DOCUMENT RESUME

ED 476 863	TM 034 957

AUTHOR	Roberts, James S.; Bao, Han; Huang, Chun-Wei; Gagne, Phill
TITLE	Exploring Alternative Characteristic Curve Approaches to Linking Parameter Estimates from the Generalized Partial Credit Model.
PUB DATE	2003-04-00
NOTE	49p.; Paper presented at the Annual Meeting of the National Council on Measurement in Education (Chicago, IL, April 22-24, 2003).
PUB TYPE	Reports - Research (143) -- Speeches/Meeting Papers (150)
EDRS PRICE	EDRS Price MF01/PC02 Plus Postage.
DESCRIPTORS	*Estimation (Mathematics); *Mathematical Models
IDENTIFIERS	*Item Characteristic Function; Operating Characteristics Estimation; *Partial Credit Model

ABSTRACT

Characteristic curve approaches for linking parameters from the generalized partial credit model were examined for cases in which common (anchor) items are calibrated separately in two groups. Three of these approaches are simple extensions of the test characteristic curve (TCC), item characteristic curve (ICC), and operating characteristic curve (OCC) methods that have been previously developed for other binary item response models. The ICC approach explicitly provides a symmetric solution for estimating linking constants whereas the TCC and OCC approaches yield an asymmetric solution. Thus, the symmetry of the result is confounded with the type of characteristic curve used to derive the result. New characteristic curve techniques are developed to estimate this confound. Specifically, symmetric versions of the TCC and OCC methods are developed within the context of the generalized partial credit model (GPCM) along with an asymmetric version of the ICC technique. The accuracy of linking constant estimates and the accuracy of rescaled GPCM parameter estimates obtained with each method was examined in a simulation study. The study suggested that the TCC method yields slightly more accurate estimates of linking constants and model parameters as do symmetric, as opposed to asymmetric, solutions. The study suggests that all of the methods yield similar linking results when GPCM parameters are estimated accurately using large samples. (Contains 4 tables, 9 figures, and 13 references.) (Author/SLD)

Exploring Alternative Characteristic Curve Approaches to Linking Parameter Estimates from the Generalized Partial Credit Model

James S. Roberts, Han Bao, Chun-Wei Huang, Phill Gagne

University of Maryland

RUNNING HEAD: Alternative Characteristic Curve Approaches to Linking

Version 1.0

Presented at the annual meeting of the National Council on Measurement in Education,

Chicago, Illinois, April 24, 2003.

Abstract

Characteristic curve approaches to linking parameters from the generalized partial credit model (GPCM) are examined for cases where common (anchor) items are calibrated separately in two groups. Three of these approaches are simple extensions of the test characteristic curve (TCC), item characteristic curve (ICC), and operating characteristic curve (OCC) methods that have been previously developed for other binary item response models. The ICC approach explicitly provides a symmetric solution for estimating linking constants whereas the TCC and OCC approaches yield an asymmetric solution. Thus, the symmetry of the result is confounded with the type of characteristic curve used to derive the result. New characteristic curve techniques are developed to eliminate this confound. Specifically, symmetric versions of the TCC and OCC methods are developed within the context of the GPCM along with an asymmetric version of the ICC technique. The accuracy of linking constant estimates and the accuracy of rescaled GPCM parameter estimates obtained with each method is examined in a simulation study. The study suggested that the TCC method yields slightly more accurate estimates of linking constants and model parameters as do symmetric, as opposed to asymmetric, solutions. The study suggests that all of the methods yield similar linking results when GPCM parameters are estimated accurately using large samples.

Exploring Alternative Characteristic Curve Approaches to Linking Parameter Estimates from the Generalized Partial Credit Model

Parametric item response theory (IRT) models generally yield parameter estimates that are identifiable up to some arbitrary change in location (i.e., origin), and perhaps an arbitrary change in scale (i.e., unit) when the model contains a discrimination parameter. Constraints must be introduced to achieve a unique solution when estimating parameters in these models. These constraints, in turn, yield parameter estimates that have a data dependent metric. To overcome this dependency, parameter estimates are typically transformed to achieve a common metric before comparing estimates from different calibrations. We refer to this transformation process as "linking" parameter estimates.

There have been several methods proposed to link parameter estimates using common (anchor) items from alternative calibrations of a given parametric IRT model (Kolen & Brennan, 1995). Linking methods based on characteristic curve approaches are generally thought to be reasonably accurate and more robust to outliers as compared to methods based simply on summary measures of parameter estimate distributions (Baker & Al-Karni, 1991; Stocking & Lord, 1983). Moreover, characteristic curve methods only require access to the IRT parameter estimates themselves, and not to the estimated item parameter covariances or to the raw item response data. Therefore, this paper will focus on characteristic curve approaches to linking IRT parameter estimates.

To date, there have been three distinct characteristic curve approaches developed in the IRT literature. These are the test characteristic curve (TCC) method (Stocking & Lord, 1983), the item characteristic curve (ICC) method (Haebara, 1980), and the option characteristic curve method (Baker, 1993). The first two of these methods were proposed in the context of binary logistic models whereas the latter was developed in the context of the nominal response model (Bock, 1972). This paper will focus on extending each of these methods to Muraki's (1992) generalized partial credit model (GPCM). The GPCM is defined in the *gth* calibration group as:

$$P_{(g)ik}(\theta_{(g)j}, a_{(g)i}, b_{(g)i}, \underline{d}_{(g)i}) = \frac{\exp\left[\sum_{v=1}^{k} a_{(g)i}(\theta_{(g)j} - b_{(g)i} + d_{(g)iv})\right]}{\sum_{c=1}^{K_i} \exp\left[\sum_{v=1}^{c} a_{(g)i}(\theta_{(g)j} - b_{(g)i} + d_{(g)iv})\right]} \qquad \textbf{(1)}$$

where:

$P_{(g)ik}$ is the operating characteristic function for the *kth* category of the *ith* item in the *gth* calibration group,

$\theta_{(g)j}$ is the location of the *jth* individual from the *gth* calibration group on the latent continuum,

$a_{(g)i}$ is the discrimination parameter for the *ith* item from the *gth* calibration group,

$b_{(g)i}$ is the location of the *ith* item from the *gth* calibration group

$\underline{d}_{(g)i}$ is the vector of K_i threshold parameters for the *ith* item in the *gth* calibration group which are constrained to sum to zero, and

K_i is the number of response categories for the *ith* item.

The GPCM is interesting in that all three characteristic curve approaches can be used to link parameter estimates that are derived from it. However, there has not been a systematic comparison of these alternative characteristic curve linking methods. This paper will, among other things, report on such a comparison.

The TCC, ICC, and OCC linking methods differ from each other with regard to symmetric or asymmetric nature of their solutions. All of these methods attempt to derive linking constants that are used to transform parameter estimates from one calibration group (i.e., the transformed group) to the metric of estimates from the other group (i.e., the target group). The TCC and OCC methods yield linking constants that are dependent on which set of estimates constitute the target metric. If these methods are used to obtain two solutions in which the target calibration group is reversed, then the two results will not be simple inverses of each other. We refer to these types of solutions as asymmetric. In contrast, the ICC method yields a symmetric solution. If the ICC method is applied twice such that the roles of the target and transformation groups are reversed across the two applications, then the resulting solutions will be simple inverses of each other. As mentioned above, the past literature on characteristic curve approaches to linking IRT parameter estimates has confounded the notion of solution symmetry with the type of characteristic curve use to derive linking constants. This has important implications for measurement practice because the choice between a symmetric or asymmetric linking solution should be based on the measurement situation and not on the characteristic curve that one wishes to use. For example, studies of IRT parameter invariance generally estimate parameters separately in two groups, link the estimates to achieve a common metric, and then compare the linked parameter estimates. The selection of a target calibration group is arbitrary in such cases, and therefore, a characteristic curve approach to linking that is symmetric would be preferred. Alternatively, in situations where new test items are calibrated and subsequently added to a pre-existing item bank in which the metric is fixed, the new estimates must first be linked to the metric of items in the bank. In this case, there is a pre-existing target metric, and an asymmetric solution is justified. Unfortunately, the previous confound between the type of characteristic curve used to develop linking constants and the symmetric or asymmetric nature of the solution limits the practitioner's ability to choose the method that is most appropriate for a given measurement situation.

In order to more easily choose among characteristic curve methods of linking IRT parameter estimates, the symmetry of a solution must be disentangled from the type of characteristic curve used to derive the solution. Accordingly, a second objective of this paper was to overcome the confound between these two features. We did this by developing symmetric versions of the TCC and OCC procedures and an asymmetric version of the ICC procedure. These developments are described below along with a description of the original TCC, ICC and OCC methods.

The Test Characteristic Curve (TCC) Method

Stocking and Lord (1983) proposed a TCC approach in which the scale and location of parameters from one calibration are linearly transformed to match the metric of parameter estimates from a second calibration based on a subset of common items (i.e., anchor items). The transformation minimized the sum of squared differences between the test characteristic curves associated with common items from the two sets of parameter estimates. Originally proposed for a 3-parameter logistic model for binary responses, the TCC method can be generalized easily to the GPCM. Let $\theta_{(2)j_2}$ denote the location of the j_2th individual from the second calibration group on the latent continuum, and let $\theta_{(21)j_2}$ represent the location of the j_2th individual from the second calibration group after it has been transformed to the metric of the first calibration group. Similarly, let $b_{(2)i}$ refer to the location of the *ith* item from the second calibration, and let $b_{(21)i}$ represent the location of the *ith* item from the second calibration after it has been transformed to the metric of the first calibration group. The item discrimination parameter ($a_{(2)i}$ and $a_{(21)i}$) and the threshold parameter ($d_{(2)ik}$ and $d_{(21)ik}$) are subscripted in an analogous fashion. The TCC procedure estimates the linking constants, A and B, that transform the metric of parameter estimates from the second calibration group to that of the first as follows:

$$\theta_{(21)j_2} = A\,\theta_{(2)j_2} + B \tag{2}$$

$$b_{(21)i} = A\,b_{(2)i} + B \tag{3}$$

$$d_{(21)ik} = A\,d_{(2)ik} \tag{4}$$

$$a_{(21)i} = \frac{a_{(2)i}}{A} \tag{5}$$

The "scale" constant A and the "location" constant B are found by minimizing the squared differences between the test characteristic curves associated with anchor items from the first and second calibration groups after transforming person parameters. Let :

$$TCC_{(2)j*} = \sum_{i=1}^{I} \sum_{k=1}^{K_i} k\,[\,P_{(2)ik}(\theta_{(2)j*}\,, b_{(2)i}\,, a_{(2)i}\,, \underline{d}_{(2)i})\,] \tag{6}$$

$$TCC_{(21)j*} = \sum_{i=1}^{I} \sum_{k=1}^{K_i} k\,[\,P_{(21)ik}(\theta_{(21)j*}\,, b_{(1)i}\,, a_{(1)i}\,, \underline{d}_{(1)i})\,] \tag{7}$$

where: I is the number of anchor items,

K_i is the number of response categories for the *ith* item,

$j*$ refers to some preselected evaluation points on the θ continuum,

$P_{(2)ik}$ is the operating characteristic function for common items in the second calibration group, and

$P_{(21)ik}$ is the operating characteristic function for common items in the first calibration group using an evaluation point on the latent continuum that has been transformed from the metric of the second calibration group to that for the first calibration group.

$TCC_{(2)j*}$ is the value of the test characteristic curve at point $j*$ for common items from the second calibration. $TCC_{(21)j*}$ is the value of the test characteristic curve at point $j*$ for common items from the first calibration after the metric of the latent continuum is transformed so that the curve closely matches $TCC_{(2)j*}$ as closely as possible. The TCC method attempts to find the values of the scale constant (A) and location constant (B) that minimize the following squared loss function:

$$Q_{(21)}^{TCC} = \sum_{j*} [\, TCC_{(2)j*} - TCC_{(21)j*} \,]^2 \tag{8}$$

In other words, the method attempts to linearly transform the metric of parameters in the second calibration group so that the TCC for anchor items matches the corresponding TCC in the first calibration group as closely as possible. In this study, $Q_{(21)}^{TCC}$ was minimized with respect to A and B using a Newton-Raphson technique. This minimization technique requires the partial derivatives of $Q_{(21)}^{TCC}$. The necessary derivatives and a description of our implementation of the Newton-Raphson algorithm are given in a technical report available from the authors (Roberts, Huang & Gagne, 2003).

