Research Report

# Studies of a Latent Class Signal Detection Model for Constructed Response Scoring II: Incomplete and Hierarchical Designs

**Lawrence T. DeCarlo**

March 2010

ETS RR-10-08

# Studies of a Latent Class Signal Detection Model for Constructed Response Scoring II: Incomplete and Hierarchical Designs

Lawrence T. DeCarlo
Teachers College, Columbia University

March 2010

**Technical Review Editor:** Dan Eignor

**Technical Reviewers:** Shelby Haberman and Sandip Sinharay

## Abstract

A basic consideration in large-scale assessments that use constructed response (CR) items, such as essays, is how to allocate the essays to the raters that score them. Designs that are used in practice are incomplete, in that each essay is scored by only a subset of the raters, and also unbalanced, in that the number of essays scored by each rater differs across the raters. In addition, all of the possible rater pairs may not be used. The present study examines the effects of these factors on parameter recovery and classification accuracy using simulations of a latent class model based on signal detection theory (SDT). Many tests also include more than one CR item, which introduces a nested or hierarchical structure into the design, in that raters are nested within essays (i.e., there are multiple raters per essay) and essays are nested within examinees (i.e., each examinee provides two or more essays). A hierarchical rater model (HRM) has previously been developed to recognize the nested structure. A version of the HRM that incorporates a latent class signal detection model in the first level, referred to as the HRM-SDT model, is presented. Parameter recovery in the HRM-SDT model is examined in simulations. The model is applied to data from several ETS tests.

Key words: Constructed responses, signal detection theory, balanced incomplete block, hierarchical rater model, latent class

## Acknowledgments

The author thanks Yoon Soo Park for his help with the unbalanced simulations, and YoungKoung Kim for her help with the hierarchical rater model study. The author wrote this report while under contract to ETS.

## Table of Contents

Page

## List of Tables

Many tests, such as the SAT® and GRE®, include constructed-response (CR) items, such as essays, in addition to multiple-choice (MC) items. The use of CR items necessitates the use of raters to score the items. A basic question is how to treat the rater scores. It is important to recognize, for example, that CR scores differ in a basic way from MC scores in that CR scores contain error, due to imperfect reliability of the raters and the possible presence of rater effects, whereas this is not the case for MC items, which can be objectively scored as right or wrong by either a person or a machine. The effects of rater reliability and rater effects in CR scoring are recognized in a latent class extension of signal detection theory (SDT; DeCarlo, 2002, 2005). In this approach, CR scores are viewed as being ordinal indicators of ordinal latent categories of essay quality. A prior report (DeCarlo, 2008a) examined parameter recovery and classification accuracy for latent class SDT models in fully crossed designs (where each essay is scored by every rater) and in incomplete designs (where each essay is scored by only some raters). The present report extends this research by examining some issues that arise when essays and raters are used in large-scale assessments.

First, in terms of the allocation of essays to raters, the designs used in large-scale assessments are necessarily incomplete, in that there are too many essays for a complete design to be used. Instead, each essay is typically scored by two raters out of a pool of perhaps several dozen raters. The effect of incompleteness was studied in prior research (DeCarlo, 2008a) by using a balanced incomplete block (BIB) design, which offers a useful baseline for comparison. In a BIB design, as applied to essay scoring, all possible rater pairs are used, each rater scores the same number of essays, and each pair of raters scores the same number of essays. Although a BIB design is statistically efficient, there are practical limitations to implementing it in large-scale assessments. For example, it is difficult to keep the number of essays scored by each rater equal given that, among other things, different numbers of examinees take the test each test day and the raters differ with respect to how long it takes to score a given set of essays. It also can be difficult to use all of the possible rater pairs, since there are a large number of possible pairs (e.g., 1,225 for 50 raters) and not all of the raters are available all of the time. So too, it is difficult to balance the number of essays scored by each pair of raters. As a result, in practice each rater typically scores a different number of essays and all of the rater pairs are not used, and so the design is not balanced with respect to the number of essays scored by each rater and with respect to the number scored by each rater pair. The present study uses simulations to examine the effects of these factors on parameter recovery and classification for a latent class SDT model.

