. One exception is the simultaneous linking condition with the 20/5 and 30/2 types, where the characteristic curve methods led to slightly less MSE than the no scaling method.

Equivalent versus Nonequivalent Groups

Focusing on the SLC criterion, the values of the squared bias for each of the slope and intercept were very small for the linking conditions of both the equivalent and nonequivalent groups, although there was some variation across the four scale linking methods. This indicates that the four scale linking methods were properly extended to mixed-format tests. Overall, the values of the squared bias and MSE for the slope and intercept were smaller for the linking with equivalent groups than those for the linking with nonequivalent groups, but it was difficult to find a regular pattern in the difference of magnitude.

The results for the CRC criterion were similar to those for the SLC criterion. The nonequivalent groups linking resulted in higher average squared bias and MSE than the equivalent groups linking.

Sample Size

For both the SLC and CRC criteria, the MSE values were smaller when the sample size was larger. Some similar results were found for the squared bias.

Test Types and Linking Types

The results for the SLC criterion are given first. Focusing on the linking through single-format items only conditions, the MSE values, as expected, tended to decrease as the number of single format items increased. This tendency was more pronounced for the slope than for the intercept. For example, in the linking through DS items only condition, in most cases the 10/10 type had the greatest MSE values and the 30/2 type the smallest. In addition, the variation in MSE among the linking methods decreased as the number of single format items increased. Ignoring the mean/mean method for the 30/2 type, the linking through PS items only condition tended to show smaller variation in MSE among the linking methods than the other linking conditions did. When it comes to the Haebara and Stocking-Lord methods, with a few exceptions, the MSE was less in the simultaneous linking condition than in the linking through single-format items only conditions. This result is reasonable because more anchor items were used for simultaneous linking. However, there seemed to be a small difference in MSE between the simultaneous linking and linking through "dominant" items conditions. For example, the PS items were dominant for the 10/10 type and thus the MSE was slightly less in the simultaneous linking condition than in the linking PS items only condition.

To compare the three test types using the CRC criterion, it should be noted that the variance of the CRC criterion for an item was relatively larger for the DS items than for the PS items (the results are not presented in the figures). This result might be expected because the 5-category PS items, on average, have less spread across the category response curves than the DS items. Therefore, in the simultaneous linking condition, the average MSE tended to be larger when the number of the DS items was larger, mainly due to the increased average variance. Specific results for the CRC criterion are given below.

Regardless of the linking types, the values of the squared bias for the 30/2 type tended to be greater than that for the 10/10 and 20/5 types, which seemed related to the parametric characteristics of the items. Regardless of the test types, with very few exceptions, the simultaneous linking condition led to smaller MSE values than the linking through single-format items only conditions, as for the SLC criterion. In the linking through single-format items only conditions, the variation in MSE among the linking methods decreased as the number of single format items increased, as for the SLC criterion. Under the condition of linking through PS items only with the 30/2 type, the average MSE of the mean/mean method was always highest among the linking methods. This is related to the fact that only two discrimination estimates from two PS items are involved in estimating the slope of the new-to-old transformation in the mean/mean method; thus, the estimates could be unstable.

From the comparison by test type between the linking through DS items only and linking through PS items only conditions, the linking through "dominant" items condition led to lower MSE and less variation in MSE among the linking methods. For example, the PS items were dominant for the 10/10 type and thus the linking through PS items only condition had smaller MSE values and less variation in MSE among the linking methods than the linking through DS items only condition. For the 20/5 type, although either DS or PS items could be viewed as the dominant type for the moment methods, the linking through PS items only condition led to lower MSE and less variation in MSE among linking methods than the linking through DS items only condition. This seems to be partly because parameters of the PS items were recovered better than those of the DS items. One noteworthy point is that there was a small difference in MSE between the simultaneous linking and linking through a dominant item type conditions, at least in the characteristic curve methods.

MULTILOG versus PARSCALE

In general, MULTILOG and PARSCALE tended to perform similarly, although some differences in MSE occurred under certain conditions. This tendency was more pronounced for the characteristic curve methods than for the moment methods.

Summary and Discussion

The primary purpose of this paper was to extend "traditional" scale linking methods—more specifically, item parameter scaling methods—to mixed-format tests. For this purpose, this paper presented general versions of the following four scale linking methods extended to the mixed format tests: the mean/mean, mean/sigma, Haebara, and Stocking-Lord methods. This paper also investigated the performance of each linking method under several simulation conditions. Five factors were considered for simulation, leading to 72 conditions under each of which the four generalized linking methods were compared.

As with any other simulation study, the findings in this study have some limitations because of the unique design for simulation and the characteristics of the evaluation criteria employed. Therefore, caution should be exercised in drawing conclusions or overgeneralizing the findings to broader situations. In particular, the population item parameters were generated assuming statistical distributions rather than being based on real data. Although most of the generated dichotomous and polytomous items appeared to be "good" items in terms of discrimination and difficulty, the reality of the three test types used in simulation might be questionable. In addition, test data were simulated in an ideal situation where sampling error was well controlled and the model fit held. A partial justification for this ideal situation, however, is that ideally simulated situations should be preferred over real situations to show that the theoretical extension has been made properly. If the four linking methods are applied to more realistic situations, perhaps further distinctions could be found.

Based on the results from the simulation study, all the extended linking methods appeared to work properly. As the sample size became larger, the linking error decreased. Similarly, as the number of common items of a single format increased, the linking error tended to decrease. In addition, the linking error was less in the equivalent groups linking condition than it was in the nonequivalent groups linking condition. In general, MULTILOG and PARSCALE tended to perform alike. In the equivalent groups linking condition, for the 10/10 type conducting scale linking through linking methods resulted in larger linking error than no scaling. However, for the 20/5 and 30/2 types, the linking error for the no scaling method tended to be greater than that for the Haebara and Stocking-Lord methods but less than that for the mean/mean and mean/sigma methods.

With a few exceptions, the characteristic curve methods showed lower linking error than the moment methods, as found in Hanson and Béguin (2002). The Haebara method usually had the lowest linking error among the four linking methods. These results suggest that the characteristic curve methods are preferable to the moment methods.