The reader should note that the TCC method described above does not yield a symmetric solution for A and B. If one switches the roles of the first and second calibration groups in the above equations and recalculates the estimates of A and B, then the new solution will not generally be the inverse of the original solution. Thus, the TCC method provides an asymmetric solution. It is also important to note that Stocking and Lord (1983) did not explicitly define what evaluation points (i.e., j*) should be used to minimize $Q_{(21)}^{TCC}$. In practice, a small number of equally-spaced points (e.g., 21) ranging from -4 to +4 are often used (Baker, 1995).

A symmetric version of the TCC method can be developed for the GPCM in a straightforward manner. Relying on Haebara's approach to forming a symmetric quadratic loss function in the item characteristic curve method (see below), we can define the following additional relationships:

$$\theta_{(12)j_1} = \frac{\theta_{(1)j_1} - B}{A} \quad . \tag{9}$$

In Equation 9, the term $\theta_{(1)j_1}$ refers to the location of the j_lth individual from the first calibration group on the latent continuum. In contrast, $\theta_{(12)j_1}$ denotes the location of the j_lth individual from the first calibration group on the latent continuum after transforming it to match the metric of the second calibration group. We can further define the following two TCCs:

$$TCC_{(1)j*} = \sum_{i=1}^{I} \sum_{k=1}^{K_i} k\,[\,P_{(1)ik}(\theta_{(1)j*}\,, b_{(1)i}\,, a_{(1)i}\,, \underline{d}_{(1)i})\,] \tag{10}$$

$$TCC_{(12)j*} = \sum_{i=1}^{I} \sum_{k=1}^{K_i} k\,[\,P_{(12)ik}(\theta_{(12)j*}\,, b_{(2)i}\,, a_{(2)i}\,, \underline{d}_{(2)i})\,] \,. \tag{11}$$

$TCC_{(1)j*}$ is the test characteristic curve at point $j*$ for common items in the first calibration group, whereas $TCC_{(12)j*}$ is the test characteristic curve at point $j*$ for common items in the second calibration group after the metric of the latent continuum is rescaled so that the curve matches $TCC_{(1)j*}$ as closely as possible. The degree of matching can be quantified with the following squared loss function:

$$Q_{(12)}^{TCC} = \sum_{j*} \left[TCC_{(1)j*} - TCC_{(12)j*} \right]^2 . \qquad \textbf{(12)}$$

With these definitions in place, we can develop the following symmetric loss function:

$$Q^{TCC} = Q_{(12)}^{TCC} + Q_{(21)}^{TCC} . \qquad \textbf{(13)}$$

The values of the scale constant, A, and the location constant, B, which minimize this function are found using a Newton-Raphson technique. The necessary derivatives are given in Roberts et al. (2003). Note that because this function minimizes differences between test characteristic curves twice - first by rescaling the metric of the curve in the second calibration group and then by rescaling the metric of the curve in the first group - it consequently yields a symmetric solution for the linking constants.

The Item Characteristic Curve Method

Haebara (1980) proposed an alternative characteristic curve approach in which the linking constants, A and B, are found that minimize the sum of squared differences between item characteristic curves (ICC) associated with common items across two calibration groups. Haebara's approach was also distinct from Stocking and Lord's (1983) method in that it produced a symmetric solution. The ICC method was originally proposed for a 3-parameter logistic model for binary responses, but it can be extended to the GPCM in a straightforward manner. In order to illustrate the squared loss function for the generalized ICC method, let us define the following functions:

$$ICC_{(1)ij*} = \sum_{k=1}^{K_i} k\,[\,P_{(1)ik}(\theta_{(1)j*}, b_{(1)i}, a_{(1)i}, \underline{d}_{(1)i})\,] \tag{14}$$

$$ICC_{(12)ij*} = \sum_{k=1}^{K_i} k\,[\,P_{(12)ik}(\theta_{(12)j*}, b_{(2)i}, a_{(2)i}, \underline{d}_{(2)i})\,]\,. \tag{15}$$

$$ICC_{(2)ij'} = \sum_{k=1}^{K_i} k\,[\,P_{(2)ik}(\theta_{(2)j'}, b_{(2)i}, a_{(2)i}, \underline{d}_{(2)i})\,] \tag{16}$$

$$ICC_{(21)ij'} = \sum_{k=1}^{K_i} k\,[\,P_{(21)ik}(\theta_{(21)j'}, b_{(1)i}, a_{(1)i}, \underline{d}_{(1)i})\,] \tag{17}$$

$ICC_{(1)ij*}$ is the value of the item characteristic curve at point *j** for the *ith* common item in the first calibration group. $ICC_{(2)ij'}$ is the value of the item characteristic curve at point *j'* for the *ith* common item in the second calibration group. $ICC_{(12)ij*}$ is the value of the item characteristic curve at point *j** for the *ith* common item in the second calibration group after the metric of the latent continuum is transformed to match the metric of the first calibration group. Conversely, $ICC_{(21)ij'}$ is the item characteristic curve for the *ith* common item at point *j'* in the first calibration group after the metric of the latent continuum is transformed to match the metric of the second calibration group. The squared loss function in Haebara's ICC method can then be written as:

$$\begin{aligned} Q^{ICC} &= \sum_{i=1}^{I} \left[\sum_{j*} (ICC_{(1)ij*} - ICC_{(12)ij*})^2 + \sum_{j'} (ICC_{(2)ij'} - ICC_{(21)ij'})^2 \right] \\ &= \sum_{i=1}^{I} \left[Q_{(12)i}^{ICC} + Q_{(21)i}^{ICC} \right] \end{aligned} \qquad \textbf{(18)}$$

The values of the scale constant, A, and location constant, B, that minimize Q^{ICC} are found using a Newton-Raphson technique. The derivatives required to perform this minimization are given in Roberts et al. (2003). If the roles of the first and second calibration groups are reversed in Equation 18, then the resulting solution is the inverse of the original solution, and thus, Haebara's solution is symmetric. Haebara originally suggested the evaluation points, j* and j', be based on the distributions of estimated θ values in the first and second calibration groups, respectively. However, any reasonable set of evaluation points could be used, and in the remaining part of this paper, we simply define $j* = j'$.

An asymmetric solution can easily be developed for the ICC method by simply minimizing either the $\sum_{i}^{I} Q_{(12)i}^{ICC}$ or the $\sum_{i=1}^{I} Q_{(21)i}^{ICC}$ term from Equation 18 with respect to A and B. To maintain comparability with the definition of the asymmetric solution in the TCC method, we chose to minimize the latter term:

$$Q_{21}^{ICC} = \sum_{i=1}^{I} \left[\sum_{j'} (ICC_{(2)ij'} - ICC_{(21)ij'})^2 \right] = \sum_{i=1}^{I} Q_{(21)i}^{ICC} . \qquad \textbf{(19)}$$

<u>Operating Characteristic Curve Method</u>

Baker (1993) developed an operating characteristic curve (OCC) method to link IRT parameter estimates. He originally proposed the method for the nominal model where calculation of an ICC or TCC need not be applicable. However, Baker noted that the technique is equally appropriate in the case of graded polytomous responses such as those modeled in the GPCM. Let:

$$OCC_{(2)ikj*} = P_{(2)ik}(\theta_{(2)j*}, b_{(2)i}, a_{(2)i}, \underline{d}_{(2)i}),\ \textit{and} \tag{20}$$

$$OCC_{(21)ikj*} = P_{(21)ik}(\theta_{(21)j*}, b_{(1)i}, a_{(1)i}, \underline{d}_{(1)i})\ . \tag{21}$$

The OCC method finds the scale and location constants, A and B, that minimized the following squared loss function:

$$Q_{(21)}^{OCC} = \sum_{i=1}^{I} \sum_{k=1}^{K_i} \sum_{j^*} \left[OCC_{(2)ikj^*} - OCC_{(21)ikj^*}\right]^2\ . \tag{22}$$

The solution to this minimization problem can be solved using the Newton-Raphson algorithm using the derivatives given in Roberts et al. (2003). Baker (1995) has suggested that a limited number of evaluation points (25 or less) that are equally spaced between $\theta \pm 4$ can be used to evaluate differences between the OCCs in the two calibration groups.

The OCC method as originally proposed by Baker (1993) is obviously asymmetric. However, a symmetric version , can be easily developed if we let:

$$OCC_{(1)ikj*} = P_{(1)ik}(\theta_{(1)j*}, b_{(1)i}, a_{(1)i}, \underline{d}_{(1)i}), \tag{23}$$

$$OCC_{(12)ikj*} = P_{(12)ik}(\theta_{(12)j*}, b_{(2)i}, a_{(2)i}, \underline{d}_{(2)i}), \quad \textit{and} \tag{24}$$

$$Q^{OCC}_{(12)} = \sum_{i=1}^{I} \sum_{k=1}^{K_i} \sum_{j^*} \left[OCC_{(1)ikj^*} - OCC_{(12)ikj^*} \right]^2 . \tag{25}$$

With these definitions in place, a squared loss function can be developed to yield a symmetric solution:

$$\begin{aligned} Q^{OCC} &= \sum_{i=1}^{I} \sum_{k=1}^{K_i} \sum_{j^*} \left[OCC_{(1)ikj^*} - OCC_{(12)ikj^*} \right]^2 + \left[OCC_{(2)ikj^*} - OCC_{(21)ikj^*} \right]^2 \\ &= Q^{OCC}_{(12)} + Q^{OCC}_{(21)} . \end{aligned} \tag{26}$$

Again, this function is minimized with respect to the linking constants, A and B using standard techniques like the Newton-Raphson method. The derivatives of the loss function required to implement the method are given in Roberts et al. (2003).

Simulation Study

A simulation study was performed to examine and compare the behavior of each of the six aforementioned linking methods. In this study, item responses from two independent groups of simulees were generated for the same 20-item test form. These responses were generated to conform with the GPCM and were subsequently used to estimate GPCM parameters separately in each of the two groups. The parameter estimates from the two groups were linked using each of the six linking methods. Linking was based on either the entire set of common test items or on randomly chosen subsets. The nature of the evaluation points chosen to compare characteristic curves was also varied. Each aspect of the simulation is described in detail below.

Method

Item Characteristics

Item parameter estimates from the GPCM published in the 1998 NAEP technical report (Allen, Donoghue & Schoeps, 2001) served as the true item parameters in the simulation. Only the 162 NAEP items with three response categories were used in an effort to simplify the experimental design. Of these initial items, only the 153 items with $| b_i | <2$ were maintained in the item pool. Elimination of items with more extreme locations was done in an effort to assure that each response category would be used at least once for every item. On a given replication, 20 items were randomly sampled from the item pool. Responses to these items were simulated independently in the two groups, and then the sampled items were returned to the pool for subsequent resampling.

Sample Size and Simulee Characteristics

In a given experimental condition, either 300 or 2000 respondents were simulated in each of the respondent groups. The true θ values for these simulees were derived in one of two different ways. In one condition, true θ values were generated from a N(0,1) distribution in both respondent groups. In the other condition, the true θ values were generated from a N(0,1) distribution for the first group, and a N(.5, 1.25^2) distribution for the second group. Given the lack of standard terminology , we will refer to the first condition as a horizontal linking scenario, whereas the second condition will be referred to as the vertical linking scenario. True θ values were independently generated on every replication within each condition.

Calculation of Model Parameter Estimates

Parameter estimates were derived from simulated item responses using the PARSCALE computer program (Version 3.2; Muraki & Bock, 1997). A marginal maximum *a posteriori* estimation algorithm was used to derive item parameter estimates. The algorithm utilized a N(0,1) prior distribution for θ, a log-normal(0,.5^2) prior distribution for slopes, and a N(0,2^2) prior for item locations. Thirty equally-spaced quadrature points between -4 and +4 were used to obtain a solution along with a convergence criterion in which the largest absolute change for any item parameter estimates was less than .001 from one iteration to the next.

