DOCUMENT RESUME

ED 457 187 TM 033 283

AUTHOR	Roberts, James S.
TITLE	Equating Parameter Estimates from the Generalized Graded Unfolding Model.
PUB DATE	2001-04-00
NOTE	36p.; Paper presented at the Annual Meeting of the American Educational Research Association (Seattle, WA, April 10-14, 2001).
PUB TYPE	Reports - Research (143) -- Speeches/Meeting Papers (150)
EDRS PRICE	MF01/PC02 Plus Postage.
DESCRIPTORS	*Estimation (Mathematics); *Item Response Theory; Simulation
IDENTIFIERS	Parameter Identification

ABSTRACT

Three common methods for equating parameter estimates from binary item response theory models are extended to the generalized grading unfolding model (GGUM). The GGUM is an item response model in which single-peaked, nonmonotonic expected value functions are implemented for polytomous responses. GGUM parameter estimates are equated using extended versions of the mean-sigma, mean-mean, and item characteristic curve methods. The former two methods are implemented using two different strategies based on alternative parameterizations of the GGUM. All of these methods attempt to estimate a scale constant (A) and a location constant (B) that can equate the metric of item response model parameters derived from separate calibrations. A small simulation is performed to provide preliminary information about the characteristics of the alternative equating methods studied. The item characteristic curve method performed best with regard to the mean squared error, bias, and standard error of equating constant estimates as well as the absence of extremely deviant estimates. It was noted that, although the average superiority of estimates produced by the item characteristic curve method was quite small, substantial outliers sometimes emerged when estimating equating constants with other methods. Consequently, the item characteristic curve method is recommended as a means to develop estimates of equating constants in the GGUM. An appendix discusses equating parameter estimates from the GGUM. (Contains 3 figures, 1 table, and 17 references.) (Author/SLD)

Equating Parameter Estimates from the Generalized Graded Unfolding Model

James S. Roberts

University of Maryland

Paper presented at the 2001 American Educational Research Association annual meeting, Seattle, Washington

Abstract

Three common methods for equating parameter estimates from binary item response theory models are extended to the generalized graded unfolding model (GGUM). The GGUM is an item response model in which single-peaked, nonmonotonic expected value functions are implemented for polytomous responses. GGUM parameter estimates are equated using extended versions of the mean-sigma, mean-mean, and item characteristic curve methods. The former two methods are implemented using two different strategies based on alternative parameterizations of the GGUM. All of these methods attempt to estimate a scale constant (A) and a location constant (B) that can equate the metric of item response model parameters derived from separate calibrations. A small simulation is performed to provide preliminary information about the characteristics of the alternative equating methods studied. The item characteristic curve method performed best with regard to the mean squared error, bias, and standard error of equating constant estimates as well as the absence of extremely deviant estimates. It was noted that although the average superiority of estimates produced by the item characteristic curve method was quite small, substantial outliers sometimes emerged when estimating equating constants with other methods. Consequently, the item characteristic curve method is recommended as a means to develop estimates of equating constants in the GGUM. *Key Words: GGUM, generalized graded unfolding model, equating, linking, item response theory, unfolding, characteristic curve, mean-sigma, mean-mean*

The generalized graded unfolding model (GGUM; Roberts, Donoghue & Laughlin, 2000) is a unidimensional, parametric, item response theory model that is applicable to either binary or polytomous responses that follow from a proximity relation (Coombs, 1964). A proximity-based response process is one in which an individual is expected to obtain higher item scores to the extent that the individual is located close to a given item on an underlying latent continuum. This notion is consistent with traditional attitude measurement applications where respondents are asked to indicate their level of disagreement or agreement with each statement on an attitude questionnaire (Roberts, Laughlin, & Wedell, 1999). It is also generally implied when measuring preferences (DeSarbo & Hoffman, 1987) and certain developmental processes in which particular cognitions or behaviors occur in distinct stages (Nöel, 1999). The remainder of this paper will presume an attitude measurement context.

The GGUM is an item response theory (IRT) model, and as such, it provides a means to develop large item banks from multiple attitude questionnaires which might subsequently be used as the foundation for computerized adaptive attitude assessments (Roberts, Lin & Laughlin, 2001). The IRT framework also enables one to examine if and how an attitude item functions differently in alternative subpopulations. These applications presuppose that characteristics of attitude items derived from separate calibrations can be expressed on a common metric. Therefore, a reliable means to equate GGUM parameter estimates across multiple calibrations is required before such applications are possible.

The GGUM is consistent with a proximity-based response process, and thus, it yields item characteristic curves that are single-peaked and nonmonotonic. These nonmonotonic item characteristics lead to a test characteristic function which is not a one-to-one function of the latent trait. Consequently, equating strategies based on raw test scores (e.g., equipercentile equating or linear equating based) are not appropriate in this case because most raw test scores are associated with at least two or more points on the latent continuum. Fortunately, the GGUM provides a means to equate tests from an IRT perspective which logically incorporates this nonmonotonic relationship between item responses and the latent trait.

The GGUM is defined by its category probability function which is equal to:

$$P[Z_i = z \mid \theta_j, \alpha_i, \delta_i, \tau_{ik}] = \frac{\exp\left(\alpha_i\left[z(\theta_j-\delta_i) - \sum_{k=0}^{z}\tau_{ik}\right]\right) + \exp\left(\alpha_i\left[(M-z)(\theta_j-\delta_i) - \sum_{k=0}^{z}\tau_{ik}\right]\right)}{\sum_{w=0}^{C}\left[\exp\left(\alpha_i\left[w(\theta_j-\delta_i) - \sum_{k=0}^{w}\tau_{ik}\right]\right) + \exp\left(\alpha_i\left[(M-w)(\theta_j-\delta_i) - \sum_{k=0}^{w}\tau_{ik}\right]\right)\right]}, \quad (1)$$

where: Z_i = an observable response to statement (item) i,

z = 0, 1, 2, ..., C ; $z = 0$ corresponds to the strongest level of disagreement and $z = C$ refers to the strongest level of agreement,

θ_j = the location of the *jth* individual on the latent continuum,

δ_i = the location of the *ith* item on the latent continuum,

α_i = the discrimination of the *ith* item,

τ_{ik} = the *kth* subjective category threshold parameter associated with the *ith* item,

C = the number of observable response categories minus 1, and

$M = 2*C + 1$.

From an IRT perspective, if responses to two sets of test items are analyzed separately using two GGUMs, then test equating is simply a matter of placing the parameter estimates from the two calibrations onto a common metric. Several methods have been suggested to establish a common metric for item response models with monotonic item characteristics. This paper will extend some of the more popular equating strategies for monotonic models to the GGUM and provide preliminary information about the adequacy of each strategy.

Linear Indeterminacy of GGUM Parameter Estimates

The GGUM yields response probabilities that are invariant with respect to the unit and origin of the latent continuum. Consequently:

$$P[Z_i = z \mid \theta_j, \alpha_i, \delta_i, \tau_{ik}] = P[Z_i = z \mid \theta_j^*, \alpha_i^*, \delta_i^*, \tau_{ik}^*] , \tag{2}$$

where:

$$\theta_j^* = A\,\theta_j + B , \tag{3}$$

$$\alpha_i^* = \frac{\alpha_i}{A} , \tag{4}$$

$$\delta_i^* = A\,\delta_i + B , \tag{5}$$

and

$$\tau_{ik}^* = A\,\tau_{ik} . \tag{6}$$

In order to remove this linear indeterminacy during parameter estimation, the A and B constants are fixed to some arbitrary values. For example, the GGUM2000 estimation software uses an N(0,1) prior distribution for θ which consequently constrains the unit and origin of the latent continuum.