With a very few exceptions, simultaneous linking (using both dichotomous and polytomous items) yielded more accurate results than linking through a single item type only, regardless of the test types. In the comparison between linking through dichotomous items only and linking through polytomous items only, linking through "dominant" types of items led to lower error and less variation in MSE among the linking methods. More important, there was a small difference in linking accuracy between simultaneous linking and linking through dominant item types. This implies that if practical situations do not permit the use of simultaneous linking, one may choose to use linking through the dominant item type after carefully considering both the number of response categories and the stability of parameter estimates of the item type.

The three types of mixed-format tests considered in this paper used only two distinct IRT models—the three-parameter logistic and generalized partial credit models. Other types of unidimensional IRT models, such as the graded response, nominal response, or multiple-choice model, could be considered as suitable IRT models for the subtests of the mixed-format tests. Consider a situation where the nominal response or multiple-choice model is used to analyze a set of polytomous items, say having four or five response categories. Also suppose that the constraints given in Equation 6 or 8 are imposed on the parameters for model identification, as is often the case. In such an event, the three linking methods other than the Haebara method are not likely to handle the nominal response or multiple-choice model properly, for different reasons, as follows.

The moment methods seem ineffective at performing scale linking because of the zero sums of a-parameter estimates as well as the instability of (reparameterized) b-parameter estimates. Actually, in a simulation study conducted independently of the present study, the moment methods did not work at all under the nominal response model with the constraints. The Stocking-Lord method also has a limitation in that it cannot incorporate the response categories from the nominal response or multiple-choice model into the criterion function unless a proper scoring function is assigned to the categories. By contrast, the Haebara method can handle any mixture of IRT models without difficulty. In fact, the Haebara method takes full advantage of the invariance property in IRT that item characteristic curves are invariant from group to group except a linear difference between proficiency scales (Lord, 1980). Fortunately, the superiority of the Haebara method has been verified in several other studies (Hanson & Béguin, 2002; Kim & Kolen, 2004; Kim & Song, 2004).

The characteristics and behavior of the four linking methods extended to mixed-format tests need to be further examined with different simulation conditions or in real situations. There are two other important research questions: (1) How do the linking methods following separate calibration perform in situations where test data do not fit the models and/or the assumptions of the IRT models involved in a mixed-format test fail to hold? (2) With mixed-format tests, how does the performance of the linking methods following separate calibration compare to the performance of concurrent calibration, which does not require linking coefficients? Further investigation is needed to answer these questions.

References

Baker, F.B. (1992). Equating tests under the graded response model. *Applied Psychological Measurement, 16*, 87-96

Baker, F.B. (1993). Equating tests under the nominal response model. *Applied Psychological Measurement, 17*, 239-251.

Baker, F.B. (1997). Empirical sampling distributions of equating coefficients for graded and nominal response instruments. *Applied Psychological Measurement, 21*, 157-172.

Baker, F.B., & Al-Karni, A. (1991). A comparison of two procedures for computing IRT equating coefficients. *Journal of Educational Measurement, 28*, 147-162.

Bennett, R.E., Rock, D.A., & Wang, M. (1991). Equivalence of free-response and multiple-choice items. *Journal of Educational Measurement, 28*, 77-92.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F.M. Lord and M.R. Novick (Eds.), *Statistical theories of mental test scores* (pp. 397-472). Reading, MA: Addison-Wesley.

Bock, R.D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika, 37*, 29-51.

Childs, R.A., & Chen, W. (1999). Obtaining comparable item parameter estimates in MULTILOG and PARSCALE for two polytomous IRT models. *Applied Psychological Measurement, 23*, 371-379.

Cohen, A.S., & Kim, S.-H. (1998). An investigation of linking methods under the graded response model. *Applied Psychological Measurement, 22*, 116-130.

Dennis, J.E., & Schnabel, R.B. (1996). *Numerical methods for unconstrained optimization and nonlinear equations*. Philadelphia: Society for Industrial and Applied Mathematics.

Divgi, D. R. (1985). A minimum chi-square method for developing a common metric in item response theory. *Applied Psychological Measurement, 9*, 413-415.

Haebara, T. (1980). Equating logistic ability scales by a weighted least squares method. *Japanese Psychological Research, 22*, 144-149.

Hanson, B.A. & Béguin, A.A. (2002). Obtaining a common scale for item response theory item parameters using separate versus concurrent estimation in the common-item equating design. *Applied Psychological Measurement, 26*, 3-24.

Kim, J.-S., & Hanson, B.A. (2000). *Test equating under the multiple-choice model.* Paper presented at the annual meeting of American Educational Research Association, New Orleans, LA.

Kim, J.-S., & Hanson, B.A. (2002). Test equating under the multiple-choice model. *Applied Psychological Measurement, 26*, 255-270.

Kim, S., & Kolen, M.J. (2004). *Optimally defining criterion functions for the characteristic curve procedures in the IRT scale linking.* Paper presented at the annual meeting of the American Educational Research Association, San Diego, CA.

Kim, S., & Song, M.-Y. (2004). *Least squares estimation of IRT scale linking coefficients under the graded response model.* Paper presented at the annual meeting of the American Educational Research Association, San Diego, CA.

Kim, S.-H., & Cohen, A.S. (1992). Effects of linking methods on detection of DIF. *Journal of Educational Measurement, 29*, 51-66.

Kim, S.-H., & Cohen, A.S.(1995). A minimum χ^2 method for equating tests under the graded response model. *Applied Psychological Measurement, 19*, 167-176.

Kolen, M.J., & Brennan, R.L. (1995). *Test equating: Methods and practices.* New York: Springer-Verlag.

Li, Y.H., Lissitz, R.W., & Yang, Y. (1999). *Estimating IRT equating coefficients for tests with polytomously and dichotomously scored items.* Paper presented at the annual meeting of the National Council on Measurement in Education, Montreal, Quebec, Canada.

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

Loyd, B.H., & Hoover, H.D. (1980). Vertical equating using the Rasch model. *Journal of Educational Measurement, 17*, 179-193.

Marco, G.L. (1977). Item characteristic curve solutions to three intractable testing problems. *Journal of Educational Measurement, 14*, 139-160.