After obtaining solutions for item parameters, expected *a posteriori* (EAP) estimates of θ were obtained for each simulee. These estimates were calculated using 30 equally spaced quadrature points between -4 and +4 along with a N(0,1) prior distribution.

Linking Parameter Estimates

The parameter estimates for the two calibration groups were linked using both the symmetric and asymmetric forms of the TCC, ICC and OCC methods. Furthermore, each of these methods was implemented using either 5, 10, 15 or 20 common items. The common items used in the linking procedure were randomly selected at the beginning of each replication. However, items were added to the set of common items in blocks of 5 items so that, within a given replication, the larger sets of common items contained the smaller sets. (In other words, 5 items were added to the original set of 5 common items to produce a set of 10 common items, and so on.) Differences between characteristic curves were evaluated at points on the latent continuum defined by one of three strategies. Curves were contrasted at 50 equally spaced points between

θ=(-4, +4), at 2*N equally spaced points between θ=(-4, +4), or at the 2*N values of θ that were estimated in the two corresponding calibration groups.

Experimental Design

The simulation conditions were structured as a 2 (sample size) x 2 (linking scenario) x 3 (characteristic curve type) x 2 (solution symmetry) x 3 (evaluation points) x 4 (anchor items) split-plot factorial design. The first two of these factors comprised the between-replication conditions, whereas the last four factors were within-replication conditions. There were 100 replications in each between-replication condition. On a given replication, the true item parameters were randomly sampled from the item pool, item responses were generated in each of the two independent groups, GPCM estimates were calculated separately in the two groups, and the parameter estimates were repeatedly linked using the 72 different procedures defined by the within-replication factors.

Data Analyses

Primary and secondary data analyses were performed to assess the accuracy of linking under the conditions that were explored. In the primary analyses, the squared error of the estimated scale constant, $(\hat{A} - A)^2$, and that for the estimated location constant $(\hat{B} - B)^2$ were analyzed using a univariate split-plot ANOVA. Consequently, the hypotheses tested by these ANOVA models were framed in terms of mean squared error (MSE). The nominal Type I error rate was set to .025 when testing each ANOVA effect due to the fact that two dependent measures were studied. The probability of each within-replication effect under the null hypothesis was adjusted using the Hyunh-Feldt (1970) procedure. Each effect in the ANOVA model, was a classified into one of 16 effect families. An effect family included all the effects tested by a given error term along with the error term itself. The proportion of familywise variance associated with each effect was calculated. This is referred to as the η^2_{WF}. In order to limit the interpretation of trivial effects, only those effects that were both statistically significant and associated with $\eta^2_{WF} > .03$ were interpreted in this paper.

The secondary data analyses explored the squared deviation of GPCM parameter estimates from their true values using ANOVA techniques. Thus, the hypotheses for each ANOVA effect was framed in terms of mean squared deviations (MSD). There were two basic types of ANOVA models constructed to examine MSD. First, the squared error of GPCM parameter estimates (i.e., $\hat{b}$, $\hat{a}$, $\hat{d}$, and $\hat{\theta}$) for the first calibration group was modeled as a function of sample size, linking scenario and their interaction. A second ANOVA model was used to investigate the squared error of parameter estimates from the second calibration group after they were rescaled to the metric of the first calibration group. These squared errors were examined using an ANOVA model which included all the between-replication and within-replication effects in the experimental design. The Type I error rate for the secondary data analyses was set to .0125 because there were four dependent measures examined in each type of ANOVA model. As in the primary analyses, probability values for within-replication effects were adjusted using the Huynh-Feldt procedure. Given that the total number of effects tested in these secondary analyses was substantially larger than that for the primary analyses, the effect size criterion for interpretation was increased to $\eta^2_{WF} > .05$. Thus, only those effects that were statistically significant and had corresponding $\eta^2_{WF} > .05$ were ultimately interpreted.

Results

Primary Analyses

Table 1 lists the ANOVA effects for the MSE of $\hat{A}$ that were deemed to be interpretable on the basis of the previously defined operational criteria. Of these interpretable effects, some were main effects that were subsumed under higher order interactions. Only the higher order effects will be described in this paper. The MSE for $\hat{A}$ was partially a function of a two-way interaction between the number of anchor items used to estimate linking constants and the sample size used to calibrate GPCM parameters. The form of this interaction is shown in the top panel of Figure 1. When parameter estimates were calibrated using large samples (*N*=2000), the MSE for $\hat{A}$ was minimal, and the number of anchor items had little effect on the MSE values. However, when the calibration sample size was small (*N*=300), then the MSE for $\hat{A}$ increased noticeably; and this increase was inversely related to the number of anchor items used in the linking solution. The mean differences associated with this interaction were quite small and were confined to the third decimal place. However, these differences were not considered to be ignorable. For example, the difference between the root of the MSE incurred in the two sample size conditions with 5 anchor items was equal to .058, and this represented 5.8% of the true θ standard deviation. For the 20 anchor item condition, the difference in root MSE decreased to .032.

Insert Table 1 About Here

Insert Figure 1 About Here

The interaction between linking scenario and calibration sample size was also classified as interpretable. This interaction is illustrated in the bottom panel of Figure 1. With a large calibration sample size, the MSE of $\hat{A}$ was quite small, and the linking scenario (i.e., horizontal or vertical linking) had virtually no effect on its magnitude. In contrast, when the calibration sample size was small, the MSE of $\hat{A}$ increased for both linking scenarios, but this increase was larger in the vertical linking condition. Again, this effect was small, but not ignorable. The difference in root MSE values for $\hat{A}$ in the vertical linking condition was approximately .05 which was equal to 5% of the standard deviation of true θ values, whereas the difference in root MSE for the horizontal linking condition was equal to .032.

There was also a three-way interaction between type of characteristic curve utilized, the symmetry of the solution, and calibration sample size. This interaction is displayed in Figure 2. The MSE for $\hat{A}$ was trivial and showed little sensitivity to the type of characteristic curve or the symmetry of the solution when a large calibration sample size was used. However, when a small calibration sample size was used, then the MSE for $\hat{A}$ increased, and a small interaction between the type of characteristic curve used and the symmetry of the solution emerged. The TCC approach produced the smallest amount of error, followed by the ICC and OCC methods, respectively. With the latter two methods, there appeared to be a small effect of symmetry such that a symmetric solution produced a slightly smaller MSE for $\hat{A}$ as compared to an asymmetric solution. The tiny magnitude of this interaction was corroborated by its sum of squares which was nearly zero. (As shown in Table 1, the sum of squares for this interaction rounded to zero at the third decimal place.) Given its very small magnitude, this interaction is probably of little practical importance.

Insert Figure 2 About Here

Table 2 lists the ANOVA effects corresponding to the MSE for $\hat{B}$ that were deemed to be interpretable. As was the case for $\hat{A}$, the MSE for $\hat{B}$ was a function of the two-way interaction between the number of anchor items used in the linking solution and the sample size used to calibrate GPCM parameter estimates. The means corresponding to this interaction are shown in the top panel of Figure 3. The pattern of means is quite similar to that found with the MSE for $\hat{A}$, and suggested that the MSE $\hat{B}$ was trivial in magnitude and insensitive to the number of anchor items used in the linking solution when the calibration sample size was large. With a small calibration sample size, the MSE for $\hat{B}$ increased, but the accuracy of $\hat{B}$ improved as the number of anchor items grew larger. The difference in the root MSE for $\hat{B}$ between the sample size conditions was equal to .037 when the number of anchor items was equal to 5. This difference decreased to .022 when the number of anchor items was 20.

Insert Table 2 About Here

Insert Figure 3 About Here

The bottom panel of Figure 3 illustrates another two-way interaction in which the MSE for $\hat{B}$ was a function of the type of characteristic curve used in the solution and the nature of the evaluation points used to contrast characteristic curves. When either 50 or 2*N* equally spaced evaluation points were used, then the TCC method produced slightly more accurate estimates of $\hat{B}$ relative to the ICC and OCC methods. The latter two methods produced similar MSE values. In contrast, when the characteristic curves were evaluated at the 2*N* points associated with $\hat{\boldsymbol{\theta}}$, then all the characteristic curve methods yielded similar MSE values and these values were slightly smaller than those in other evaluation point conditions. Although this effect was interesting, it was so small that it would have few, if any, pragmatic consequences.

Secondary Analyses

Group 1 Analyses. The two-way ANOVA on the squared difference of GPCM parameters for Group 1 revealed intuitive results. The main effect of calibration sample size on MSD led to η^2_{WF} values equal to .63, .65 and .52 for estimates of $\hat{b}$, $\hat{a}$, and $\hat{d}$, respectively. These sample size effects were all statistically significant at the p<.0001 level and all were in the predicted direction; larger sample sizes led to smaller MSDs (.003 versus .017 for $\hat{b}$, .004 versus .021 for $\hat{a}$, and .007 versus .043 for $\hat{d}$). These differences were very small in the metric of MSD. However, when differences were expressed in terms of root MSD, they appeared more substantial (.052 versus .131 for $\hat{b}$, .059 versus .144 for $\hat{a}$, and .083 versus .208 for $\hat{d}$). Moreover, when these differences were interpreted in the light of the root pooled variance of true item parameters within replications (.659 for $\hat{b}$, .293 for $\hat{a}$, and .937 for $\hat{d}$) they seem more substantial. Specifically, the differences in root MSD represented 11.99%, 29.01%, 13.34% of the root pooled within-replication variance of true b, a, d parameters, respectively.

The MSD for $\hat{\theta}$ in Group 1 was not statistically related to calibration sample size, linking scenario or their interaction. This was expected because the number of responses to informative items determines the accuracy of an EAP estimate, and the $\hat{\theta}$ were calculated from 20 item responses in all conditions. Additionally, the $\hat{\theta}$ values for Group 1 were not rescaled via the linking constants, and therefore, one would not necessarily expect their accuracy to be dependent on the linking scenario if the test provided ample information across the latent continuum.

Group 2 Analyses. The ANOVA on the squared difference for rescaled GPCM item parameter estimates from Group 2 revealed several interesting findings. Table 3 lists the effects that were deemed to be interpretable. With regard to the MSD for $\hat{b}$, there was a two-way interaction between calibration sample size and the number of anchor items used to solve for linking coefficients. This interaction is depicted in the upper panel of Figure 4. As shown in the figure, when the calibration sample size was large, the MSD for $\hat{b}$ was minimal and insensitive to the number of anchor items used to solve for linking constants. In contrast, when the calibration sample size was small, the MSD for $\hat{b}$ increased, and slight reductions in MSD emerged as the number of anchor items increased. The difference in root MSD for $\hat{b}$ between the two sample size conditions was equal to .084 when linking was based on 5 anchor items, and it decreased to .072 when 20 anchor items were used. These differences represented 12.7% and 11.0% of the root pooled variance of true b values within replications. Thus, the increase in MSD due to small calibration sample size was mitigated very slightly by an increase in anchor items. It is also important to note that the MSD for $\hat{b}$ in the small calibration condition decreased very little after the number of anchor items reached 15.

Insert Table 3 About Here

Insert Figure 4 About Here

The ANOVA for the MSD of $\hat{b}$ also revealed an interpretable interaction between the nature of evaluation points and the number of anchor items used in the linking solution. This effect is illustrated in the bottom panel of Figure 3. The interaction suggested that using estimated theta values from both calibration groups as points to evaluate characteristic curve differences was slightly preferable when a small number anchor items was employed. However with 15 or more anchor items, using theta estimates as evaluation points produced slightly higher MSD values for $\hat{b}$. Although this interaction met the criteria for interpreting an effect, the corresponding mean differences were negligible as suggested by the nearly zero sum of squares for this effect as shown in Table 3.