Equating GGUM parameter estimates is a matter of choosing equating constants, A and B, that transform the metric associated with one set of parameter estimates to the target metric of another set of estimates. The methods of estimating equating constants presented in this paper presume that there are common test items among the multiple forms to be equated. Consequently, the methods are appropriate for two basic types of designs. In a multiple-group, common-form design, two or more groups of respondents receive the same test form. However, the researcher, by choice or circumstance, calibrates the responses from each group separately. In this situation, the metric of the GGUM parameter estimates will be different to the extent that the multiple groups of respondents have different θ distributions. In the multiple-group, multiple-form, anchor-item design, two or more groups of respondents receive alternative forms of a test, but pairs of forms are related to each other through a set of common items (i.e., anchor items). The common feature of both equating designs is that either the entire test form or some subset of test items is identical for two or more respondent groups in a given application.

Extending Some Common IRT Equating Methods to the GGUM

The Mean-Sigma Method

Marco (1977) introduced the mean-sigma method of equating IRT parameter estimates in the 3-parameter logistic model. This method can be extended to the GGUM as follows. Let $\hat{\delta}_{i,1}$ be the item location estimate for the *ith* common item from the first calibration. Similarly, let $\hat{\delta}_{i,2}$ be the item location estimate for the *ith* common item from the second calibration. Suppose the goal of the equating procedure is to transform the metric of the GGUM parameter estimates from the second calibration to that for the first calibration. The equating constants can be estimated from the means and standard deviations of the item location estimates as follows:

$$\hat{B} = \bar{\hat{\delta}}_{i,1} - \frac{s_{\hat{\delta}_{i,1}}}{s_{\hat{\delta}_{i,2}}} \bar{\hat{\delta}}_{i,2} , \tag{7}$$

and

$$\hat{A} = \frac{s_{\hat{\delta}_{i,1}}}{s_{\hat{\delta}_{i,2}}} , \tag{8}$$

where $\bar{\hat{\delta}}_{i,1}$ and $\bar{\hat{\delta}}_{i,2}$ are the means of the item location estimates for the first and second calibrations, respectively; and $s_{\hat{\delta}_{i,1}}$ and $s_{\hat{\delta}_{i,2}}$ are the corresponding standard deviations. With these equating constants estimated, the transformed parameters can be obtained using the following formulae:

$$\hat{\alpha}_{i,21} = \frac{\hat{\alpha}_{i,2}}{\hat{A}} , \tag{9}$$

$$\hat{\delta}_{i,21} = \hat{A}\,\hat{\delta}_{i,2} + \hat{B} , \tag{10}$$

$$\hat{\tau}_{ik,21} = \hat{A}\,\hat{\tau}_{ik,2} , \tag{11}$$

and

$$\hat{\theta}_{j_2,21} = \hat{A}\,\hat{\theta}_{j_2,2} + \hat{B} , \tag{12}$$

where $\hat{\alpha}_{i,21}$, $\hat{\delta}_{i,21}$, $\tau_{ik,21}$ and $\hat{\theta}_{j_2,21}$ denote the parameter estimates from the second calibration after they have been transformed to the metric of those from the first calibration.

One characteristic of the equating constants estimated with Equations 7 and 8 is that they ignore information contained in all parameter estimates other than $\hat{\delta}_i$. An alternative extension of the method can be constructed that takes information about both $\hat{\delta}_i$ and $\hat{\tau}_{ik}$ into account.

Specifically, the GGUM can be re-expressed as:

$$P[Z_i = z \mid \theta_j, \alpha_i, \xi_{is}] = \frac{\exp\left(\alpha_i \sum_{s=0}^{z} (\theta_j - \xi_{is})\right) + \exp\left(\alpha_i \sum_{s=0}^{(M-z)} (\theta_j - \xi_{is})\right)}{\sum_{w=0}^{C}\left[\exp\left(\alpha_i \sum_{s=0}^{w} (\theta_j - \xi_{is})\right) + \exp\left(\alpha_i \sum_{s=0}^{(M-w)} (\theta_j - \xi_{is})\right)\right]}, \tag{13}$$

where

$$\xi_{is} = \delta_i + \tau_{is} , \tag{14}$$

and

$$\tau_{is} = \tau_{ik} , \; for \;\; s = k \leq C , \tag{15}$$

$$\tau_{is} = 0 , \; for \;\; s = C+1 , \tag{16}$$

and

$$\tau_{is} = -\tau_{i(M-s+1)} , \; for \;\; s > C+1 , \tag{17}$$

due to the assumption of symmetric subjective response category thresholds (Roberts, Donoghue & Laughlin, 2000). Note that in Equation 13, $\xi_{i0}=0$ and $\sum_{k=0}^{0} (\theta_j - \xi_{ik}) = 0$ by definition. These constraints imply that there are C+1 estimable ξ_{is} parameters; namely $\xi_{i1}, \ldots, \xi_{i(C+1)}$. The estimable ξ_{is} parameters can be used to develop equating constants with the mean-sigma procedure:

$$\hat{B} = \bar{\xi}_{is,1} - \frac{s_{\xi_{is,1}}}{s_{\xi_{is,2}}} \bar{\xi}_{is,2} , \tag{18}$$

and

$$\hat{A} = \frac{s_{\hat{\xi}_{is,1}}}{s_{\hat{\xi}_{is,2}}} , \tag{19}$$

where the means and standard deviations are taken across all estimable ξ_{is} in a given calibration. These constants can then be used to rescale the original parameters from the second calibration as indicated in Equations 9 through 12. Alternatively, Equations 9 and 12 can be used in conjunction with the following formula to rescale parameters using the ξ_{is} parameterization:

$$\hat{\xi}_{is,21} = \hat{A}\,\hat{\xi}_{is,2} + \hat{B} \ . \tag{20}$$

Although this formulation of the mean-sigma method includes information about both $\hat{\delta}_i$ and $\hat{\tau}_{ik}$, the former estimates are generally estimated more accurately than the latter (Roberts, Donoghue, & Laughlin, 2001). Therefore, it remains to be seen which version of the mean-sigma technique yields more accurate or more stable estimates of equating constants.

The Mean-Mean Method

Loyd and Hoover (1980) introduced the mean-mean method of estimating equating constants in the context of the 1-parameter logistic model. The mean-mean method can be extended to the GGUM as follows:

$$\hat{B} = \overline{\hat{\delta}}_{i,1} - \left(\frac{\overline{\hat{\alpha}}_{i,2}}{\overline{\hat{\alpha}}_{i,1}} \right) \overline{\hat{\delta}}_{i,2} , \tag{21}$$

and

$$\hat{A} = \frac{\overline{\hat{\alpha}}_{i,2}}{\overline{\hat{\alpha}}_{i,1}} . \tag{22}$$

This formulation uses information about both $\hat{\delta}_i$ and $\hat{\alpha}_i$, and one would generally expect that this added source of information would help better represent the metric difference between the two calibrations. Another potential advantage of this method is that, unlike the mean-sigma approach, the mean-mean technique involves only means of parameter estimates, and means of parameter estimates may be more robust to outliers than standard deviations (Baker & Al-Karni, 1991). However, Roberts et al. (2001) have shown that α_i are more difficult to estimate than δ_i, so the advantage of including information about mean $\hat{\alpha}_i$ could potentially be outweighed by estimation error.

The mean-mean method formulation above can be easily adapted to incorporate information about all item parameter estimates. If the GGUM is parameterized using Equation 13, then an estimate of B can be obtained from:

$$\hat{B} = \bar{\hat{\xi}}_{is,1} - \left(\frac{\bar{\hat{\alpha}}_{i,2}}{\bar{\hat{\alpha}}_{i,1}} \right) \bar{\hat{\xi}}_{is,2} \,. \tag{23}$$

Note that the estimate of A is still derived using Equation 22. This formulation includes information about all item parameters, but it is not clear that inclusion of these parameters will increase the accuracy of the equating constant estimates given that both τ_{ik} and α_i are generally more difficult to estimate than δ_i (Roberts, et al, 2001).