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement, 16*, 159-176.

Muraki, E., & Bock, R.D. (1997). *PARSCALE 3: IRT based item analysis and test scoring for rating-scale data.* Chicago, IL: Scientific Software International, Inc.

Ogasawara, H. (2001). Least squares estimation of item response theory linking coefficients. *Applied Psychological Measurement*, 25(4), 373-383.

Rosa, K., Swygert, K.A., Nelson, L., & Thissen, D. (2001). Item response theory applied to combinations of multiple-choice and constructed-response items – Scale scores for patterns of summed scores. In D. Thissen, & H. Wainer (Eds.), *Test scoring* (pp. 253-292). Mahwah, NJ: Lawrence Erlbaum Associates.

Samejima, F. (1969). Estimation of a latent ability using a response pattern of graded scores. *Psychometrika Monograph Supplement, 17.*

Samejima, F. (1972). A general model for free response data. *Psychometrika Monograph Supplement, 18.*

Stocking, M., & Lord, F.M. (1983). Developing a common metric in item response theory. *Applied Psychological Measurement, 7*, 207-210.

Swygert, K.A., McLeod, L.D., & Thissen, D. (2001). Factor analysis for items or testlets scored in more than two categories. In D. Thissen, & H. Wainer (Eds.), *Test scoring* (pp. 217-250). Mahwah, NJ: Lawrence Erlbaum Associates.

Tate, R.L. (1999). A cautionary note on IRT-based linking of tests with polytomous items. *Journal of Educational Measurement, 36*, 336-346.

Tate, R.L. (2000). Performance of a proposed method for the linking of mixed format tests with constructed response and multiple choice items. *Journal of Educational Measurement, 37*, 329-346.

Thissen, D. (1991). *MULTILOG user's guide: Multiple, categorical item analysis and test scoring using item response theory*. Chicago, IL: Scientific Software International, Inc.

Thissen, D., & Steinberg, L. (1984). A response model for multiple choice items. *Psychometrika, 49*, 501-519.

Thissen, D., & Steinberg, L. (1997). A response model for multiple-choice items. In W.J. van der Linden & R.K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 51-65). New York: Springer-Verlag.

Thissen, D., Wainer, H., & Wang, X.-B. (1994). Are tests comprising both multiple-choice and free-response items necessarily less unidimensional than multiple-choice tests? An analysis of two tests. *Journal of Educational Measurement, 31*, 113-123.

Traub, R.E. (1993). On the equivalence of the traits assessed by multiple-choice and constructed-response tests. In R.E. Bennett, & W.C. Ward (Eds.), *Construction versus choice in cognitive measurement* (pp. 29-44). Hillsdale, NJ: Lawrence Erlbaum Associates.

Wainer, H., & Thissen, D. (1993). Combining multiple-choice and constructed-response test scores: Toward a Marxist Theory of test construction. *Applied Measurement in Education, 6*, 103-118.

Appendix A

This appendix gives the MULTILOG command files used for the 10/10 type with a sample size of 3000. When the other test types and the sample size of 500 are used, the keywords, "NI" and "NE" in the command "PRO" should be properly given their values according to the number of items and the number of examinees, respectively. In addition, a series of alphanumeric codes after the command "END" should also be specified correctly in a given condition.

Simultaneous Calibration with Both DS and PS Items (MULTILOG)

```
>PRO RA, IN, NI=20, NG=1, NE=3000, DATA='data filename';
>TES IT=(1(1)10), L3;
>TES IT=(11(1)20), NO, NC=(5(0)10), HI=(5(0)10);
>PRI IT=(1(1)10), AJ PA=(1.7, 1.0);
>PRI IT=(1(1)10), CJ PA=(-1.4, 1.0);
>TMA IT=(11(1)20), AK, POLY;
>TMA IT=(11(1)20), CK, TRIA;
>FIX IT=(11(1)20), AK=(2,3,4), VAL=0.0;
>TGR NU=40,
QP=(-4.0000, -3.7950, -3.5900, -3.3850, -3.1790,
    -2.9740, -2.7690, -2.5640, -2.3590, -2.1540,
    -1.9490, -1.7440, -1.5380, -1.3330, -1.1280,
    -0.9231, -0.7179, -0.5128, -0.3077, -0.1026,
     0.1026,  0.3077,  0.5128,  0.7179,  0.9231,
     1.1280,  1.3330,  1.5380,  1.7440,  1.9490,
     2.1540,  2.3590,  2.5640,  2.7690,  2.9740,
     3.1790,  3.3850,  3.5900,  3.7950,  4.0000);
>SAV;
>EST NC=200, IT=10;
>END ;
5
01234
11111111111111111111
22222222222222222222
00000000003333333333
00000000004444444444
00000000005555555555
(20A1)
```

Appendix B

This appendix gives the PARSCALE command files used for the 10/10 type with a sample size of 3000. When the other test types are used, the keywords, "NTOTAL" and "LENGTH" in the command "INPUT", the keyword "NBLOCK" in the command "TEST", and the keyword "REP" in the command "BLOCK" should be properly given their values according to the number of items. The FORTRAN variable format statements should also be specified properly according to the structure of item response data.

Simultaneous Calibration with Both DS and PS Items (PARSCALE)

```
>COMMENT;
>FILE    DFNAME='data filename', SAVE;
>SAVE    PARM='parameter filename';
>INPUT   NIDW=4, NTOTAL=20, NTEST=1, LENGTH=20, NFMT=1;
(4A1,T1,20A1)
>TEST    TNAME=MIXTEST, ITEM=(1(1)20), NBLOCK=20;
>BLOCK   BNAME=DSITEMS, NITEMS=1, NCAT=2,
         ORI=(0,1), MOD=(1,2), GPARM=0.2, GUESS=(2,EST), REP=10;
>BLOCK   BNAME=PSITEMS, NITEMS=1, NCAT=5,
         ORI=(0,1,2,3,4), MOD=(1,2,3,4,5), REP=10;
>CALIB   PAR, LOG, SCALE=1.7, NQPTS=40, ESTORDER,
         CYCLES=(200,10,10,10,1,1), NEWTON=5, SPRI, GPRI, PRI;
>PRIORS SSI=(0.6(0)20);
>SCORE NOSCORE;
```