Second, many tests include more than one CR item, and so another issue that arises is how to model the multiple sets of CR scores for each examinee. It is important to recognize that, for multiple CR items, there are not only multiple raters nested within each CR item (i.e., each essay is scored by more than one rater), but also multiple CR items nested within examinees (e.g., each examinee writes more than one essay). The nested structure means that the sets of CR scores should not be treated as being independent. To recognize the nested structure, a hierarchical rater model (HRM) has been introduced (see Patz, Junker, Johnson, & Mariano, 2002). In the version of the model presented here, which will be referred to as the HRM-SDT model, a latent class SDT model is used in the first level and an item response theory (IRT) model is used in the second level (DeCarlo, 2008b). Specifically, in the first level of the model, the CR scores are used as ordinal indicators of the quality of an essay, as in the usual latent class SDT model. In the second level the essay qualities are used as ordinal indicators of a continuous underlying ability, as in the usual IRT model. The approach recognizes the dependence that arises from having each examinee provide two essays; it also provides information about the difficulty and discriminability of the CR items. The current study presents results for simulations where parameter recovery for the HRM-SDT was examined. Also presented are applications of the model to large-scale tests that include two essays or three problem-solving exercises.

This report is organized as follows: First, the latent class signal detection model is reviewed. Some issues that arise with incomplete and unbalanced designs are discussed, followed by a presentation of results for simulations of unbalanced incomplete designs. Next, the HRM-SDT model is introduced, followed by simulations that examine parameter recovery. The last section applies the model to several real-world data sets.

## Latent Class Signal Detection Theory

Latent class signal detection theory has been discussed in a previous research report (DeCarlo, 2008a) and in several publications (DeCarlo, 2002, 2005), and so it is only briefly described here. Each essay is viewed as belonging to one of several latent categories of quality, with the latent categories defined by the scoring rubric. The task for each rater is to determine which category each essay belongs to. A basic idea of SDT is that a rater's judgment depends on

(a) his or her perception of the quality of the essay and (b) his or her use of response criteria, which reflect what the rater considers to be, for example, no mastery, little mastery, fair mastery, and so on; thus it is recognized in SDT that there may be individual differences in the way raters use the response categories. A rater is viewed as arriving at a judgment for each essay by using his or her perception of the essay together with response criteria.

SDT summarizes the data by providing two rater measures — a discrimination parameter *d,* which indicates how well the rater discriminates between the latent categories, and response criteria $c_k$, which reflect how the rater uses the response categories. In SDT, *d* has an interpretation as the distance between underlying probability distributions (e.g., of perception), whereas $c_k$ are criteria located relative to the underlying distributions. Figure 1 shows an example with four latent categories (and so there are four perceptual distributions, one for each latent category) and responses of one to four (and so there are three response criteria).

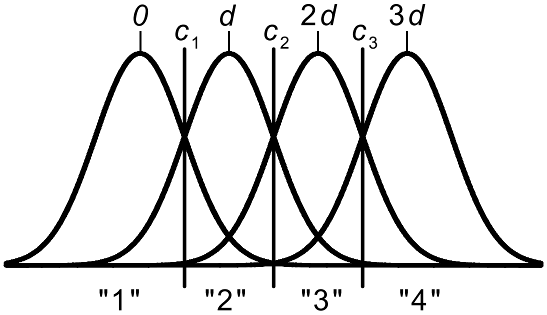

***Figure 1.* A representation of signal detection theory.**

### The Latent Class SDT Model

The above assumptions lead to a statistical model where the observed score of a rater serves as an ordinal indicator of the quality of an essay; the quality of an essay in turn is viewed as being a latent categorical variable on an ordinal scale. More specifically, the latent class SDT model for rater *j* is a model of the cumulative response probability given the latent category,