Characteristic Curve Methods

A variety of characteristic curve methods have been proposed for monotonic models in the IRT literature. These methods attempt to find the equating constants that minimize discrepancies between characteristic curves developed from items that are common across two calibrations.

This presumably leads to more accurate estimates because deviant estimates have relatively less impact on the results as compared to procedures that incorporate only summary measures of item parameter distributions corresponding to the common items (Baker & Al-Karni, 1991; Stocking & Lord, 1983). Characteristic curve methods differ in the specific type of curves that are contrasted. Differences between test (Stocking & Lord, 1983; Baker, 1992), item (Haebara, 1980) and category (Baker, 1993) characteristic curves have been proposed for alternative monotonic IRT models. Haebara's (1980) item characteristic curve method is especially attractive because it produces symmetric results when the target metric is that for the second calibration rather than the first. It is also unique in that it explicitly incorporates information about the distributions of $\hat{\theta}_{j_1, 1}$ and $\hat{\theta}_{j_2, 2}$ when evaluating differences between characteristic curves.

Haebara originally proposed the item characteristic curve method for the 3-parameter logistic model. It can be extended to the GGUM as follows. Let $E_{i, 1}(\hat{\theta}_{j_1, 1})$ be the expected value of the *ith* common item from the first calibration given the j_1*th* individual's latent trait estimate derived in that calibration:

$$E_{i, 1}(\hat{\theta}_{j_1, 1}) = \sum_{z=0}^{C} z \, P[Z_i = z \mid \hat{\theta}_{j_1, 1}, \hat{\alpha}_{i, 1}, \hat{\delta}_{i, 1}, \hat{\tau}_{ik, 1}] \; . \tag{24}$$

Similarly, let $E_{i, 2}(\hat{\theta}_{j_2, 2})$ be the expected value of the *ith* common item from the second calibration given the j_2*th* individual's latent trait estimate derived in that calibration. Let $E_{i, 1}(\hat{\theta}_{j_2, 21})$ denote the expected value of the *ith* common item from the first calibration given the j_2*th* individual's transformed latent trait estimate. The transformation takes the individual's trait estimate from the second calibration and rescales it to the metric of the first:

$$\hat{\theta}_{j_2,21} = A\,\hat{\theta}_{j_2,2} + B \ . \tag{25}$$

In an analogous fashion, let $E_{i,2}(\hat{\theta}_{j_1,12})$ denote the expected value of the *ith* common item from the second calibration given the j_1th individual's transformed latent trait estimate. This transformation is given by:

$$\hat{\theta}_{j_1,12} = \frac{\hat{\theta}_{j_1,1} - B}{A} \ . \tag{26}$$

Note that if there were 1) no sampling error, 2) no differential item functioning, and 3) perfect fit of the GGUM to the data in each calibration, then values of A and B could be found to make the following identity hold for all i, j_1, and j_2:

$$E_{i,2}(\hat{\theta}_{j_2,2}) - E_{i,1}(\hat{\theta}_{j_2,21}) = E_{i,1}(\hat{\theta}_{j_1,1}) - E_{i,2}(\hat{\theta}_{j_1,12}) = 0 \ . \tag{27}$$

However, under less idealistic circumstances, this identity will not hold. Instead, estimates of A and B are developed to minimize the following loss function:

$$Q = \sum_{i=1}^{I} \left(\sum_{j_2=1}^{J_2} \left[E_{i,2}(\hat{\theta}_{j_2,2}) - E_{i,1}(\hat{\theta}_{j_2,21}) \right]^2 + \sum_{j_1=1}^{J_1} \left[E_{i,1}(\hat{\theta}_{j_1,1}) - E_{i,2}(\hat{\theta}_{j_1,12}) \right]^2 \right) . \tag{28}$$

The values of $\hat{A}$ and $\hat{B}$ are calculated by solving the following system of equations:

$$\frac{\partial Q}{\partial B} = 0 \ , \tag{29}$$

and

$$\frac{\partial Q}{\partial A} = 0 \ . \tag{30}$$

The solution can be found using the Newton-Raphson procedure in which the estimates are updated iteratively. On iteration $t + 1$, the estimates are calculated using the following formula:

$$[B \quad A]_{t+1} = [B \quad A]_t - \begin{bmatrix} \frac{\partial Q}{\partial B} & \frac{\partial Q}{\partial A} \end{bmatrix}_t \begin{bmatrix} \frac{\partial^2 Q}{\partial B^2} & \frac{\partial^2 Q}{\partial B \partial A} \\ \frac{\partial^2 Q}{\partial A \partial B} & \frac{\partial^2 Q}{\partial A^2} \end{bmatrix}_t^{-1} . \tag{31}$$

The derivatives required to solve Equation 31 are given in the appendix.

Simulation Study

The forgoing methods of estimating equating constants have not yet been investigated in the context of the GGUM. Therefore, a small simulation study was conducted to provide preliminary information about the relative performance of each method. The goal of this study was to determine if any of these methods produced substantially different estimates than their counterparts under a limited number of simulated conditions.

Method

Design

The simulation was based on a 2 (equating condition) x 2 (sample size) factorial design. In each cell of the design, responses to 20 polytomous (6-category) items were generated for two groups of subjects based on Equation 1. Parameter estimates were calibrated separately for each group using GGUM2000 software (Roberts, 2001). Solutions were derived using an N(0,1) prior distribution for θ, 30 quadrature points, and a convergence criterion of .001. The resulting parameter estimates for the second group were equated to those for the first group using the mean-sigma method based on $\hat{\delta}_i$ (MS1), the mean-sigma method based on $\hat{\xi}_{is}$(MS2), the mean-mean method based on $\hat{\delta}_i$ (MM1), the mean-mean method based on $\hat{\xi}_{is}$(MM2), and the item characteristic curve method (ICC). This process of generating responses, estimating GGUM parameters and estimating equating constants was replicated 30 times in each cell of the design.

Item Characteristics

True item parameter values used in this simulation were similar to those found in real data. The true δ_i for the 20 items were randomly sampled from a uniform distribution ranging from (-2, +2). True α_i were randomly sampled from a uniform distribution on the interval of (.5, 2). Threshold parameters (τ_{ik}) corresponding to a 6-category response were generated independently for each item. For a given item, the true τ_{iC} parameter was generated from a uniform (-1.4, -.4) distribution. Successive true τ_{ik} parameters were then generated with the following recursive equation:

$$\tau_{i\,k-1} = \tau_{ik} - .25 + e_{i\,k-1}\,, \quad for\ k = 2,3,\ldots,C \tag{32}$$

where $e_{i\,k-1}$ denotes a random error term generated from a N(0, .04) distribution. The τ_{ik} parameters derived with this formula were not necessarily ordered across the continuum within an item.

The true item parameters were independently sampled on every replication in each condition. However, these parameters were held constant for the two groups of responses simulated on each replication. Therefore, the simulation was consistent with a situation in which parameter estimates from a common form were equated across two calibrations. It was also similar to a 2-group, 2-form, anchor item equating situation in which 20 anchor items were used. (However, one would generally expect more precise estimates of θ_j in the latter situation due to the larger number of total test items.)

Equating Conditions

The simulated equating condition was either a horizontal or vertical equating scenario. In the horizontal condition, true θ values were normally distributed with $\bar{X} = 0$ and $s = 1$ in both respondent groups. Consequently, the true values for A and B were 1 and 0, respectively, in the horizontal equating condition. In the vertical equating condition, true θ values were normally distributed with $\bar{X} = 0$ and $s = 1$ in the first respondent group, and they were normally distributed with $\bar{X} = .5$ and $s = 1.25$ in the second respondent group. Given that an N(0,1) prior distribution for θ was used to estimate parameters in both groups, the origin and scale of parameter estimates from the second respondent group were translated to those of the prior distribution, whereas the origin and scale of parameter estimates from the first respondent group remained unchanged. Consequently, the true values of A and B in the vertical equating condition equaled 1.25 and .5, respectively, and these values served to reestablish the original metric of parameters in the second respondent group.