TABLE 1

Item Parameters for Dichotomous and Polytomous Items Used for Simulations

Dichotomous	Parameters			Types of test		
Items	*a*	*b*	*c*	10/10	20/5	30/2
1	1.150	0.661	0.216	x	x	x
2	1.452	-0.012	0.096	x	x	x
3	0.819	0.629	0.154	x	x	x
4	1.092	0.008	0.161	x	x	x
5	0.844	-1.632	0.249	x	x	x
6	1.273	-0.964	0.183	x	x	x
7	0.839	-0.105	0.202	x	x	x
8	1.130	1.330	0.257	x	x	x
9	0.896	0.264	0.186	x	x	x
10	1.042	0.814	0.253	x	x	x
11	0.915	-1.492	0.082		x	x
12	0.857	1.130	0.276		x	x
13	0.778	1.725	0.114		x	x
14	1.260	0.052	0.162		x	x
15	1.312	0.983	0.232		x	x
16	0.951	-1.410	0.145		x	x
17	1.190	1.071	0.214		x	x
18	1.222	0.837	0.164		x	x
19	1.194	1.492	0.229		x	x
20	0.678	0.103	0.218		x	x
21	0.885	-1.910	0.236			x
22	1.212	0.433	0.218			x
23	1.101	-0.143	0.137			x
24	0.804	0.931	0.161			x
25	0.882	1.420	0.141			x
26	1.016	0.701	0.137			x
27	0.932	0.928	0.135			x
28	0.776	-1.572	0.255			x
29	1.022	-0.646	0.073			x
30	1.448	-1.051	0.097			x

Polytomous	Parameters					Types of test		
Items	*a*	b_2	b_3	b_4	b_5	10/10	20/5	30/2
1	0.972	-1.442	-0.321	0.506	1.552	x	x	x
2	1.091	-1.567	-0.512	0.602	1.430	x	x	x
3	0.954	-1.420	-0.579	0.385	1.802	x	x	
4	1.201	-1.523	-0.365	0.365	1.332	x	x	
5	0.874	-1.745	-0.447	0.694	1.173	x	x	
6	1.120	-1.680	-0.557	0.628	1.474	x		
7	1.076	-1.302	-0.446	0.365	1.654	x		
8	0.871	-1.171	-0.249	0.523	1.281	x		
9	1.065	-1.862	-0.135	0.187	1.503	x		
10	1.090	-1.523	-0.754	0.622	1.732	x		

Note: "x" indicates items that are assigned to the type.

***FIGURE 1*. Squared Bias and MSE for Scale Linking Coefficient (SLC) Criterion Simulated with Equivalent Groups, Analyzed with MULTILOG**

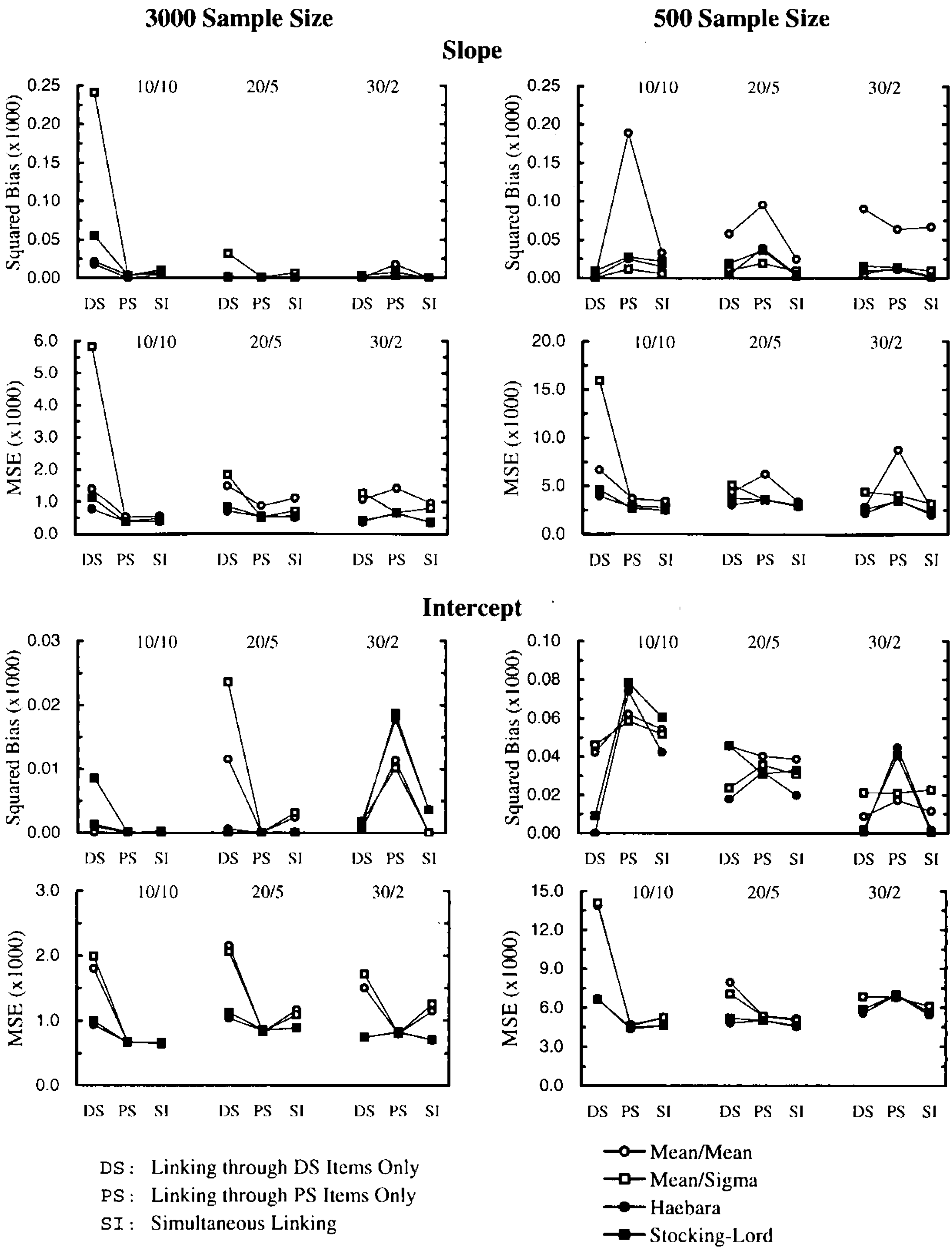

***FIGURE 2*. Squared Bias and MSE for Scale Linking Coefficient (SLC) Criterion Simulated with Equivalent Groups, Analyzed with PARSCALE**