Insert Figure 5 About Here

A three-way interaction of the characteristic curve, the symmetry, and the evaluation points inherent in a given solution also emerge when evaluating MSD for $\hat{b}$. The interaction is portrayed in Figure 5. When equally spaced evaluation points were used, then the symmetric or asymmetric nature of the solution had little effect on the MSD for $\hat{b}$ unless an ICC method was used. In this case, an asymmetric solution seemed slightly more preferable. However, a symmetric solution led to slightly smaller MSD for the ICC and OCC methods when $\hat{\theta}$ values served as evaluation points. The symmetry of a solution had little impact of the MSD for $\hat{b}$ when the TCC method was used, regardless of the nature of evaluation points. Given the very small size of these mean differences and the small sum of squares associated with the interaction, it is doubtful that this finding will have any pragmatic consequences for practitioners.

With regard to the MSD for $\hat{a}$, there were two ANOVA effects that met the criteria for interpretation. The first of these was a two-way interaction between calibration sample size and number of anchor items which is shown in Figure 6. As was the case with $\hat{b}$, the MSD for $\hat{a}$ was negligible when the calibration sample size was large, regardless of the number of anchor items used in the linking solution. However, for small calibration samples, the MSD for $\hat{a}$ was noticeably larger, and it decreased very slightly as the number of anchor items increased. These decreases in MSD attenuated when the number of anchor items was 15 or more. The difference in root MSD for $\hat{a}$ between the two sample size conditions was equal to .096 when linking was based on 5 anchor items. The difference decreased to .081 when 20 items were used in the linking procedure. These values represented 32.9% and 27.6% of the root pooled variation in true a values within replications. Thus, the noticeable ill effects of a small calibration sample size were mitigated slightly by an increase in the number of anchor items used in the linking solution.

Insert Figures 6 and 7 About Here

The MSD for $\hat{a}$ was also dependent on a three-way interaction between the calibration sample size, the type of characteristic curve employed, and the symmetric nature the solution. This interaction is shown in Figure 7. With large calibration samples, the MSD for $\hat{a}$ was quite small, and the remaining two factors had virtually no impact on its value. With small calibration samples, the MSD for $\hat{a}$ increased noticeably. Additionally, there was a very slight advantage for the symmetric solution when using either an ICC or OCC method, as opposed to the TCC method. However, the size of this simple interaction effect was extremely small and is probably of no practical importance.

The MSD for $\hat{d}$ was primarily a function of the calibration sample size and the number of anchor items used to solve for linking constants. Both of these main effects are displayed in Figure 8. As shown in the top panel of the figure, the large calibration sample led to an MSD for $\hat{d}$ that was .032 smaller than that in the small calibration sample condition. The difference in the corresponding root MSD values was equal to .115 and represented 12.2% of the root pooled variation in true d parameters within replications. Therefore, this effect was considered to be small, but still meaningful.

Insert Figure 8 About Here

The bottom panel of Figure 8, illustrates the main effect of the number of anchor items in the linking solution on the MSD for $\hat{a}$. The MSD decreased as the number of anchor items increased, although the size of the decrease diminished once the number of anchor items had reached 15 or more. The MSD for $\hat{a}$ decreased by .002 when the number of anchor items increased from 5 to 10. The corresponding difference in root MSD values was also quite small (.006), and thus, this effect was thought to have little, if any, practical implications.

With regard to the MSD for rescaled $\hat{\theta}$ values from Group 2, there were several interpretable effects that emerged from the ANOVA, and each of these is listed in Table 4. The strongest effect was the main effect of linking scenario. The MSD for rescaled $\hat{\theta}$ values was larger in the vertical linking scenario (.144) relative to the horizontal linking condition (.128). When transformed into root MSD, the magnitude of this difference was small and represented 2.2% of the standard deviation of true θ.

Insert Table 4 About Here

Insert Figure 9 About Here

The MSD for $\hat{\theta}$ was also dependent on the type of characteristic curved used to estimate linking constants. As shown in the top panel of Figure 9, the TCC method led to MSD values that were smaller than that for the ICC and OCC methods. However, given the very small size of these differences, their impact was deemed negligible.

There was also a two-way interaction between calibration sample size and number of anchor items used in the linking solution. This interaction is shown in the bottom panel of Figure 9. When the sample size was large, the number of anchor items used to estimate linking constants had little effect on the accuracy of the rescaled $\hat{\theta}$. However, the MSD for $\hat{\theta}$ increased slightly in the small sample size conditions, and these increases were mitigated as the number of anchor items increased. When 5 anchor items were utilized, the difference in root MSD between sample size conditions was equal to .013. This difference fell to .003 when 20 anchor items were used in the linking process. Again, these differences were very small, and thought to have little practical importance.

Discussion

Results from both the primary and secondary analyses suggested several convergent findings. First, all the differences found among various linking conditions studied here were small. In many cases, the differences were extremely tiny and thought to have little pragmatic implications. This may be due to the fact that the simulated data fit the GPCM, and thus, reasonable parameter estimates were obtained in all between-replication conditions. If accurate parameter estimates are available, then linking parameters can be easily recovered and there is less opportunity for linking methods to differ (Cohen & Kim, 1998). Nonetheless, when any new method is introduced, the method must first be shown to work well in ideal situations before exploring its characteristics under more realistic conditions. In this regard, the accuracy of linking achieved with methods that have been generalized or created in this study was consistently reasonable.

Although all the effects discovered in the simulation were generally quite small, there were, nonetheless, consistent findings that emerged. As the calibration sample size increased, the accuracy of the linking methods increased, and sensitivity to variables inherent in a specific linking situation had little, if any, effect on the resulting scale transformation. Again, this was presumably do to the fact that IRT item parameters were estimated very well with large samples. When one has accurate item parameter estimates of anchor items in two groups, then finding the linear transformation that relates the metrics of the two groups is relatively easy and can be accomplished accurately using a variety of solution schemes.

When the calibration sample size was small, the accuracy of the linking procedure degraded somewhat, and other factors began to affect the quality of the linking results ever so slightly. The most important of the factors studied here was the number of anchor items used to estimate linking constants. There was noticeable improvement in linking accuracy when the number of anchor items was increased from 5 to 10 items in the small calibration condition. These improvements attenuated once the number of linking items reached 15. The generality of this result is no doubt highly dependent on the particular tests that are calibrated in the two respondent groups. In our simulation study, the item pool consisted of all non-extreme 1998 NAEP items with 3 response categories. One could conjecture that between 10-15 anchor items will be needed to achieve optimal linking results for similar item pools. Unfortunately, determining the similarity between a given item pool and the one used in this study is fraught with difficulty, and thus, further study is required before specific recommendations can be made.

There was also an effect of the linking scenario on the accuracy of estimated linking constants when small calibration sample sizes were used. In this case, the scale constant estimated by the linking procedure was more accurate when the distributions of the two respondent groups were identical. This suggests that achieving a common metric for item parameters may be more difficult in vertical testing situations when the calibration sample size is small. Again, when the sample size is large, then this issue is of little concern.

The type of characteristic curves used to develop linking estimates and the points at which the curves are evaluated had systematic, but trivial effects on the accuracy of linking. The TCC method generally produced linking results that were slightly more accurate then the ICC method, which, in turn, produced slightly more accurate results when compared to the OCC method. However, the impact of these differences will probably have negligible impact in practice. Nonetheless, if one is forced to choose among these methods based only on the results of this simulation, then the choice seems clear. The TCC method consistently provided equal or better linking accuracy in every condition studied here. The choice between the type of evaluation points to use is a bit less clear. Again, the effects of this variable were very small and of little practical importance. However, evaluation of characteristic curve differences at $\hat{\theta}$ points usually led to more accurate linking. The fact that this advantage was quite small may, itself, be practically meaningful because linking is sometimes required in cases where there in no access to $\hat{\theta}$ (e.g., linking item parameters obtained from different published studies). In such cases, a limited number of equally spaced evaluation points should suffice as long as the item parameters are estimated well.

One of the primary contributions of this work was the introduction of alternative characteristic curve methods that allow the practitioner to choose between a symmetric or asymmetric solution regardless of the type of characteristic curve that is employed to link parameter estimates. The simulation showed that use of a symmetric linking procedure generally led to equal or slightly better linking accuracy relative to an asymmetric procedure. Again, these differences were so slight that they seemed trivial. We find the comparable accuracy of symmetric and asymmetric solutions encouraging because we believe that the characteristics of the measurement situation should determine the method used to link parameter estimates. It is comforting to know that this choice can be made on substantive, rather than methodological, grounds with little, if any, impact on linking accuracy.

As with any simulation study, the particular results that have emerged here require further investigation before they can be adequately generalized. Future research should include an assessment of the robustness of these linking methods in situations where the data are systematically misfit by the GPCM or when the assumptions of the GPCM fail to hold (e.g., minor deviations from unidimensionality and/or respondent independence). Perhaps further distinctions among these linking strategies can be made under less ideal circumstances than those simulated here. Future research should also explore how the number of response categories affects the number of anchor items required to produce adequate linking accuracy.

References

Allen, N. L., Donoghue, J. R., & Schoeps, T. L. (2001), *The NAEP 1998 technical report.* Washington, DC: NCES.

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

Baker, F. B. (1995). EQUATE computer program Version 2.1 user documentation. Madison, WI: University of Wisconsin.

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

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

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

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

Huynh, H., & Feldt, L. S. (1970). Conditions under which mean square ratios in repeated measurement designs have exact F-distributions. *Journal of the American Statistical Association, 65,* 1582-1589.

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

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: IRT item analysis and test scoring for rating-scale data*. Chicago: Scientific Software.

Roberts, J. S., Huang, C., & Gagne, P. (2003). *Solving for linking constants in the generalized partial credit model using alternative characteristic curve methods.* Technical report in preparation.

Stocking, M. L., & Lord, F. M. (1983). Developing a common metric in item response theory. *Applied Psychological Measurement, 7*, 201-210.

Author Notes

This material is based upon work supported by the National Science Foundation under Grant No. 0133019. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

Further information can be obtained from James S. Roberts, Department of Measurement, Statistics & Evaluation, University of Maryland, 1230F Benjamin Building, College Park, Maryland 20742. E-mail: jr245@umail.umd.edu.

Table 1. Interpretable ANOVA effects on MSE for $\hat{A}$.

Source	df_{num}	df_{den}	SS	F	p^a	η^2_{WF}
N	1	396	.102	145.47	.0001	.249
LS	1	396	.017	24.53	.0001	.042
N*LS	1	396	.013	18.81	.0001	.032
A	3	1188	.047	59.41	.0001	.119
A*N	3	1188	.029	36.84	.0001	.074
CC	2	792	.001	14.06	.0001	.033
S*CC	2	792	.000	17.42	.0001	.039
S*CC*N	2	792	.000	14.80	.0001	.033

N=calibration sample size, LS=linking scenario, CC=characteristic curve method, S=symmetry of solution, A=number of anchor items.

[a] p-values for within-replication effects are adjusted with the Huynh-Feldt procedure.

Table 2. Interpretable ANOVA effects on MSE for $\hat{B}$

Source	df_{num}	df_{den}	SS	F	p^a	η^2_{WF}
N	1	396	.027	119.09	.0001	.229
A	3	1188	.013	65.22	.0001	.133
A*N	3	1188	.006	28.79	.0001	.059
CC	2	792	.000	13.79	.0001	.033
CC*EP	4	1584	.000	16.72	.0001	.039

N=calibration sample size, A=number of anchor items, CC=characteristic curve method, EP=nature of evaluation points.

[a] p-values for within-replication effects are adjusted with the Huynh-Feldt procedure.

Table 3. Interpretable ANOVA effects on MSD for item parameter estimates.