Sample Size

Responses from either 300 or 1000 simulees were used in each calibration. Recovery simulations by Roberts et al. (2001) have suggested that approximately 750 respondents may be required to produce very accurate GGUM item parameter estimates when 15 to 20 uniformly-spaced items with six response categories per item are used. Therefore, the 1000 simulee condition was thought to represent a satisfactory sample size. Roberts et al. have also shown that 300 respondents can lead to GGUM estimates with noticeably lower levels of accuracy. Consequently, the 300 simulee condition was expected to produce estimates with substantially larger amounts of error with a higher potential for outliers.

Analysis of Simulation Results

The adequacy of each equating constant estimation method was assessed by studying the mean squared error associated with the corresponding estimates. Specifically, the squared difference (i.e., squared error) between an estimated constant and its true value was calculated for each $\hat{A}$ and $\hat{B}$ produced with a particular estimation method. There were 30 squared error scores associated with a given type of estimate in each cell of the simulation design. Differences in mean squared error scores for $\hat{B}$ were analyzed using a 2 (equating condition) x 2 (sample size) x 5 (estimation method) split-plot analysis of variance (ANOVA). The first two factors in this analysis were between-replication factors whereas the third was a within-replication factor. A similar analysis was run for $\hat{A}$ although the MM2 estimates were not used because they were mathematically identical to the MM1 estimates. Therefore, only 4 levels of the within-replication factor were present in the analysis of $\hat{A}$. The Type I error rate used in each ANOVA was set at .025 to control for the fact that two dependent measures were studied using the same analytical design. Probabilities associated with tests of within-replication effects were corrected with the Huynh-Felt procedure. The proportion of total between-replication variance attributable to each between-replication effect in a given ANOVA (i.e., η^2) was calculated. A similar quantity was also calculated for all within-replication effects based only on the total within-replication variance.

Descriptive analyses were performed to supplement the primary analysis of squared error. Specifically, the empirical bias and the standard error inherent in each type of estimate was examined. A graphical analysis of the variability in the estimates under each simulation condition was also conducted.

Results

Mean Squared Error

The ANOVA on the squared error associated with $\hat{A}$ revealed statistically significant main effects of sample size (F (1, 116)=21.22, MSe=.0001, p=.0001; η^2=.147) and estimation method (F (3, 348)=13.97, MSe=.0001, $p_{adjusted}$ = .0001; η^2=.096), and an interaction between sample size and estimation method (F (3, 348)=10.66, MSe=.0001, $p_{adjusted}$ = .0001; η^2=.073). The main effect of sample size was in the expected direction with slightly larger mean squared error occurring with the smaller sample size. These mean differences were confined to the third decimal place (.006 versus .001). The main effect of estimation method was due to the fact that the ICC method produced the smallest mean squared error (.0003), followed by the MM1 method (.001), and the MS1 and MS2 methods (.004 and .005). These mean differences were, again, confined to the third decimal place and were, thus, quite small. The top panel of Figure 1 illustrates the interaction between sample size and estimation method for $\hat{A}$. The mean squared error was generally small and similar for all estimation methods when the sample size was equal to 1000. However, when the sample size was equal to 300, the mean squared error in $\hat{A}$ increased for all methods. Post-hoc comparisons showed that these differences were statistically significant for all estimation methods, although they were smallest for the ICC and most noticeable for MS2 and MS1.

Insert Figure 1 About Here

The ANOVA on squared error for $\hat{B}$ revealed statistically significant effects for all between-replication factors. The main effect of equating condition (F (1, 116)=14.39, MSe=.0002, p=.0002; η^2=.093) was due to a slightly higher mean squared error in estimates derived in the vertical equating scenario (.006 versus .002). The main effect of sample size (F (1, 116)=17.45, MSe=.0002, p=.0001; η^2=.113) was such that slightly higher mean squared errors were found for the smaller sample (.006 versus .001). The interaction of these factors (F (1,116)=6.69, MSe=.0002, p=.0109; η^2=.043) suggested that the mean squared error was relatively high (.010) when vertical equating was performed and the sample size was small. The mean squared error obtained in the remaining between-replication conditions was consistently smaller (i.e., between .001 and .003).

There were also reliable within-replication effects for squared errors of $\hat{B}$. In particular, there was a statistically significant main effect of estimation method (F (4, 464)=8.02, MSe=.0001, $p_{adjusted}$=.0003; η^2=.061). The mean squared error values corresponding to this main effect were equal to .001 for ICC, .003 for MS1, .003 for MM1, .004 for MS2, and .007 for MM2. A post-hoc analysis showed that the ICC method produced a statistically smaller mean squared error than did the MS2 and MM2 methods. Additionally, the MM2 method produced a slightly larger degree of error than did the MM1 procedure. The interaction of sample size and estimation method was also statistically significant (F (4, 464)=3.77, MSe=.0001, $p_{adjusted}$=.0207; η^2=.029).

The lower panel of Figure 1 shows the cell means corresponding to this interaction. The mean square error was consistently larger when the sample size was 300. However, this was especially true for the MM2 and MS2 methods.

Insert Table 1 About Here

Empirical Bias and Standard Deviation of Estimates

The means and standard deviations of $\hat{A}$ and $\hat{B}$ across the 30 replications are given in Table 1 for each simulated condition. The empirical bias in $\hat{A}$ was typically negligible and unsystematic with the largest discrepancy of -.036 occurring for the MS2 method in the vertical equating condition with 300 simulees. The standard deviation of $\hat{A}$ was smallest for the ICC method across all conditions, and it was generally largest for the MS1 and MS2 methods. These differences in standard errors were more apparent in the small sample size conditions.

The empirical bias in $\hat{B}$ was generally negligible in the horizontal equating conditions. The largest degree of bias in these conditions occurred for the MS2 method when the sample size was small, in which case, the bias was only .014. Noticeably more bias in $\hat{B}$ occurred in the vertical equating conditions; especially when the sample size was small. Across these conditions, the degree of bias observed for the ICC and MS2 methods was slightly less than that seen with the other methods. The standard error in $\hat{B}$ estimates was consistently smallest for the ICC method and largest for the MM2 procedure. Again, these differences were more apparent in the small sample size condition.

Graphical Analysis of Estimates

Figures 2 and 3 illustrate the estimates of A and B obtained with each method on every replication of the simulation. Figure 2 gives a scatterplot of the $\hat{A}$ separately for each simulation condition. The horizontal line drawn on each scatterplot represents the true value of $\hat{A}$ in the given condition. As shown in Figure 2, the variability in the $\hat{A}$ values was relatively greater in the small sample conditions. The increased variability induced by small calibration samples was particularly evident in the vertical equating condition. More importantly , the scatterplot emphasizes that although average measures of estimation accuracy reported in previous sections generally suggested only minor differences between estimates produced by alternative methods, substantial differences among estimates emerged on several replications. When such differences occurred, it was generally the case that the MS1 and MS2 estimates of A were the most disparate.

Insert Figures 2 and 3 Here

Figure 3 provides a scatterplot of $\hat{B}$ values by replication for each simulated condition. The figure illustrates the increased variability in $\hat{B}$ values reported previously for vertical equating. It also depicts the increased variability associated with smaller calibration samples as well as the multiplicative effect of small samples in a vertical equating scenario. Again, the scatterplot shows that substantial differences did emerge occasionally among the $\hat{B}$ values produced by alternative methods. Such differences occurred even in the large sample conditions, albeit much less frequently. Furthermore, when large discrepancies occurred among the alternative estimates, it was generally the case that the MM2 method produced the most aberrant $\hat{B}$.