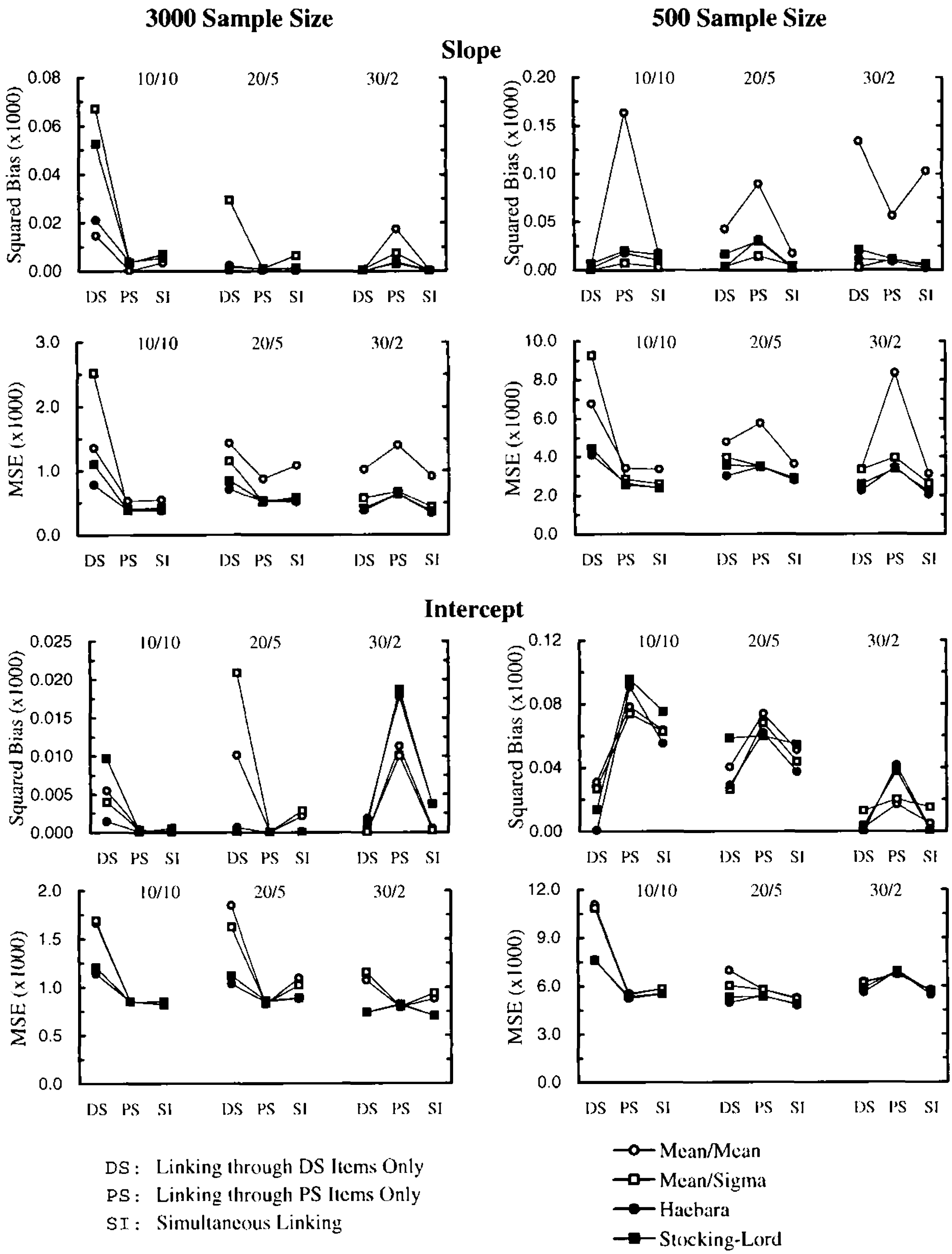

FIGURE 3. **Squared Bias and MSE for Scale Linking Coefficient (SLC) Criterion Simulated with Nonequivalent Groups, Analyzed with MULTILOG**