Source	Parameter	df_{num}	df_{den}	SS	F	p[a]	η^2_{WF}
N	$\hat{b}$	1	396	1.413	640.16	.0001	.613
A	$\hat{b}$	3	1188	.038	88.96	.0001	.167
A*N	$\hat{b}$	3	1188	.020	46.29	.0001	.087
A*EP	$\hat{b}$	6	2376	.000	11.92	.0002	.531
CC*S	$\hat{b}$	2	792	.000	23.49	.0001	.054
CC*S*EP	$\hat{b}$	4	1584	.000	24.96	.0001	.056
N	$\hat{a}$	1	396	2.229	877.88	.0001	.670
A	$\hat{a}$	3	1188	.046	67.79	.0001	.132
A*N	$\hat{a}$	3	1188	.032	46.86	.0001	.092
CC*S	$\hat{a}$	2	792	.000	30.56	.0001	.068
CC*S*N	$\hat{a}$	2	792	.000	24.45	.0001	.054
N	$\hat{d}$	1	396	7.582	426.79	.0001	.518
A	$\hat{d}$	3	1188	.042	33.51	.0001	.074

N=calibration sample size, A=number of anchor items, CC=characteristic curve method, EP=nature of evaluation points, S=symmetry of solution.

[a] p-values for within-replication effects are adjusted with the Huynh-Feldt procedure.

Table 4. Interpretable ANOVA effects on MSD for $\hat{\theta}$.

Source	df_{num}	df_{den}	SS	F	p^a	η^2_{WF}
LS	1	396	1.897	147.06	.0001	.265
A	3	1188	.101	90.31	.0001	.165
A*N	3	1188	.056	49.92	.0001	.091
CC	2	792	.001	19.04	.0001	.044

LS=linking scenario, CC=characteristic curve method, A=number of anchor items, N=calibration sample size .

[a] p-values for within-replication effects are adjusted with the Huynh-Feldt procedure.

Figure Captions

Figure 1. Two-way interactions for the MSE of $\hat{A}$. The top panel illustrates the calibration sample size by number of anchor items interaction. The bottom panel illustrates the calibration sample size by linking scenario interaction.

Figure 2. Three-way interaction of characteristic curve by symmetry of solution by calibration sample size on the MSE of $\hat{A}$.

Figure 3. Two-way interactions for the MSE of $\hat{B}$. The top panel illustrates the calibration sample size by number of anchor items interaction. The bottom panel depicts the characteristic curve by nature of evaluation points interaction.

Figure 4. Two-way interactions for the MSD of $\hat{b}$. The top panel illustrates the calibration sample size by number of anchor items interaction. The bottom panel depicts the nature of evaluation points by number of anchor items interaction.

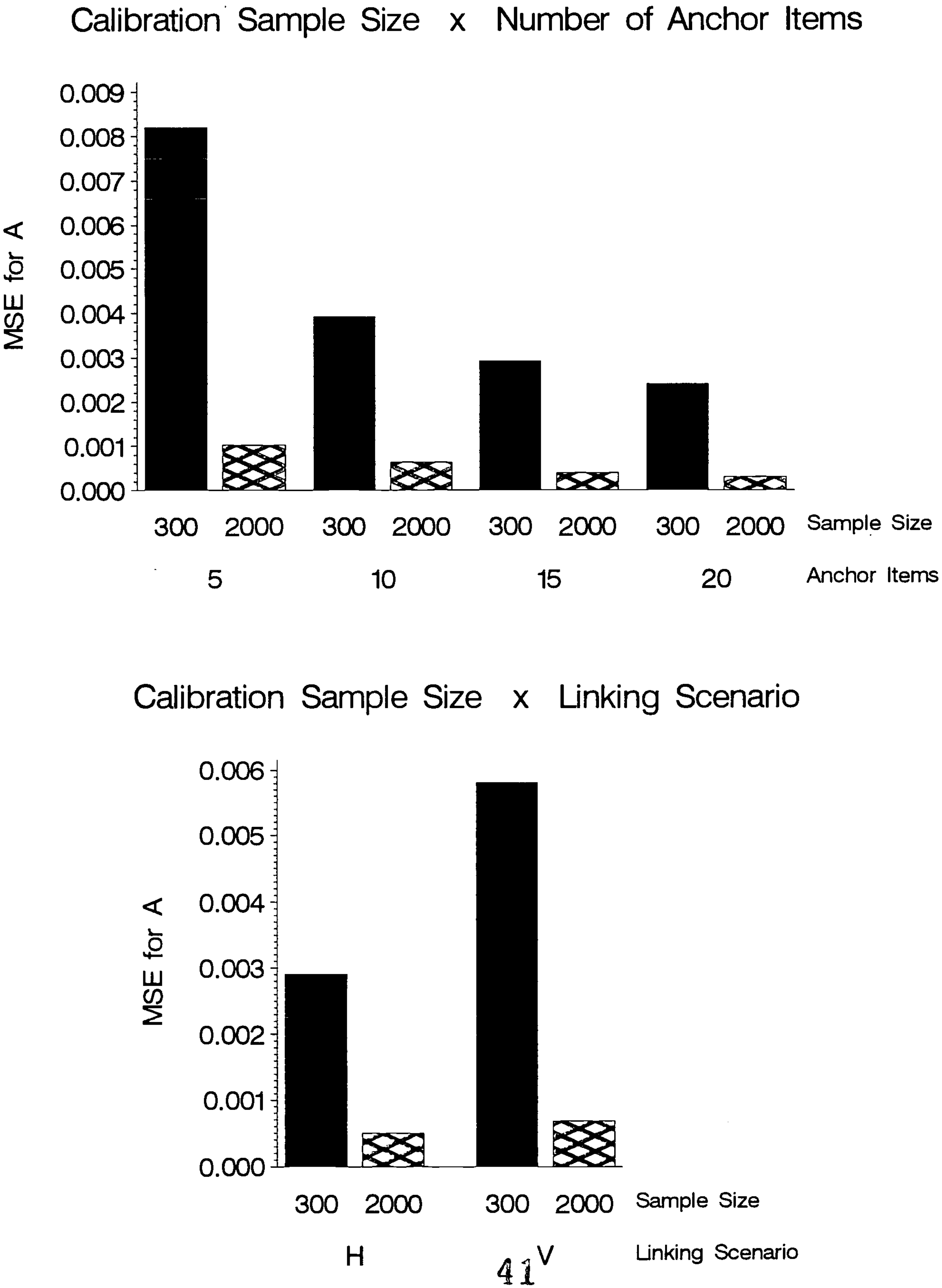
Figure 5. Three-way interaction of characteristic curve by symmetry of the solution by nature of evaluation points on the MSD of $\hat{b}$.

Figure 6. Two-way interaction of calibration sample size by number of anchor items on the MSD of $\hat{a}$.

Figure 7. Three-way interaction of characteristic curve by symmetry of the solution by calibration sample size on the MSD of $\hat{a}$.

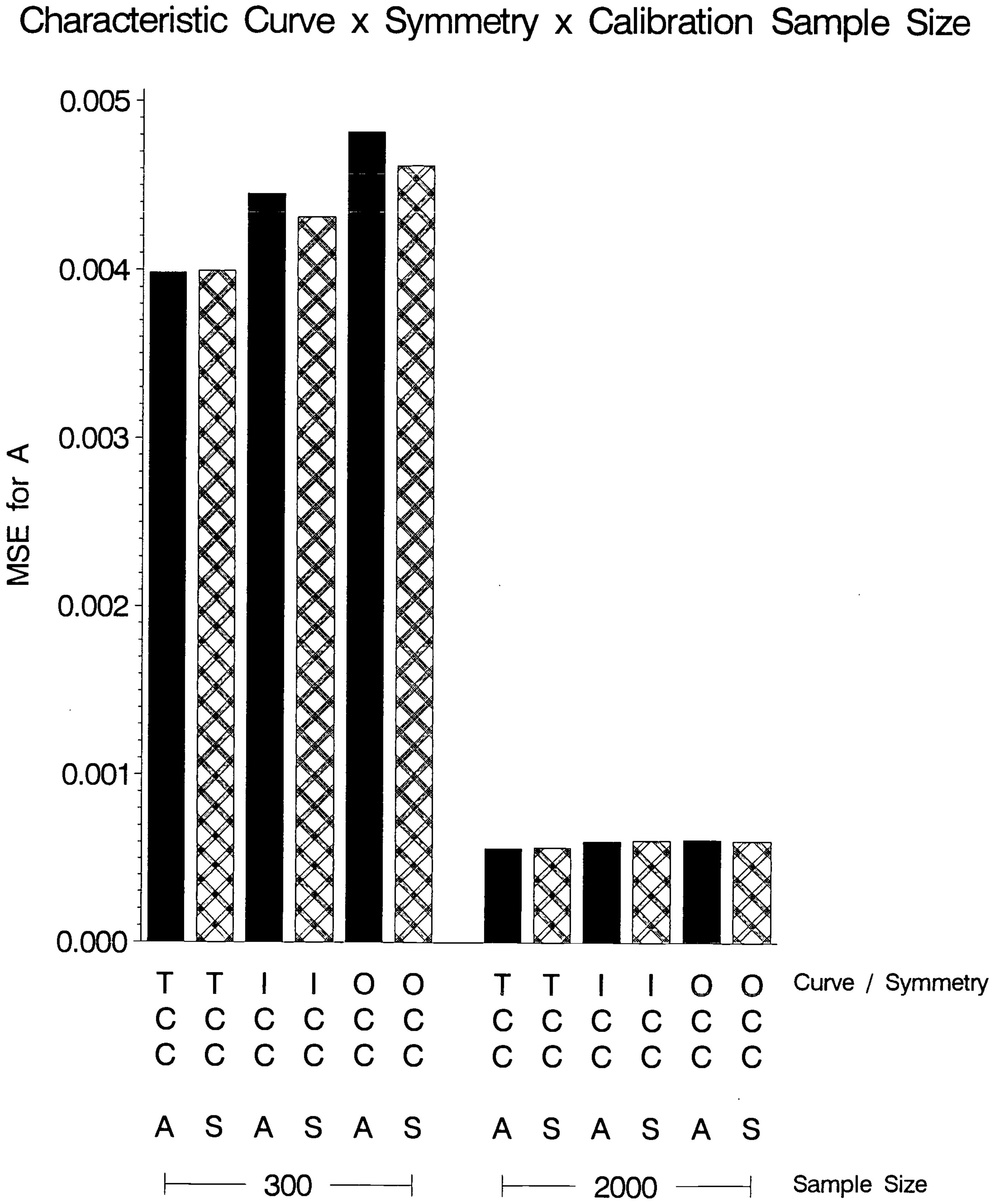
Figure 8. Main effects for the MSD of $\hat{a}$. The top panel illustrates the main effect of calibration sample size. The bottom panel portrays the main effect of the number of anchor items.