Discussion

The foregoing results suggest that with large calibration samples, the alternative estimates of equating constants developed for the GGUM may be reasonably similar. Nonetheless, even in the large sample conditions, the estimates produced by the ICC generally showed a slight advantage relative to those from the other methods. This advantage was manifested primarily by a slightly smaller standard error. In the small sample size conditions, the relatively greater accuracy of estimates produced by the ICC method became more apparent. Estimates produced by the ICC method were generally more efficient, and in the vertical equating condition, they showed less bias than several rival estimates.

These findings are consistent with those from past studies of IRT equating methods for cumulative models. Such studies indicate that mean-sigma and mean-mean estimates of equating constants are generally similar to those produced by characteristic curve methods when the IRT parameters are estimated well (Baker & Al-Karni, 1991; Cohen & Kim, 1998). This was the case in the large sample conditions in the current simulation. The psychometric literature also suggests that characteristic curve methods will be more robust than mean-sigma and mean-mean methods when some item parameter estimates are deviant (i.e., when outliers are present; Baker & Al-Karni, 1991; Stocking & Lord, 1983). The current simulation used only 300 simulees in the small sample conditions, and Roberts et al. (2001) have previously shown that samples of this size lead to relatively inaccurate estimates of GGUM item parameters. Thus, the increased accuracy noted with the ICC estimates in the small sample condition is likely due to the robustness of the ICC method in the midst of degraded GGUM item parameter estimates.

Meaningful differences in estimates of equating constants can occasionally occur even when summary measures like the mean squared error or the standard error of the estimate differ only slightly. Therefore, it is important to understand which estimation methods produce the highest frequency of outliers. The MS1 and MS2 methods exhibited the strongest tendency to produce extreme estimates of A, whereas the MM2 exhibited the strongest tendency to produce extreme estimates of B. To the extent that the results of this preliminary simulation are generalizable, then one should avoid using these methods to equate parameters of the GGUM. In contrast, the ICC and MM1 methods produced a smaller number of outliers and can be recommended on those grounds. The ICC method appeared to yield slightly more accurate estimates than those from the MM1 method, and thus, it should be the method of choice. However, in situations where θ estimates are not readily available, the MM1 could still be used to equate item parameter estimates.

The simulation reported in this paper was preliminary in nature. A number of interesting variables were not explored in the simulation including the roles of test length and the proportion of anchor items on a given test. The present work was also limited to 6-category responses, and it did not vary the degree of difference between θ distributions in the vertical equating scenarios. The ICC method evaluates differences in item characteristic curves, but differences could also be determined at the test or category levels. Furthermore, this implementation of the ICC evaluated characteristic curve differences at every θ_j observed in either respondent group. Other evaluation strategies could certainly be used (e.g., a fixed number of equally spaced θ_j points). The impact of these variables are left for future exploration.

Conclusions

This study suggests that the item characteristic curve method provides a robust means to estimate equating constants in the GGUM. Moreover, the preliminary simulation suggests that the item characteristic method can provide relatively more accurate estimates than the mean-sigma or mean-mean methods. The ability to accurately equate GGUM parameter estimates should facilitate the development of alternative test forms and item banks in situations where responses to questionnaire items unfold. This, in turn, will make other applications such as computerized adaptive attitude testing more practical; especially if item banks are shared among social science researchers. The development of sound equating methods for the GGUM is a fundamental step in the pursuit of these benefits.

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

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

Coombs, C. (1964). *A theory of data.* New York: Wiley.

DeSarbo, W. S., & Hoffman, D. L. (1987). Constructing MDS joint spaces from binary choice data: A multidimensional unfolding threshold model for marketing research. *Journal of Marketing Research, 24,* 40-54.

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

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.

Nöel, Y. (1999). Recovering unimodal latent patterns of change by unfolding analysis: Application to smoking cessation. *Psychological Methods, 4,* 173-191.

Press, W. H., Teukolsky, S. A., Vetterling, W. T., Flannery, B. P. (1992). *Numerical recipes in FORTRAN: The art of scientific computing* (2nd ed.). Cambridge: Cambridge University Press.

Roberts, J. S. (2001). GGUM2000: Software to Estimate Parameters in the Generalized Graded Unfolding Model. *Applied Psychological Measurement*, in press.

Roberts, J. S., Donoghue, J. R., & Laughlin, J. E. (2001). Characteristics of MML/EAP parameter estimates in the generalized graded unfolding model. *Applied Psychological Measurement,* conditionally accepted.

Roberts, J. S., Donoghue, J. R., & Laughlin, J. E. (2000). A general item response theory model for unfolding unidimensional polytomous responses. *Applied Psychological Measurement, 24*, 3-32.

Roberts, J. S., Laughlin, J. E., & Wedell, D. H. (1999). Validity issues in the Likert and Thurstone approaches to attitude measurement. *Educational and Psychological Measurement, 59,* 211-233.

Roberts, J. S., Lin, Y., & Laughlin, J. E. (2001). Computerized adaptive testing with the generalized graded unfolding model. *Applied Psychological Measurement*, in press.

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

Appendix

Let the loss function be defined as:

$$
\begin{aligned}
Q &= \sum_{i=1}^{I} Q_{1i} + Q_{2i} \\
&= \sum_{i=1}^{I} \left(\sum_{j_2=1}^{J_2} \left[E_{i,2}(\hat{\theta}_{j_2,2}) - E_{i,1}(\hat{\theta}_{j_2,21}) \right]^2 + \sum_{j_1=1}^{J_1} \left[E_{i,1}(\hat{\theta}_{j_1,1}) - E_{i,2}(\hat{\theta}_{j_1,12}) \right]^2 \right) ,
\end{aligned}
\tag{33}
$$

where I is the number of common items across the two calibrations, J_1 is the number of respondents in the first calibration sample and J_2 is the number of respondents in the second calibration sample. The partial derivative of the loss function with respect to A is equal to:

$$
\frac{\partial Q}{\partial A} = \sum_{i=1}^{I} \left(\frac{\partial Q_{1i}}{\partial A} + \frac{\partial Q_{2i}}{\partial A} \right) ,
\tag{34}
$$

where

$$
\frac{\partial Q_{1i}}{\partial A} = \sum_{j_2=1}^{J_2} -2 \left[E_{i,2}(\hat{\theta}_{j_2,2}) - E_{i,1}(\hat{\theta}_{j_2,21}) \right] \left[\sum_{z=0}^{C} z \, \frac{\partial P_{iz,1}(\hat{\theta}_{j_2,21})}{\partial A} \right] ,
\tag{35}
$$

and

$$
\frac{\partial Q_{2i}}{\partial A} = \sum_{j_1=1}^{J_1} -2 \left[E_{i,1}(\hat{\theta}_{j_1,1}) - E_{i,2}(\hat{\theta}_{j_1,12}) \right] \left[\sum_{z=0}^{C} z \, \frac{\partial P_{iz,2}(\hat{\theta}_{j_1,12})}{\partial A} \right] .
\tag{36}
$$

The partial derivative with respect to B is equal to:

$$
\frac{\partial Q}{\partial B} = \sum_{i=1}^{I} \left(\frac{\partial Q_{1i}}{\partial B} + \frac{\partial Q_{2i}}{\partial B} \right) ,
\tag{37}
$$

where

$$\frac{\partial Q_{1i}}{\partial B} = \sum_{j_2=1}^{J_2} -2\left[E_{i,2}(\hat{\theta}_{j_2,2}) - E_{i,1}(\hat{\theta}_{j_2,21})\right]\left[\sum_{z=0}^{C} z \, \frac{\partial P_{iz,1}(\hat{\theta}_{j_2,21})}{\partial B} \right], \tag{38}$$

and

$$\frac{\partial Q_{2i}}{\partial B} = \sum_{j_1=1}^{J_1} -2\left[E_{i,1}(\hat{\theta}_{j_1,1}) - E_{i,2}(\hat{\theta}_{j_1,12})\right]\left[\sum_{z=0}^{C} z \, \frac{\partial P_{iz,2}(\hat{\theta}_{j_1,12})}{\partial B} \right]. \tag{39}$$

Consequently, these calculations depend on particular partial derivatives of the category probability functions given in the right most terms of Equations 35, 36, 38, and 39. Each of these partial derivatives will be determined below.