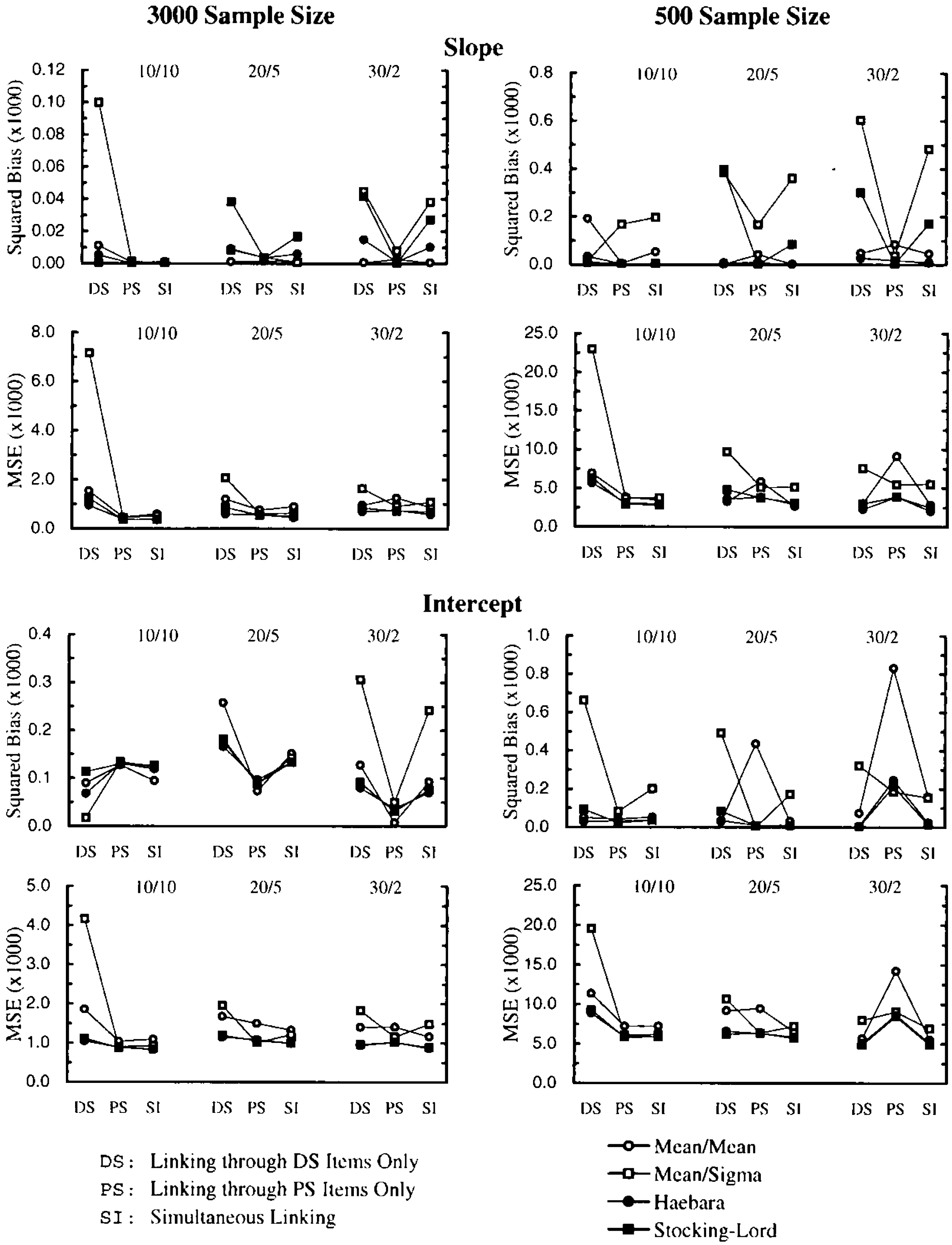

***FIGURE 4*. Squared Bias and MSE for Scale Linking Coefficient (SLC) Criterion Simulated with Nonequivalent Groups, Analyzed with PARSCALE**

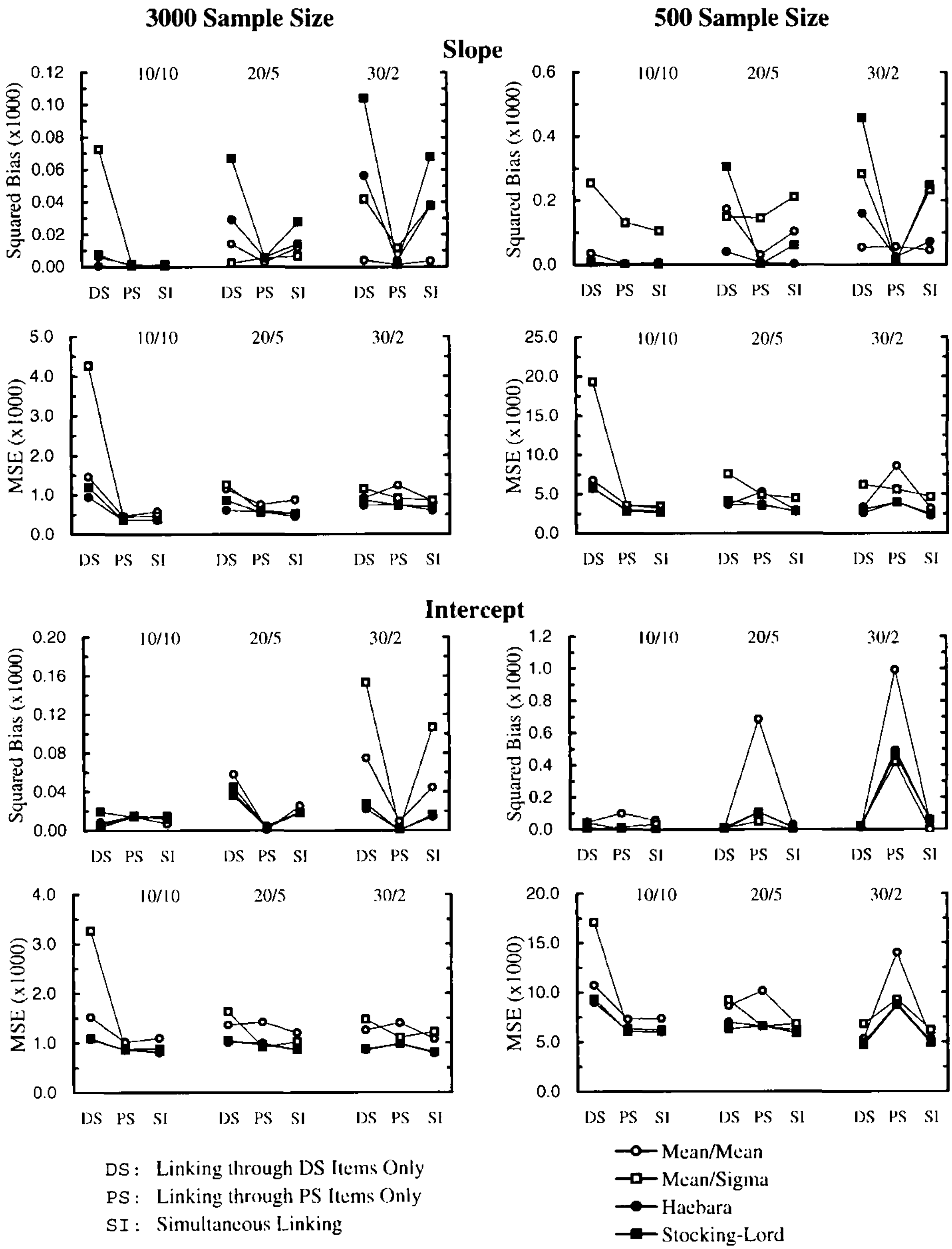

***FIGURE 5*. Squared Bias and MSE for Category Response Curve (CRC) Criterion Simulated with Equivalent Groups, Analyzed with MULTILOG**