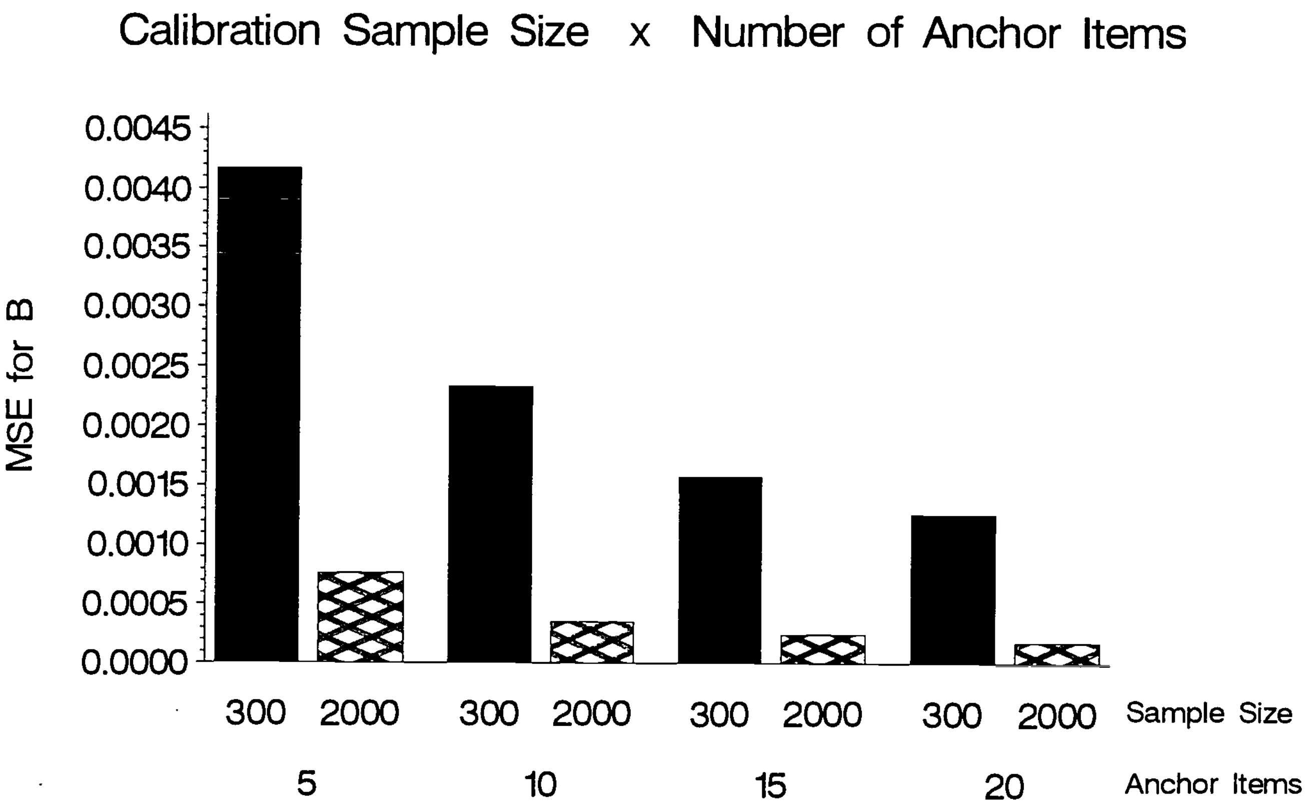
Figure 9. Interpretable effects for the MSD of $\hat{\theta}$. The top panel illustrates the main effect of characteristic curve. The bottom panel depicts the calibration sample size by number of anchor items interaction.

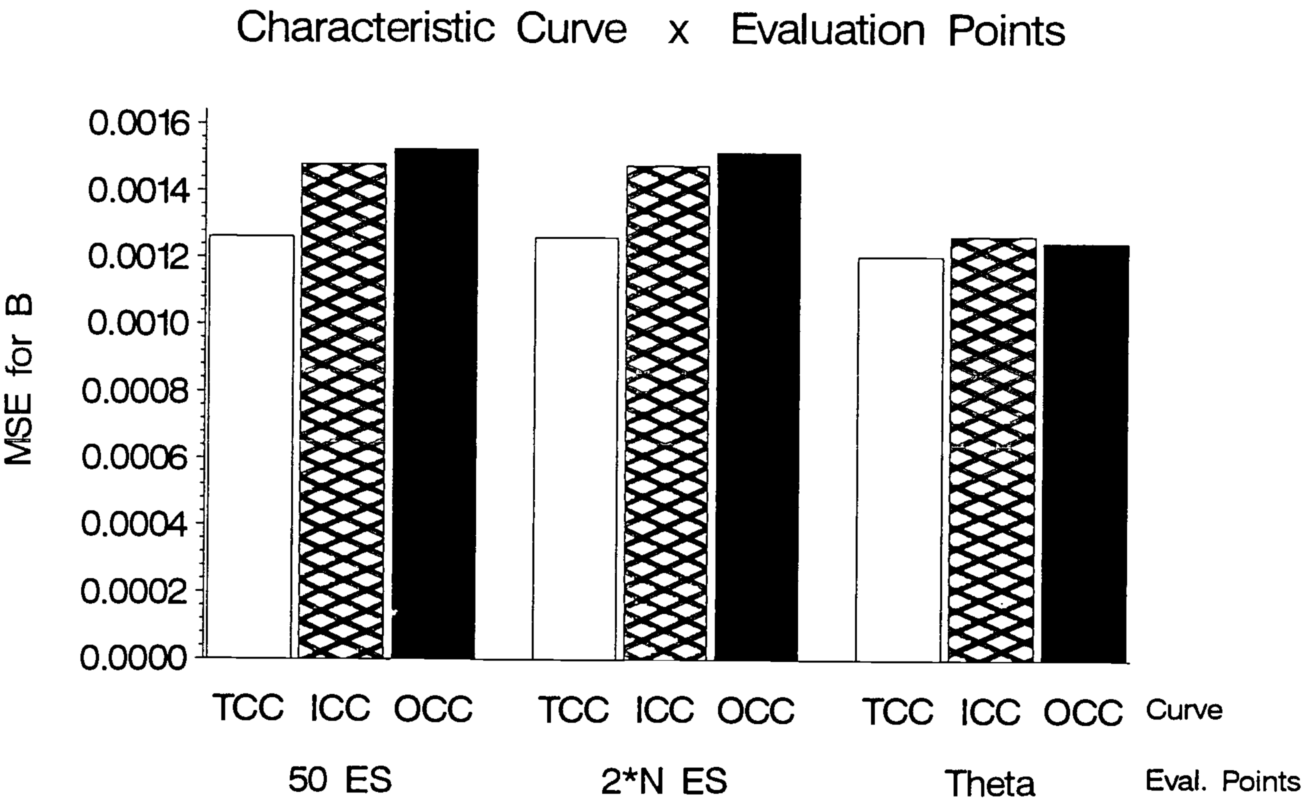


Calibration Sample Size x Number of Anchor Items

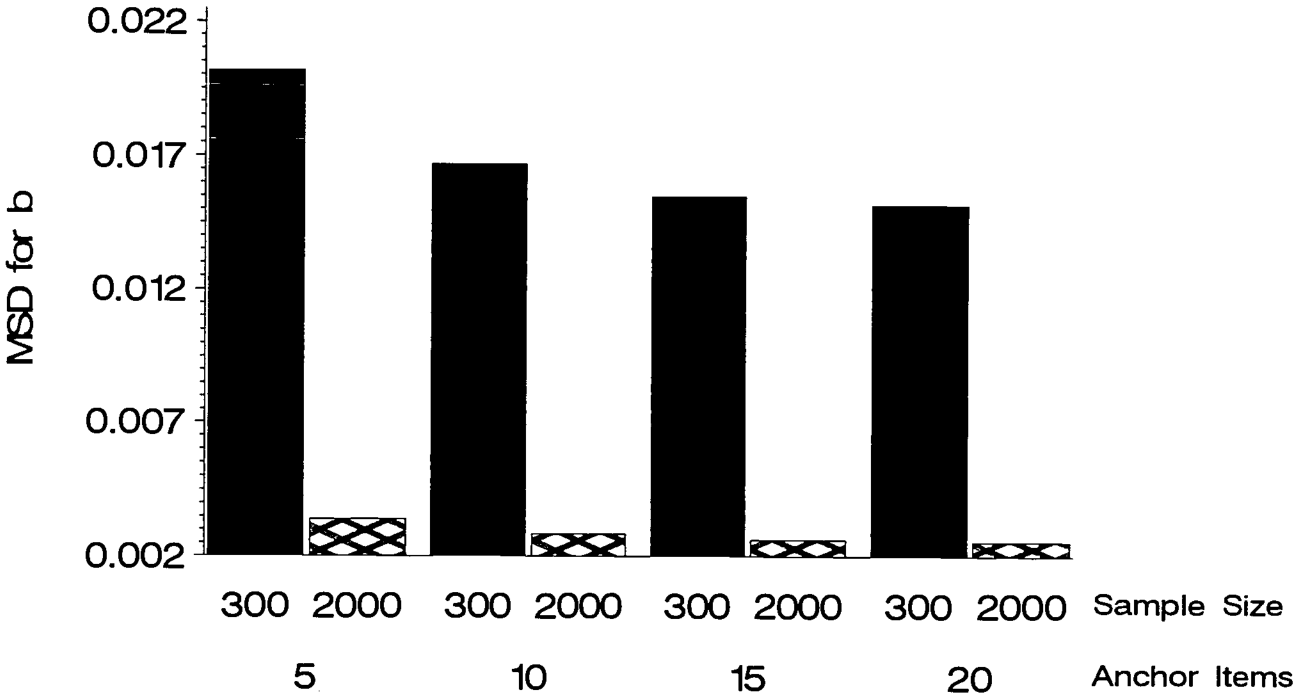
Calibration Sample Size x Linking Scenario


Characteristic Curve x Symmetry x Calibration Sample Size
MSE for A
0.005
0.004
0.003
0.002
0.001
0.000
T T I I O O
C C C C C C
C C C C C C
A S A S A S
300
T T I I O O
C C C C C C
C C C C C C
A S A S A S
2000
Curve / Symmetry
Sample Size

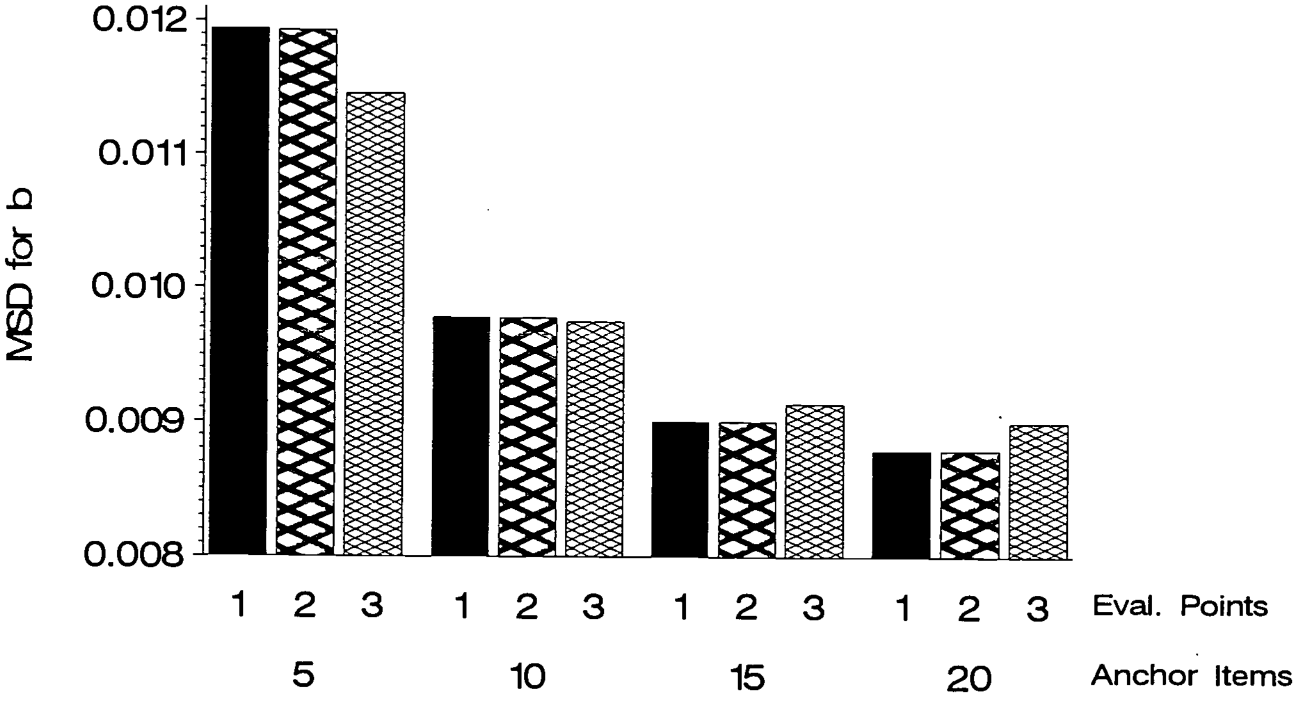

Calibration Sample Size x Number of Anchor Items
MSE for B
0.0045
0.0040
0.0035
0.0030
0.0025
0.0020
0.0015
0.0010
0.0005
0.0000
300
2000
300
2000
300
2000
300
2000
Sample Size
5
10
15
20
Anchor Items