The following definitions will be useful for calculating the partial derivatives of the category probability function:

$$b_{i,1} = \exp\left(\alpha_{i,1}\left[z\,(A\,\hat{\theta}_{j_2,2} + B - \delta_{i,1}) - \sum_{k=0}^{z} \tau_{ik,1} \right]\right), \tag{40}$$

$$\tilde{b}_{i,1} = \exp\left(\alpha_{i,1}\left[w\,(A\,\hat{\theta}_{j_2,2} + B - \delta_{i,1}) - \sum_{k=0}^{w} \tau_{ik,1} \right]\right), \tag{41}$$

$$b_{i,2} = \exp\left(\alpha_{i,2}\left[z\left(\frac{\hat{\theta}_{j_1,1} - B}{A} - \delta_{i,2} \right) - \sum_{k=0}^{z} \tau_{ik,2} \right]\right), \tag{42}$$

$$\tilde{b}_{i,2} = \exp\left(\alpha_{i,2}\left[w\left(\frac{\hat{\theta}_{j_1,1} - B}{A} - \delta_{i,2} \right) - \sum_{k=0}^{w} \tau_{ik,2} \right]\right), \tag{43}$$

$$c_{i,1} = \exp\left(\alpha_{i,1}\left[(M - z)\,(A\,\hat{\theta}_{j_2,2} + B - \delta_{i,1}) - \sum_{k=0}^{z} \tau_{ik,1} \right]\right), \tag{44}$$

$$\tilde{c}_{i,1} = \exp\left(\alpha_{i,1}\left[(M - w)\,(A\,\hat{\theta}_{j_2,2} + B - \delta_{i,1}) - \sum_{k=0}^{w} \tau_{ik,1} \right]\right), \tag{45}$$

$$c_{i,2} = \exp\left(\alpha_{i,2}\left[(M-z)\left(\frac{\hat{\theta}_{j_1,1} - B}{A} - \delta_{i,2} \right) - \sum_{k=0}^{z} \tau_{ik,2} \right] \right) , \tag{46}$$

and

$$\tilde{c}_{i,2} = \exp\left(\alpha_{i,2}\left[(M-w)\left(\frac{\hat{\theta}_{j_1,1} - B}{A} - \delta_{i,2} \right) - \sum_{k=0}^{w} \tau_{ik,2} \right] \right) . \tag{47}$$

With these definitions in place, the partial derivatives of the category probability functions can be calculated as follows:

$$\begin{aligned} \frac{\partial P_{iz,1}(\hat{\theta}_{j_2,21})}{\partial A} = {} & \Bigg[\left(\left(b_{i,1} z + c_{i,1}(M-z) \right) \left(\sum_{w=0}^{C} [\, \tilde{b}_{i,1} + \tilde{c}_{i,1} \,] \right) \right. \\ & \left. - (b_{i,1} + c_{i,1}) \left(\sum_{w=0}^{C} [\, \tilde{b}_{i,1} w + \tilde{c}_{i,1}(M-w) \,] \right) \right) \alpha_{i,1}\, \hat{\theta}_{j_2,2} \Bigg] \\ & / \left(\sum_{w=0}^{C} [\, \tilde{b}_{i,1} + \tilde{c}_{i,1} \,] \right)^2 , \end{aligned} \tag{48}$$

$$\begin{aligned} \frac{\partial P_{iz,2}(\hat{\theta}_{j_1,12})}{\partial A} = {} & \Bigg[\left(\left(b_{i,2} z + c_{i,2}(M-z) \right) \left(\sum_{w=0}^{C} [\, \tilde{b}_{i,2} + \tilde{c}_{i,2} \,] \right) \right. \\ & \left. - (b_{i,2} + c_{i,2}) \left(\sum_{w=0}^{C} [\, \tilde{b}_{i,2} w + \tilde{c}_{i,2}(M-w) \,] \right) \right) \left(\frac{-\alpha_{i,2}(\hat{\theta}_{j_1,1} - B)}{A^2} \right) \Bigg] \\ & / \left(\sum_{w=0}^{C} [\, \tilde{b}_{i,2} + \tilde{c}_{i,2} \,] \right)^2 , \end{aligned} \tag{49}$$

$$\frac{\partial P_{iz,1}(\hat{\theta}_{j_2,21})}{\partial B} = \left[\left(\left(b_{i,1} z + c_{i,1}(M-z) \right) \left(\sum_{w=0}^{C} [\tilde{b}_{i,1} + \tilde{c}_{i,1}] \right) - (b_{i,1} + c_{i,1}) \left(\sum_{w=0}^{C} [\tilde{b}_{i,1} w + \tilde{c}_{i,1}(M-w)] \right) \right) \alpha_{i,1} \right] / \left(\sum_{w=0}^{C} [\tilde{b}_{i,1} + \tilde{c}_{i,1}] \right)^2, \tag{50}$$

and

$$\frac{\partial P_{iz,2}(\hat{\theta}_{j_1,12})}{\partial B} = \left[\left(\left(b_{i,2} z + c_{i,2}(M-z) \right) \left(\sum_{w=0}^{C} [\tilde{b}_{i,2} + \tilde{c}_{i,2}] \right) - (b_{i,2} + c_{i,2}) \left(\sum_{w=0}^{C} [\tilde{b}_{i,2} w + \tilde{c}_{i,2}(M-w)] \right) \right) \frac{-\alpha_{i,2}}{A} \right] / \left(\sum_{w=0}^{C} [\tilde{b}_{i,2} + \tilde{c}_{i,2}] \right)^2. \tag{51}$$

Implementation of the Newton-Rapshon solution given in Equation 31 also involves the inversion of the matrix of second order partial derivatives of Q with respect to both A and B. Because the solution for $\hat{A}$ and $\hat{B}$ is based on a nonlinear least squares method, the second order partial derivatives can be approximated using the first order partial derivatives (Press, Teukolsky, Vetterling & Flannery, 1992). This is accomplished as follows:

$$\frac{\partial^2 Q}{\partial B^2} = 2 \sum_{i=1}^{I} \left(\left[\sum_{j_2=1}^{J_2} \left(\sum_{z=0}^{C} z \frac{\partial P_{iz,1}(\hat{\theta}_{j_2,21})}{\partial B} \right)^2 \right] + \left[\sum_{j_1=1}^{J_1} \left(\sum_{z=0}^{C} z \frac{\partial P_{iz,2}(\hat{\theta}_{j_1,12})}{\partial B} \right)^2 \right] \right), \tag{52}$$

$$\frac{\partial^2 Q}{\partial A^2} = 2\sum_{i=1}^{I}\left(\left[\sum_{j_2=1}^{J_2}\left(\sum_{z=0}^{C} z\,\frac{\partial P_{iz,1}(\hat{\theta}_{j_2,21})}{\partial A}\right)^2\right] + \left[\sum_{j_1=1}^{J_1}\left(\sum_{z=0}^{C} z\,\frac{\partial P_{iz,2}(\hat{\theta}_{j_1,12})}{\partial A}\right)^2\right]\right), \qquad \textbf{(53)}$$

and

$$\frac{\partial^2 Q}{\partial B\,\partial A} = \frac{\partial^2 Q}{\partial A\,\partial B} = 2\sum_{i=1}^{I}\left(\left[\sum_{j_2=1}^{J_2}\left(\sum_{z=0}^{C} z\,\frac{\partial P_{iz,1}(\hat{\theta}_{j_2,21})}{\partial B}\right)\left(\sum_{z=0}^{C} z\,\frac{\partial P_{iz,1}(\hat{\theta}_{j_2,21})}{\partial A}\right)\right]\right.$$

$$\left.+\left[\sum_{j_1=1}^{J_1}\left(\sum_{z=0}^{C} z\,\frac{\partial P_{iz,2}(\hat{\theta}_{j_1,12})}{\partial B}\right)\left(\sum_{z=0}^{C} z\,\frac{\partial P_{iz,2}(\hat{\theta}_{j_1,12})}{\partial A}\right)\right]\right). \qquad \textbf{(54)}$$

Table 1. Means and standard deviations of alternative equating constant estimates by sample size and equating condition. Standard deviations are given in parentheses below each mean value.