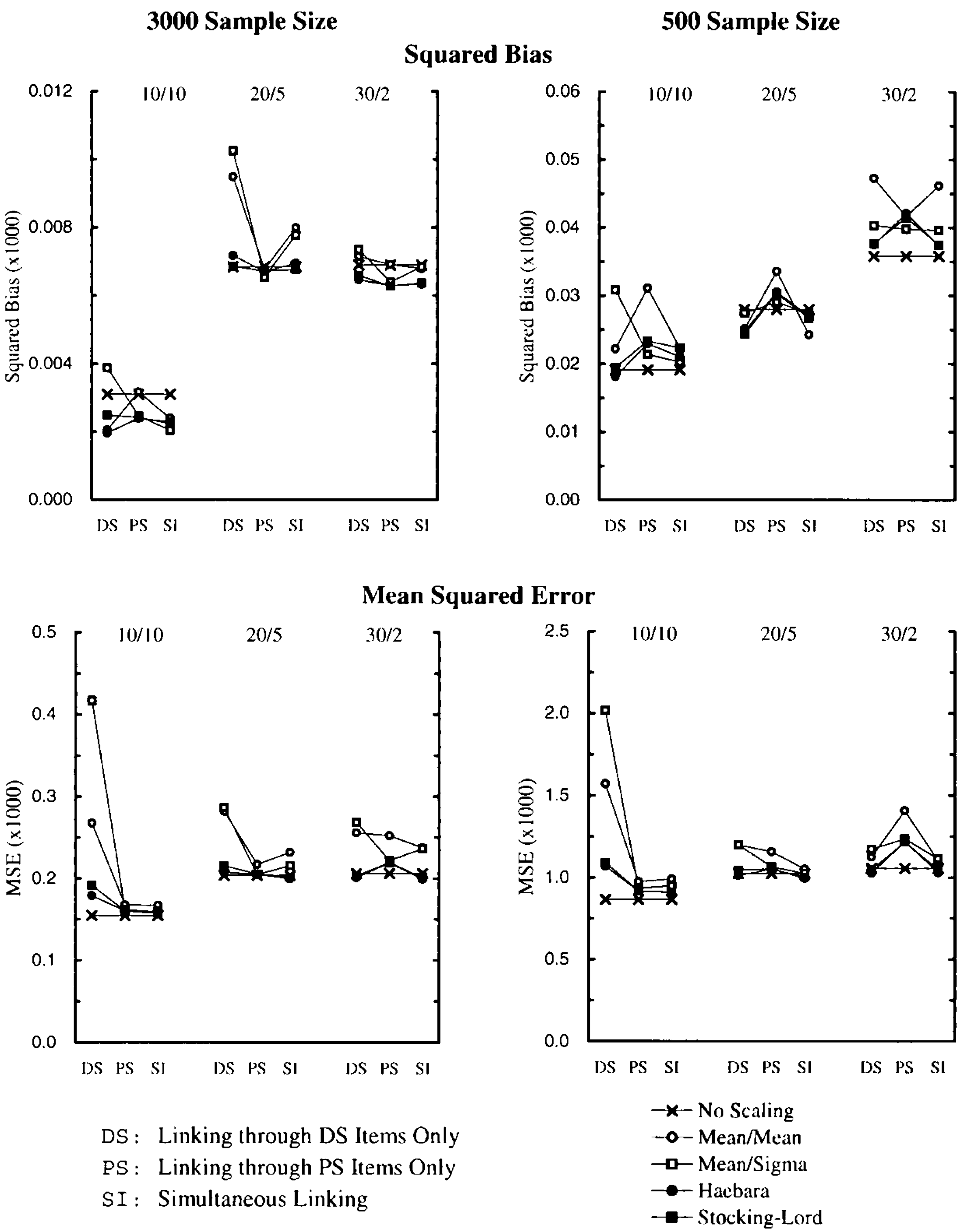

FIGURE 6. **Squared Bias and MSE for Category Response Curve (CRC) Criterion Simulated with Equivalent Groups, Analyzed with PARSCALE**