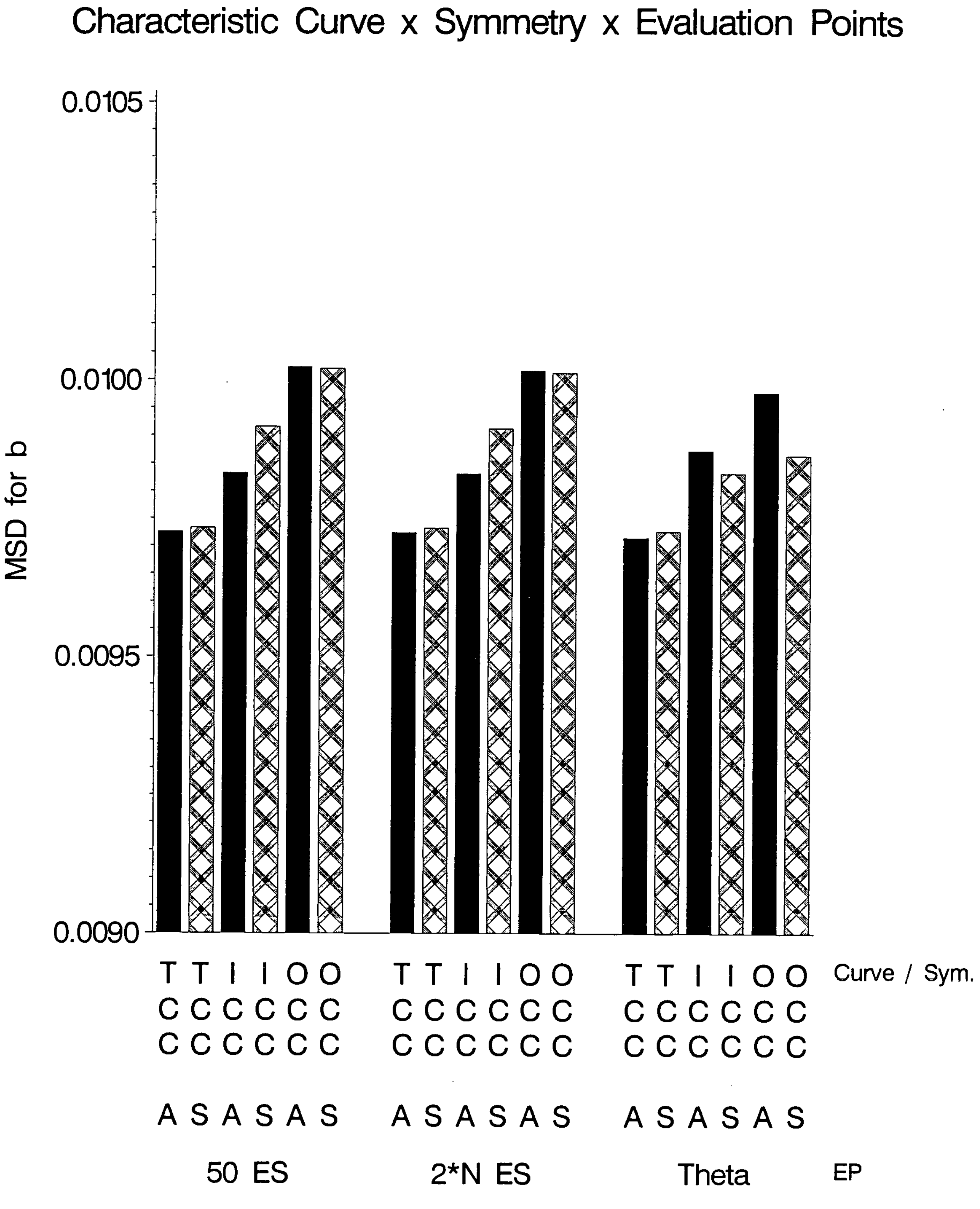

Characteristic Curve x Evaluation Points
MSE for B
0.0016
0.0014
0.0012
0.0010
0.0008
0.0006
0.0004
0.0002
0.0000
TCC
ICC
OCC
TCC
ICC
OCC
TCC
ICC
OCC
Curve
50 ES
2*N ES
Theta
Eval. Points

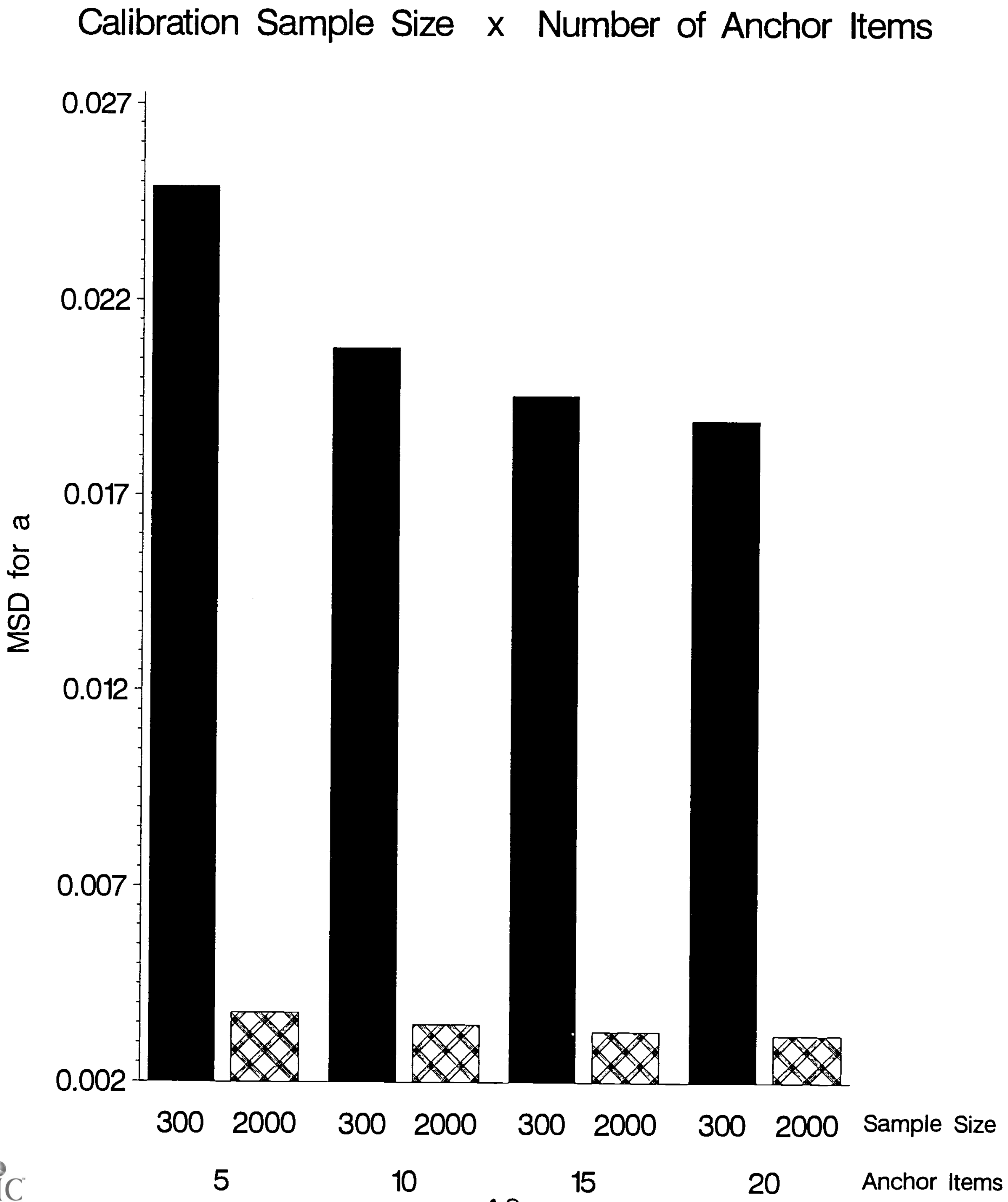
Calibration Sample Size x Number of Anchor Items



Evaluation Points x Number of Anchor Items




Characteristic Curve x Symmetry x Evaluation Points
MSD for b
0.0105
0.0100
0.0095
0.0090
T T I I O O
C C C C C C
C C C C C C
A S A S A S
50 ES
T T I I O O
C C C C C C
C C C C C C
A S A S A S
2*N ES
T T I I O O
C C C C C C
C C C C C C
A S A S A S
Theta
Curve / Sym.
EP


Calibration Sample Size x Number of Anchor Items
MSD for a
0.027
0.022
0.017
0.012
0.007
0.002
300
2000
300
2000
300
2000
300
2000
Sample Size
5
10
15
20
Anchor Items