Horizontal Equating Condition

	A					B				
N	ICC	MS1	MM1	MS2	MM2	ICC	MS1	MM1	MS2	MM2
300	.998	1.002	.998	1.009	N/A	.007	.004	.003	.014	.001
	(.020)	(.070)	(.040)	(.072)	N/A	(.023)	(.040)	(.046)	(.047)	(.079)
1000	.997	.999	1001	1.003	N/A	-.002	-.003	-.002	-.001	-.001
	(.009)	(.030)	(.017)	(.031)	N/A	(.010)	(.020)	(.022)	(.016)	(.045)

Vertical Equating Condition

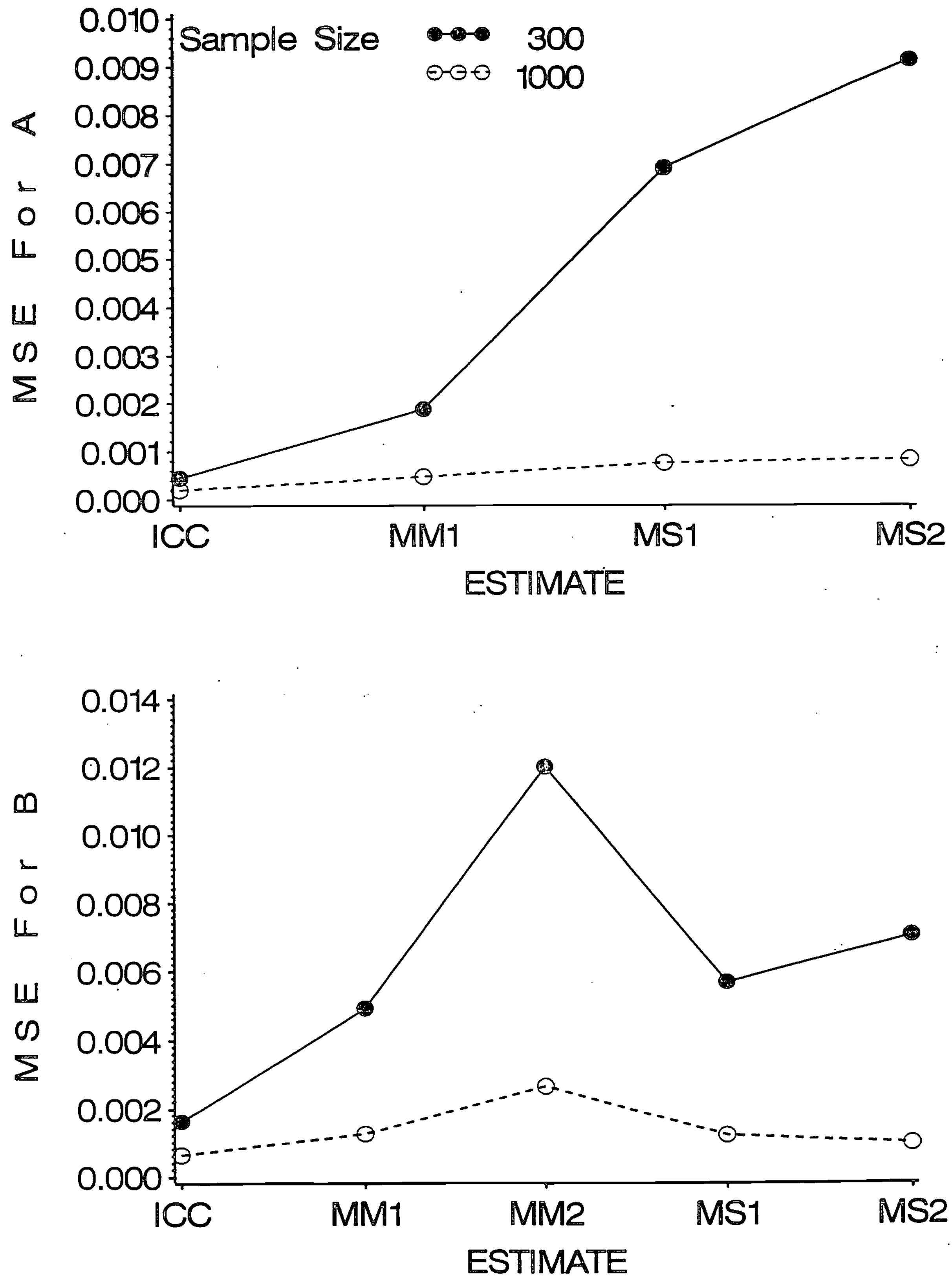
	A					B				
N	ICC	MS1	MM1	MS2	MM2	ICC	MS1	MM1	MS2	MM2
300	1.240	1.246	1.250	1.214	N/A	.505	.539	.536	.497	.545
	(.021)	(.097)	(.048)	(.111)	N/A	(.053)	(.093)	(.082)	(.111)	(.128)
1000	1.240	1.241	1.240	1.236	N/A	.512	.522	.523	.517	.523
	(.015)	(.025)	(.025)	(.024)	N/A	(.033)	(.040)	(.040)	(.038)	(.054)

Figure Captions

Figure 1. Mean squared error for $\hat{A}$ (top panel) and $\hat{B}$ (bottom panel) by calibration sample size and type of estimate.