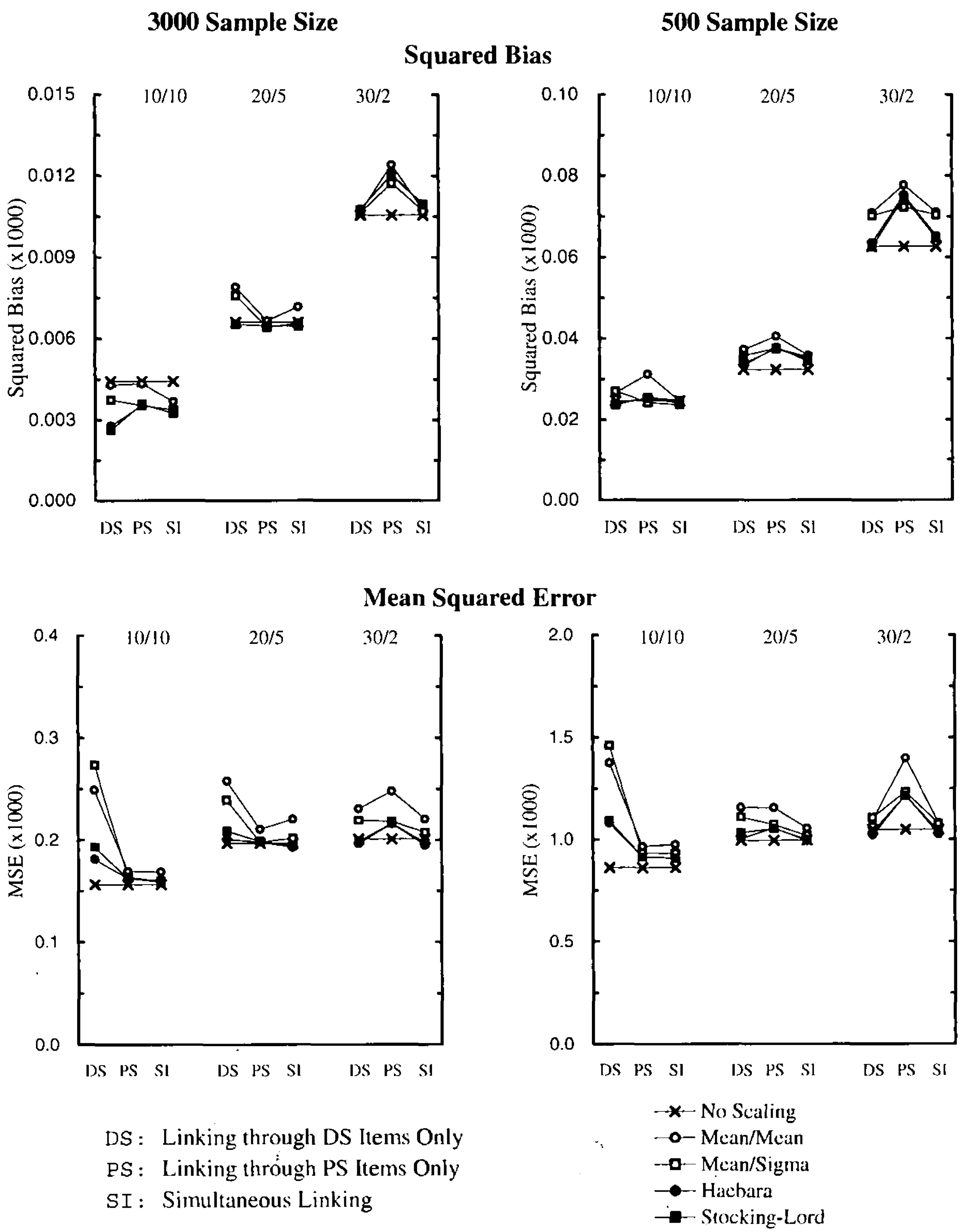

FIGURE 7. **Squared Bias and MSE for Category Response Curve (CRC) Criterion Simulated with Nonequivalent Groups, Analyzed with MULTILOG**

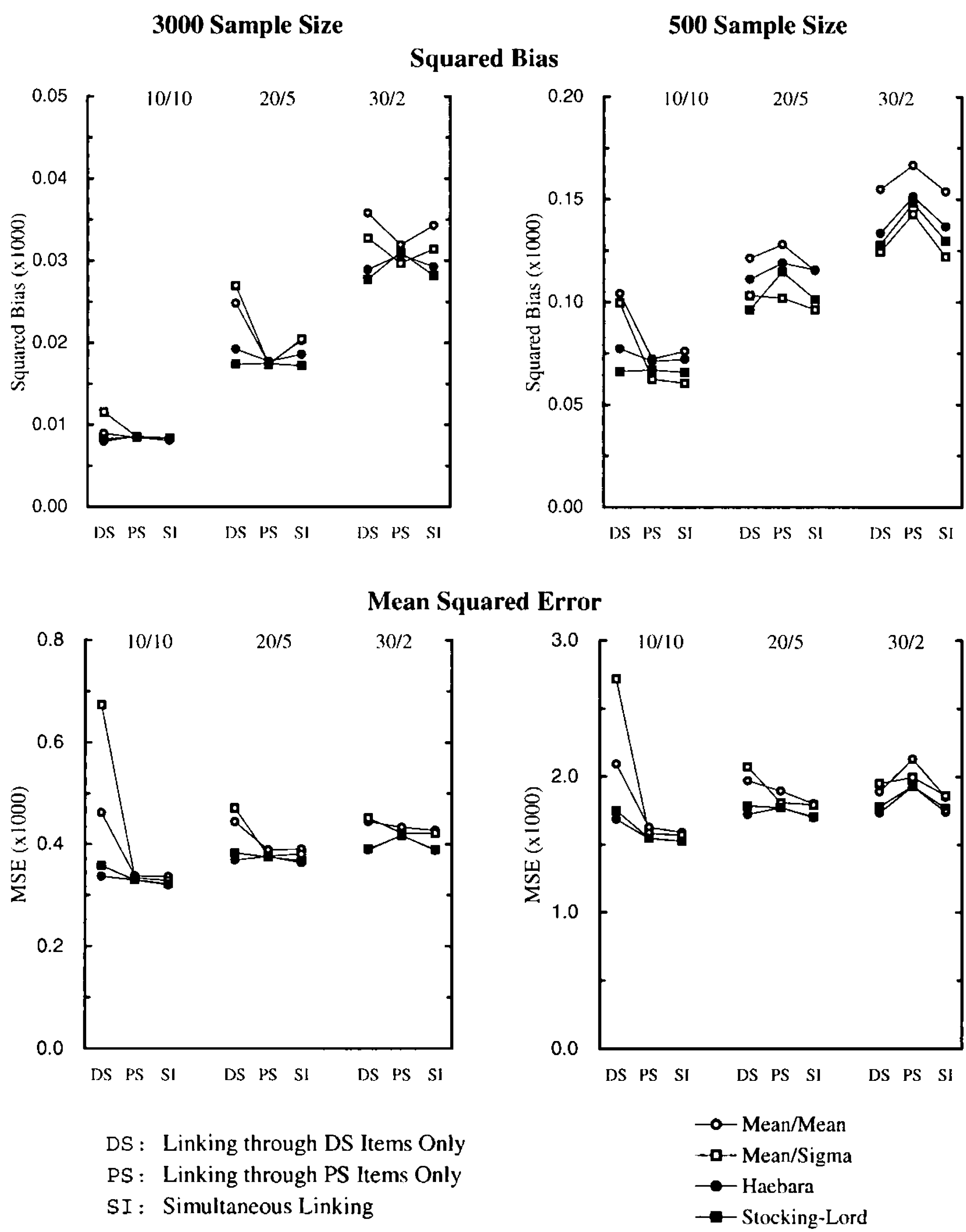

***FIGURE 8*. Squared Bias and MSE for Category Response Curve (CRC) Criterion Simulated with Nonequivalent Groups, Analyzed with PARSCALE**