Characteristic Curve x Symmetry x Calibration Sample Size

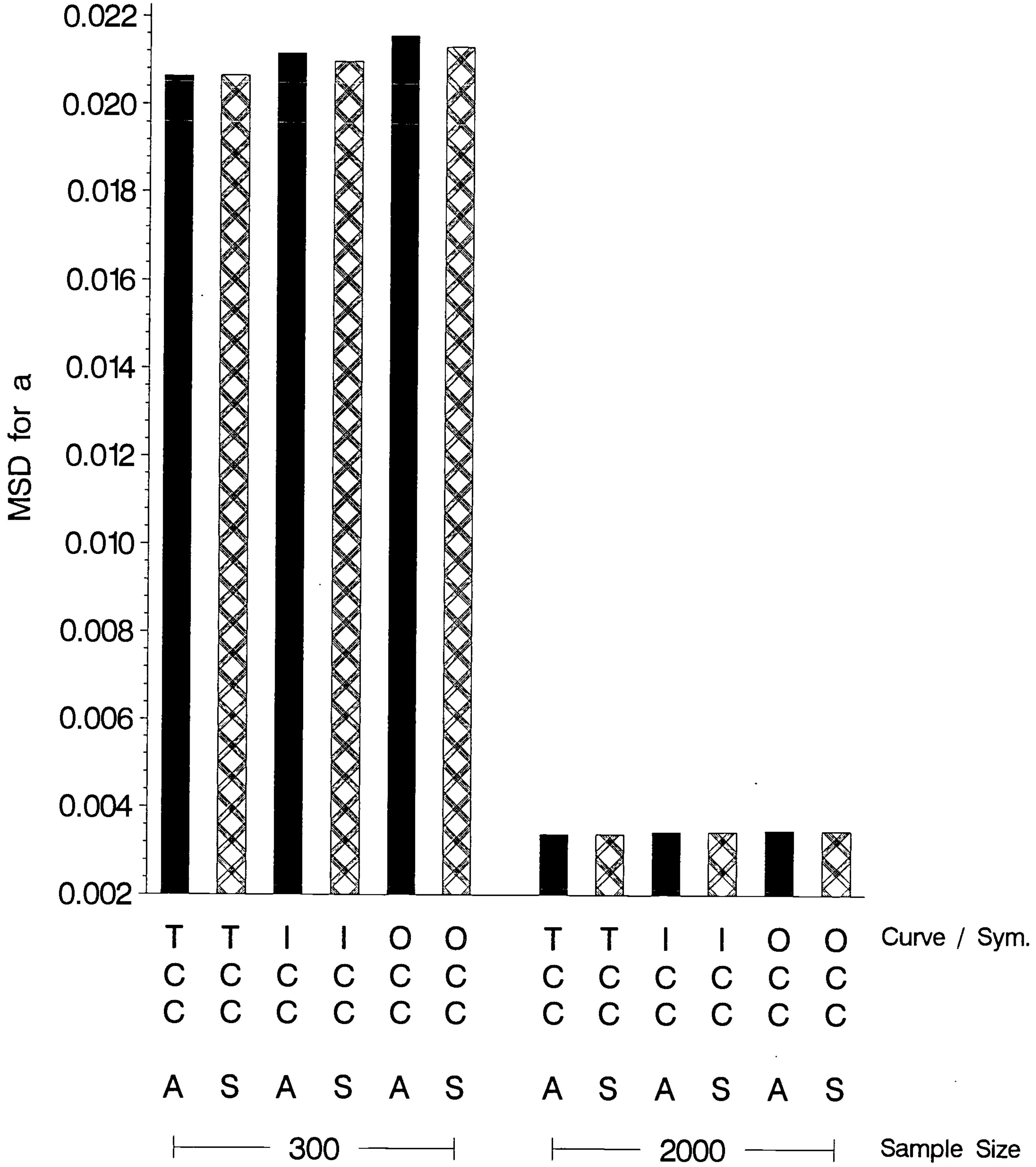
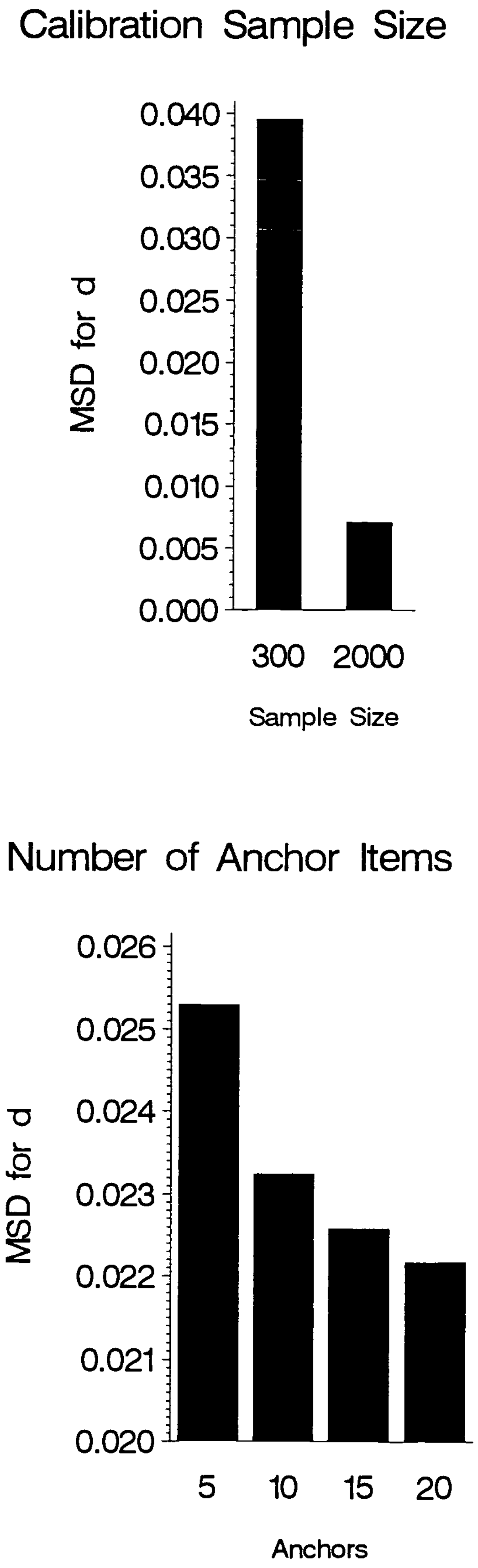
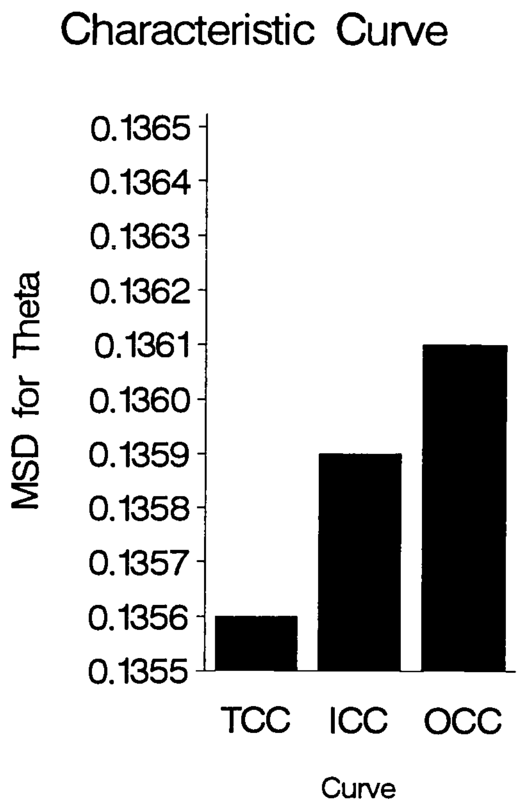


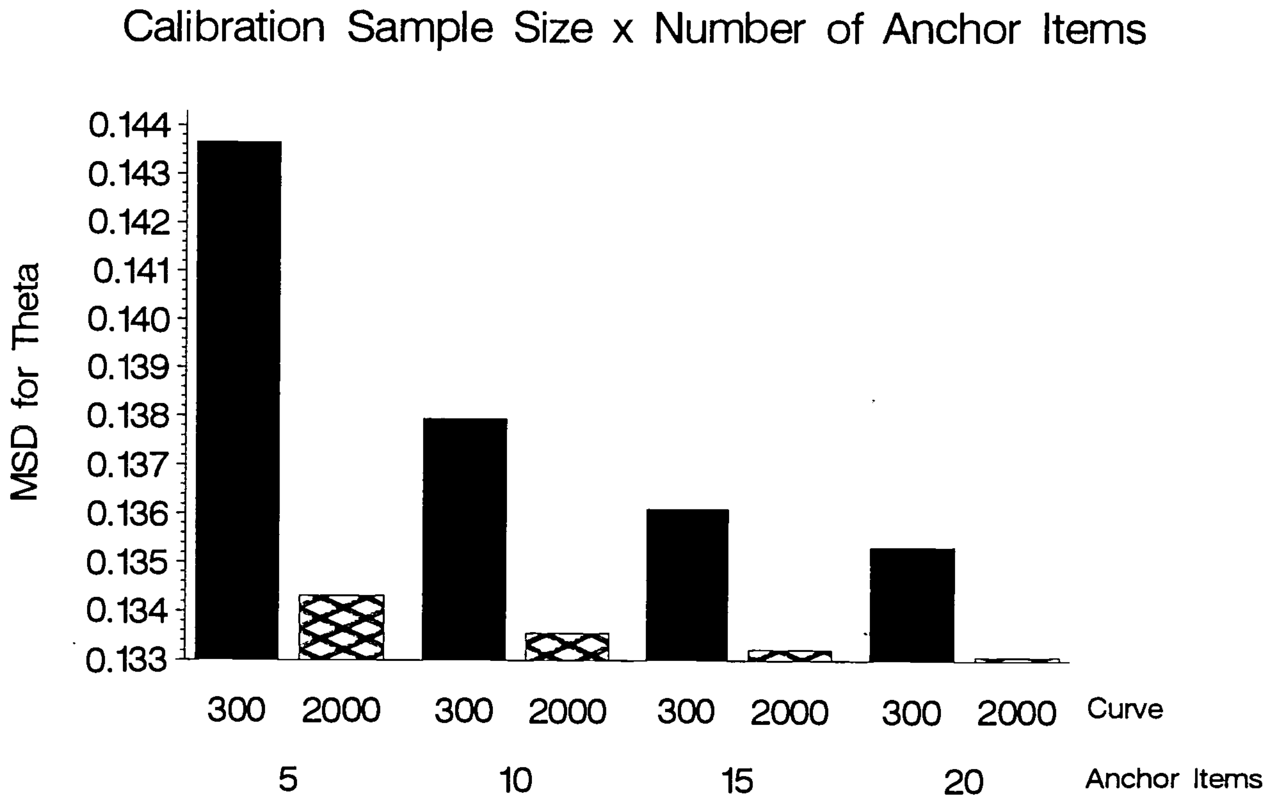
Figure 8



Calibration Sample Size

Number of Anchor Items


Characteristic Curve
MSD for Theta
0.1365
0.1364
0.1363
0.1362
0.1361
0.1360
0.1359
0.1358
0.1357
0.1356
0.1355
TCC
ICC
OCC
Curve


Calibration Sample Size x Number of Anchor Items
MSD for Theta
0.144
0.143
0.142
0.141
0.140
0.139
0.138
0.137
0.136
0.135
0.134
0.133
300
2000
300
2000
300
2000
300
2000
Curve
5
10
15
20
Anchor Items

U.S. Department of Education
Office of Educational Research and Improvement (OERI)
National Library of Education (NLE)
Educational Resources Information Center (ERIC)

REPRODUCTION RELEASE

(Specific Document)

TM034957

I. DOCUMENT IDENTIFICATION:

Title: Exploring Alternative Characteristic Curve Approaches to Linking Parameter Estimates from the Generalized Partial Credit Model.

Author(s): Roberts, J. S., Bao, H., Huang, C. & Gagne, P.

Corporate Source:	Publication Date:

II. REPRODUCTION RELEASE:

In order to disseminate as widely as possible timely and significant materials of interest to the educational community, documents announced in the monthly abstract journal of the ERIC system, *Resources in Education* (RIE), are usually made available to users in microfiche, reproduced paper copy, and electronic media, and sold through the ERIC Document Reproduction Service (EDRS). Credit is given to the source of each document, and, if reproduction release is granted, one of the following notices is affixed to the document.

If permission is granted to reproduce and disseminate the identified document, please CHECK ONE of the following three options and sign at the bottom of the page.

The sample sticker shown below will be affixed to all Level 1 documents

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY

Sample

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

1

Level 1

☑

Check here for Level 1 release, permitting reproduction and dissemination in microfiche or other ERIC archival media (e.g., electronic) *and* paper copy.

The sample sticker shown below will be affixed to all Level 2A documents

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE, AND IN ELECTRONIC MEDIA FOR ERIC COLLECTION SUBSCRIBERS ONLY, HAS BEEN GRANTED BY

Sample

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

2A

Level 2A

□

Check here for Level 2A release, permitting reproduction and dissemination in microfiche and in electronic media for ERIC archival collection subscribers only

The sample sticker shown below will be affixed to all Level 2B documents

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE ONLY HAS BEEN GRANTED BY

Sample

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

2B

Level 2B

□

Check here for Level 2B release, permitting reproduction and dissemination in microfiche only

Documents will be processed as indicated provided reproduction quality permits.
If permission to reproduce is granted, but no box is checked, documents will be processed at Level 1.

I hereby grant to the Educational Resources Information Center (ERIC) nonexclusive permission to reproduce and disseminate this document as indicated above. Reproduction from the ERIC microfiche or electronic media by persons other than ERIC employees and its system contractors requires permission from the copyright holder. Exception is made for non-profit reproduction by libraries and other service agencies to satisfy information needs of educators in response to discrete inquiries.

Sign here, → please

Signature: James S. Roberts	Printed Name/Position/Title: James S. Roberts, Assistant Professor	
Organization/Address: 1230 F Benjamin Bldg. University of Maryland College Park, MD 20742	Telephone: (301) 405-3630	FAX: (301) 314-9245
	E-Mail Address: JR245@umail.umd.edu	Date: 5/18/03

(Over)

III. DOCUMENT AVAILABILITY INFORMATION (FROM NON-ERIC SOURCE):

If permission to reproduce is not granted to ERIC, *or*, if you wish ERIC to cite the availability of the document from another source, please provide the following information regarding the availability of the document. (ERIC will not announce a document unless it is publicly available, and a dependable source can be specified. Contributors should also be aware that ERIC selection criteria are significantly more stringent for documents that cannot be made available through EDRS.)

Publisher/Distributor:
Address:
Price:

IV. REFERRAL OF ERIC TO COPYRIGHT/REPRODUCTION RIGHTS HOLDER:

If the right to grant this reproduction release is held by someone other than the addressee, please provide the appropriate name and address:

Name:
Address:

V. WHERE TO SEND THIS FORM:

Send this form to the following ERIC Clearinghouse: University of Maryland
ERIC Clearinghouse on Assessment and Evaluation
1129 Shriver Lab, Bldg 075
College Park, MD 20742
Attn: Acquisitions

However, if solicited by the ERIC Facility, or if making an unsolicited contribution to ERIC, return this form (and the document being contributed) to:

University of Maryland
ERIC Clearinghouse on Assessment and Evaluation
1129 Shriver Lab, Bldg 075
College Park, MD 20742
Attn: Acquisitions