$$p(Y_j \leq k_j \mid X^{\#} = x^{\#}) = F(c_{jk} - d_j x^{\#}) \qquad (1)$$

(cf. DeCarlo, 1998), where $Y_j$ is the response variable for rater $j$, with values $k_j$ that range from 1 to $K_j$, $X^{\#}$ is a latent categorical variable with values of $x^{\#} = 0$ to $K-1$ (i.e., it is assumed that there are $K$ latent categories, defined by the scoring rubric), $c_{jk}$ are $K-1$ response criteria for the $j$th rater and $k$th response category, with $c_{j0} = -\infty$, $c_{jK} = \infty$, and $c_{j1} < c_{j2} <\ldots< c_{j,K-1}$, $d_j$ is a discrimination parameter for the $j$th rater, and $F$ is a cumulative distribution function (CDF). Note that the use of values of 0, 1,..., $K-1$ for $x^{\#}$ in Equation 1 implements an equal distance restriction, in that it constrains the distances between the underlying distributions to be equal for adjacent distributions, and so the distances are multiples of each other (i.e., the first distance is $d$, the second is $2d$, the third is $3d$; see DeCarlo, 2002, 2005, 2008a). As noted earlier, the model can be viewed as a type of discrete factor model, and is also related to located latent class models and discrete IRT models (DeCarlo, 2002, 2005).

The latent class SDT model can be incorporated into a restricted latent class model by using differences between the cumulative response probabilities to get the probability for each response category. Thus, for $K$ response categories,

$$\begin{array}{ll} p(Y_j = k_j \mid X^{\#} = x^{\#}) = F(c_{jk} - d_j x^{\#}) & k_j = 1 \\ p(Y_j = k_j \mid X^{\#} = x^{\#}) = F(c_{jk} - d_j x^{\#}) - F(c_{jk-1} - d_j x^{\#}) & 1 < k_j < K_j. \\ p(Y_j = k_j \mid X^{\#} = x^{\#}) = 1 - F(c_{jk-1} - d_j x^{\#}) & k_j = K_j \end{array} \qquad (2)$$

The above probabilities are used in a restricted latent class model (see Clogg, 1995; Dayton, 1998), which is a model for the response patterns across raters,

$$p(Y_1 = k_1, Y_2 = k_2,\ldots, Y_j = k_j) = \Sigma_{x\#}\, p(X^{\#} = x^{\#})\, \Pi_j\, p(Y_j = k_j \mid X^{\#} = x^{\#}), \qquad (3)$$

where the summation is over the values of the latent classes (i.e., $x^{\#}$), the product is over the $J$ raters, and an assumption of local independence is made, that is,

$$p(Y_1 = k_1, Y_2 = k_2,\ldots, Y_j = k_j \mid X^{\#}) = \Pi_j\, p(Y_j = k_j \mid X^{\#}).$$

The above represents a standard restricted latent class model, and so it can be fit with several software packages, such as LEM (Vermunt, 1997), Latent Gold (Vermunt & Magidson, 2007), or

Mplus (Muthén and Muthén, 2007; note that the latent class SDT model can be implemented in Mplus as a nonparametric confirmatory factory analysis model).

### Incomplete Designs

Complete (fully crossed) designs, where each essay is scored by every rater, are not used in practice for large-scale assessments because there are typically a large number of essays to be scored and there are limitations to the number of essays each rater can score. This necessitates the use of incomplete designs; for example, each essay in large-scale assessments typically is scored by two raters out of a pool of raters. The essays can be allocated to the raters according to different rating designs (see Hombo, Donoghue, & Thayer, 2001). Prior research on the latent class SDT model has examined balanced incomplete block (BIB) designs (DeCarlo, 2008a), which are efficient (Fleiss, 1986) and serve as a useful baseline. In the current context, a BIB design uses all possible pairings of the raters, balances the number of essays scored by each rater, and balances the number of essays scored by each pair of raters. The present report extends prior research by examining, in addition to BIB, unbalanced designs, where raters score different numbers of essays and not all rater pairs are used; this is closer to the type of design that is used in real-world assessments, as noted below.

*Estimation*. The essays are allocated to the raters, and so the data in rating designs are missing by design. To use the terminology of Rubin (1976), data that are missing by design are *missing completely at random* (MCAR; note that the results are also valid with the weaker assumption of missing at random). The approach to fitting the model with missing data is to maximize the likelihood for the various subsets of raters that score each essay. First, consider the situation where three raters grade each and every essay in a fully crossed design. The log likelihood is

$$\log L = \Sigma_i \log \Sigma_{x\#}\, p(X^{\#} = x^{\#})\, p(Y_{i1} = k_1 |\, X^{\#} = x^{\#})\, p(Y_{i2} = k_2 |\, X^{\#} = x^{\#})\, p(Y_{i3} = k_3 |\, X^{\#} = x^{\#}), \qquad (4)$$