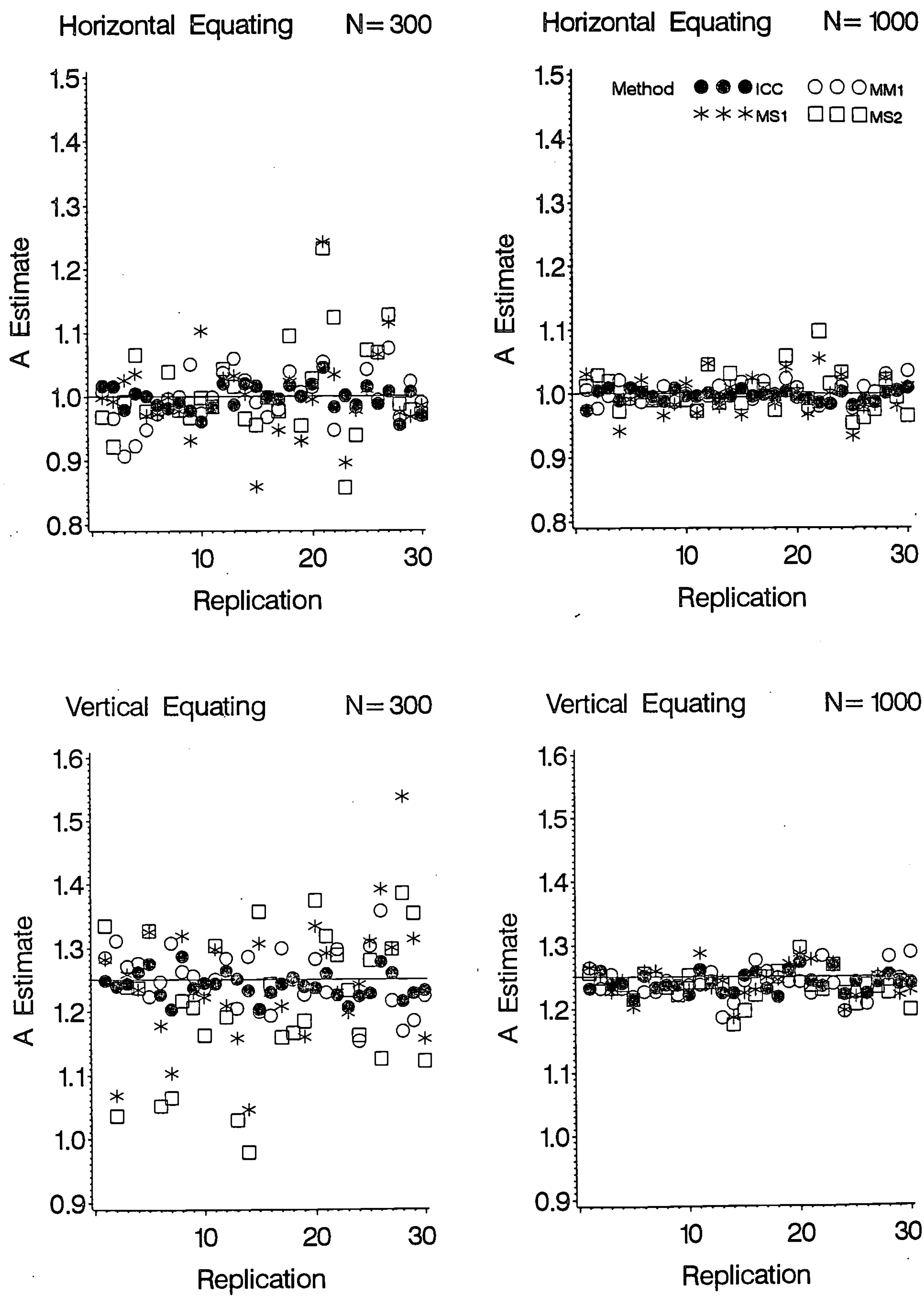
Figure 2. Scatterplots of $\hat{A}$ by replication and type of estimate. The horizontal reference line on each vertical axis indicates the true value of A.

Figure 3. Scatterplots of $\hat{B}$ by replication and type of estimate. The horizontal reference line on each vertical axis indicates the true value of B.



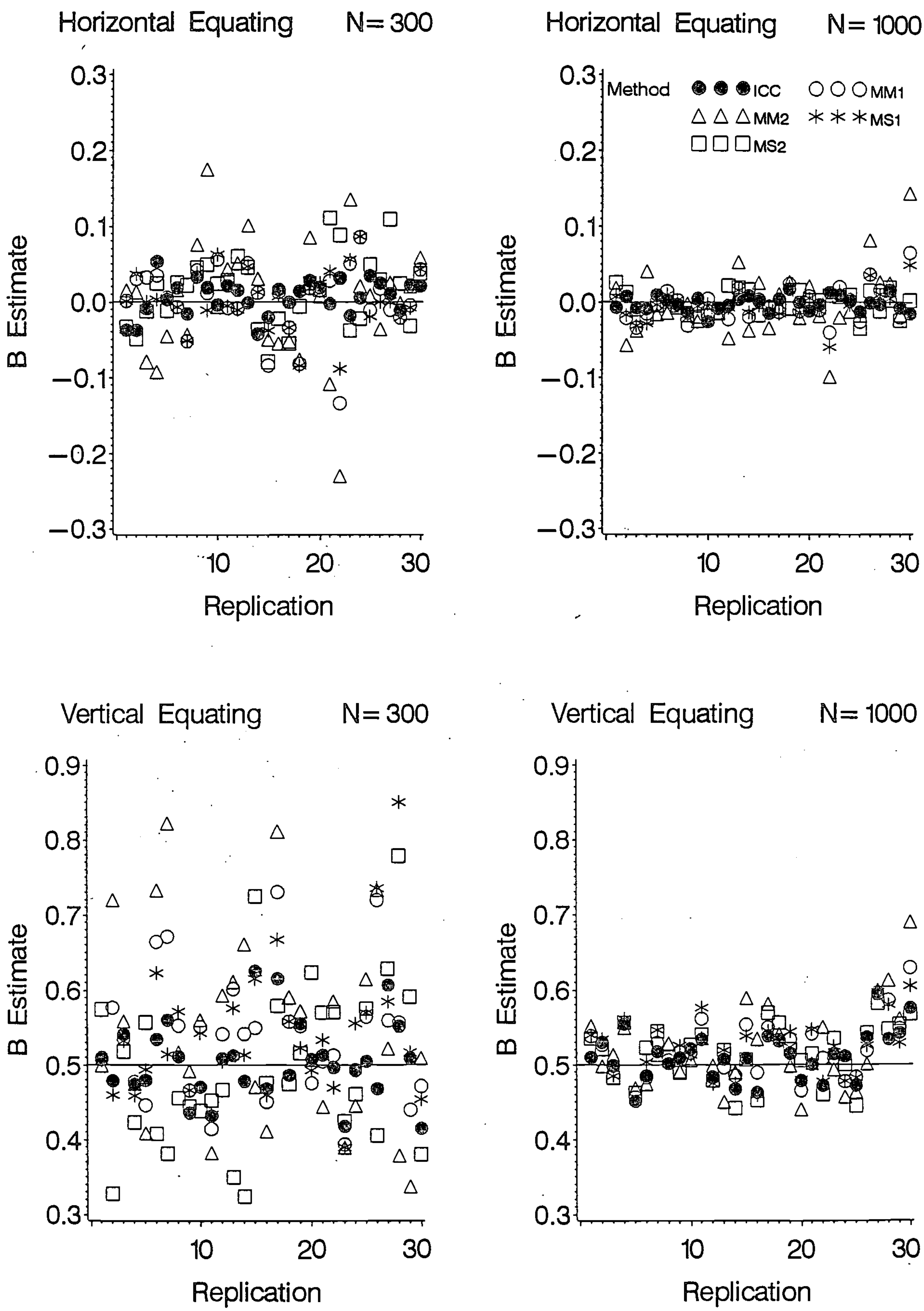
Sample Size
300
1000
MSE For A
0.010
0.009
0.008
0.007
0.006
0.005
0.004
0.003
0.002
0.001
0.000
ICC
MM1
MS1
MS2
ESTIMATE
MSE For B
0.014
0.012
0.010
0.008
0.006
0.004
0.002
0.000
ICC
MM1
MM2
MS1
MS2
ESTIMATE

Figure 2



Horizontal Equating N=300
Horizontal Equating N=1000
Method ICC MM1 MS1 MS2
A Estimate
Replication
Vertical Equating N=300
Vertical Equating N=1000

Figure 3



Horizontal Equating N= 300
Horizontal Equating N= 1000
Method ICC MM1 MM2 MS1 MS2
B Estimate
0.3
0.2
0.1
0.0
−0.1
−0.2
−0.3
10 20 30
Replication
Vertical Equating N= 300
Vertical Equating N= 1000
0.9
0.8
0.7
0.6
0.5
0.4
0.3

Signature: James S. Roberts	Printed Name/Position/Title: James S. Roberts / Assistant Professor	
Organization/Address: Department of Measurement, Statistics & Evaluation, University of Maryland, 1230F Benjamin Bldg., College Park, MD 20742	Telephone: 301-405-3630	Fax: 301-314-9245
	E-mail Address: jr245@umail.umd.edu	Date: 8/16/01

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:	
ERIC Clearinghouse on Assessment and Evaluation 1129 Shriver Laboratory (Bldg 075) College Park, Maryland 20742	**Telephone: 301-405-7449 Toll Free: 800-464-3742 Fax: 301-405-8134 ericae@ericae.net http://ericae.net**

EFF-088 (Rev. 9/97)