where $Y_{ij}$ is the score for examinee (essay) *i* for rater *j*. Next, consider an incomplete design where Essay 1 is scored by only Raters 1 and 2, Essay 2 is scored by only Raters 1 and 3, and so on. Then the log likelihood for the first case, with Raters 1 and 2, is

$$\log L_1 = \log \Sigma_{x\#}\, p(X^{\#} = x^{\#})\, p(Y_{11} = k_1 |\, X^{\#} = x^{\#})\, p(\mathrm{Y}_{12} = k_2 |\, X^{\#} = x^{\#}),$$

and for the second case, with Raters 1 and 3, is

$$\log L_2 = \log \Sigma_{x\#}\, p(X^{\#} = x^{\#})\, p(Y_{21} = k_1 | X^{\#} = x^{\#})\, p(Y_{23} = k_3 | X^{\#} = x^{\#}),$$

and similarly for the remaining cases. Thus, in the presence of missing data, the log likelihood is based on the subsets of raters for which we have observations, and so all available information is used for each case. See the technical manual of Latent Gold (Vermunt & Magidson, 2005) for further details on estimation in latent class models with missing data using maximum likelihood (or posterior mode) estimation.

**Connectedness.** When using fewer than all of the possible pairs of raters, one has to take care that the raters are all *connected*, so that *d*, for example, is on a common scale and can be compared across all of the raters; any non-connected raters cannot be compared to the other raters. For example, if Raters 1 and 2 grade one set of essays and Raters 3 and 4 grade a second set of essays with no overlap of essays, then Raters 1 and 2 can be compared to each other, but they cannot be compared to Raters 3 or 4 (without making further assumptions). On the other hand, all of the raters can be compared if Rater 1 has some overlapping essays with Rater 3, for example (i.e., at least three rater pairs are needed for all of the raters to be connected when there are four raters). The issue is related to earlier discussions in the literature about comparing treatment effects in block designs, in terms of whether treatment contrasts are estimable (e.g., Eccleston & Hedayat, 1974; Weeks & Williams, 1964); note that in the current context, the essays (examinees) correspond to blocks and the raters correspond to treatments.

For the situation with *J* raters and two raters per block (essay), at least *J*−1 rater pairs must be used in order for all of the raters to be connected. For example, one can simply pair *J*−1 of the raters with the *J*th rater, that is, for 10 raters, one can use the nine pairs (1, 10), (2, 10), (3, 10), and so on, with the result that all of the raters will be fully connected. Simply put, each rater is paired with the same rater (e.g., Rater 10), and so all of the raters can be compared to each other. Another option is to use a spiral-like design, with the nine pairs being (1,2), (2,3), (3,4), and so on (which is simple to implement and helps with balancing). This design was used here for the unbalanced condition because it is simple to implement, is used in practice, and allows one to create unbalanced data with some control over how many essays each rater scores.

**Lack of balance.** The present report examines the effect of a lack of balance, that is, situations where the raters score different numbers of essays instead of an equal number. The type of design used is shown in Table 1, with the rows referring to groups of essays and the columns to raters. For example, the first row of Table 1 shows that Raters 2 and 5 score the same 20 essays. The last row of the table shows that the 10 raters score a total of 50, 60, 120, 140, 200, 230, 280, 310, 370, or 400 essays each. Note that the basic nine adjacent pairs needed for the raters to be fully connected are used, along with a 10th pair that was added so that each rater is paired with two other raters. Thus (10/45) × 100 = 22% of the possible 45 rater pairs are used, which is close to the lower limit for connectedness of 20% (i.e., for 10 raters, a minimum of nine rater pairs are required, giving (9/45) × 100 = 20%). The total number of essays scored is 1,080, with each rater scoring an average of 216 essays (to match the BIB design). Note that for the BIB condition (and in the original fully crossed data), the population values of *d* were ordered, in increasing magnitude, from Rater 1 to 10 (and the incomplete data were created from the original fully crossed data, see below). Thus for the incomplete design the raters were randomly allotted to the 10 columns of the design, as shown in Table 1, so that the sample size was not systematically related to the value of *d*.

**Table 1**

***An Unbalanced Incomplete Design (N = 1,080)***

| | Rater | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | 2 | 5 | 6 | 4 | 3 | 9 | 10 | 1 | 8 | 7 | Total |
| | 20 | 20 | | | | | | | | | |
| | | 40 | 40 | | | | | | | | |
| | | | 80 | 80 | | | | | | | |
| | | | | 60 | 60 | | | | | | |
| | | | | | 140 | 140 | | | | | |
| | | | | | | 90 | 90 | | | | |
| | | | | | | | 220 | 220 | | | |
| | | | | | | | | 150 | 150 | | |
| | | | | | | | | | 250 | 250 | |
| | 30 | | | | | | | | | 30 | |
| | | | | | | | | | | | 1080 |
| Total/Rater | 50 | 60 | 120 | 140 | 200 | 230 | 310 | 370 | 400 | 280 | |

**Linkage.** As noted above, a design with two raters per essay can be fully connected with the use of only $J$–1 rater pairs out of a possible [($J$ × ($J$–1)]/2 rater pairs. For example, if there are 40 raters, then the raters will be fully connected with the use of 39 rater pairs out of a possible 780 pairs (i.e., 5% of the total possible unique pairs). Another issue that arises has to do with the *degree of linkage* in the design. The degree of linkage refers (in part) to the percent of the possible rater pairs used in the design. A basic question of interest is whether there is any benefit to using more than the minimum number of rater pairs needed for connectedness. This will be examined in the present report in both balanced and unbalanced designs by increasing the number of rater pairs and examining the effects on estimation and classification.

An aspect of the degree of linkage in incomplete designs that has not, to my knowledge, been noted before is that it affects the relative sparseness of the data. For example, consider the situation with 40 raters using a 1 to 5 response. For a BIB design with two raters per essay, there are $5^2$ × (40 × 39)/2 = 19,500 possible response patterns, and so unless the sample size is very large, there likely will be many patterns with few or no observations, and so the data will be quite sparse. However, for a spiral design, using the minimum number of rater pairs for connectedness (i.e., 39), there are only $5^2$ × 39 = 975 possible response patterns, and so the number of patterns with no observations will be smaller and the frequencies per response pattern will be larger. In addition, with the spiral design the number of essays scored by each rater pair will be larger than for the BIB design. Both of these factors might have effects on estimation that offset any loss due to having fewer rater pairs. The effects of linkage will be examined here using conditions where both the balance and linkage are varied.

**Boundary problems and posterior mode estimation.** It was noted in a previous research report that a problem arises when the latent class SDT model is applied to data from incomplete designs (DeCarlo, 2008a). The problem is that estimates of the parameters that are on the boundary are often found, such as estimates of zero (or one) for one or more of the latent class sizes, or estimates of $d$ that are infinite or large with indeterminate standard errors. Problems of this sort are well known in latent class analysis (e.g., Clogg & Eliason, 1987; Maris, 1999), and several solutions have been proposed. One approach is (partly) Bayesian, in that the parameters are treated as random variables instead of as fixed values to be estimated. For example, the latent classes have a multinomial distribution with parameters, $p(X^{\#})$, that are the latent class sizes; the parameters can be assumed to also have a probability distribution, such as the Dirichlet, which is a conjugate prior for the multinomial, in that the posterior is also Dirichlet.

This approach is used in *posterior mode estimation* (PME, also called *maximum a posteriori estimation*; see Galindo-Garre & Vermunt, 2006; Gelman, Carlin, Stern, & Rubin, 1995; Maris, 1999; Schafer, 1997; Vermunt & Magidson, 2005). In posterior mode estimation, the maximum of the log posterior function is found, where the log posterior function is the log likelihood function plus a log prior. Posterior mode estimation can be viewed as maximum likelihood estimation (MLE) with a penalty for solutions that are close to the boundary; it basically smooths the estimates away from the boundary. Galindo-Garre and Vermunt (2006) presented a simulation that suggested PME gave more reliable parameter estimates and standard errors than either MLE or parametric bootstrapping. Galindo-Garre, Vermunt, and Bergsma (2004) presented a simulation that suggested advantages of PME over a full Bayesian approach via Monte Carlo Markov chain (MCMC) methods; they also noted that PME is less computationally intense than MCMC. In the context of latent class SDT, simulations presented in DeCarlo (2008a) suggest that PME (using Bayes constants of 1, see below) ameliorated boundary problems and led to good recovery of the population parameters and standard errors (for a range of values found in real-world data).

With respect to priors, Galindo-Garre and Vermunt (2006) examined several noninformative priors (Jeffreys, normal, and Dirichlet) in a simulation and suggested versions of the Dirichlet prior that appeared to perform well (and are simple to implement); these priors are implemented in Latent Gold 4.5, which was used here. With respect to the latent class sizes, the prior smooths the estimates, slightly, towards equality, which helps prevent the occurrence of class sizes of zero or one. With respect to the rater responses, the prior makes the conditional response probabilities slightly more equal and so, in the present context, the prior smooths estimates of $d$ towards zero and thereby prevents indeterminate values; the version of the Dirichlet prior used in Latent Gold for the response variables also preserves their marginal distributions (for details see Vermunt & Magidson, 2005; in the context of SDT this basically means that the smoothing pertains primarily to $d$, and not $c_{jk}$). The hyperparameters of the Dirichlet prior are specified in Latent Gold through the use of Bayes constants (which also have an interpretation in terms of adding observations to the data). The simulations presented below offer additional information about the performance of PME with Dirichlet priors in latent class SDT and in a hierarchical SDT rater model.

## Simulations: Unbalanced, Incomplete Designs

This section presents simulations of BIB designs and unbalanced designs with both full and partial linkage. It should be noted that the design of the simulations was guided in part by the observations that, for several large-scale tests that were examined, anywhere from 200 to10,000 essays were obtained on any given test day; the essays for a given test day were in turn scored by anywhere from 10 to 80 raters, with each rater scoring anywhere from 1 to 600 essays. Thus for the simulations, the basic design consisted of 10 raters scoring 1,080 essays (because the BIB is fully balanced for 1,080), with each rater scoring 216 essays (or an average of 216 in the unbalanced condition), so that the number of raters, number of essays, and essays scored per rater are all comparable to those found in practice.

For the BIB design with full linkage, all 45 possible rater pairs were used, each rater scored 216 essays, and each rater pair scored 24 essays. Also examined is an unbalanced design with full linkage; the number of essays scored per rater ranged from 60 to 400, and all 45 rater pairs were used. In BIB and unbalanced designs with near minimum linkage, close to the minimum number of rater pairs was used (10 out of 45, or 22%; note that 9 out of 45, or 20%, is the minimum for connectedness) to obtain information about estimation with a near-minimum number of rater pairs (while still maintaining balance in that each rater pair appears twice). BIB and unbalanced designs with moderate linkage, 20 out of 45 rater pairs, also were used to examine the effects of degree of linkage.

### Methods

The simulated data were generated using SAS macros written by the author; earlier macros were modified as needed for the current studies. Data for 10 raters discriminating between six latent classes by giving 1-6 responses were simulated. The latent class sizes were chosen to approximate a normal distribution (.08, .17, .25, .25, .17, .08), which is consistent with results found in previous research. Prior research has found a mean value of *d* for several large-scale assessments in the vicinity of 3 to 4 with an approximate normal distribution. Thus for all conditions a range of values of *d* from 1 to 5 were used, which covers a range of detection from poor to excellent (for the logistic model) and is consistent with that found for real-world data.

The distribution of *d* was approximately normal, with values for the 10 raters of 1, 2, 2, 3, 3, 3, 3, 4, 4, and 5. As in prior research (DeCarlo, 2008a), the response criteria were located at the intersection points of adjacent distributions, which has the convenient property that the relative locations of the criteria remain the same as *d* varies; it also appears to be a reasonable approximation to what is found with real-world raters (see DeCarlo, 2008a). Specifically, relative to *d*, the first through last criteria are located at ½*d*, 1½*d*, 2½*d*, and so on, because the intersection points for symmetrical distributions are midway between two adjacent distributions. So, for example, for six latent classes and a *d* of 2, the six distributions will be at 0, 2, 4, 6, 8, and 10, and the five response criteria will be at 1, 3, 5, 7, and 9, and similarly for other values of *d*.

Specifics of data generation are given in a prior report (DeCarlo, 2008a). Data were first generated for a fully crossed design. The BIB and unbalanced incomplete data were then created from the fully crossed data by inserting missing values according to the design. For example, for the BIB, blocks of all possible pairs were created and all other data points were set to missing. For the unbalanced simulation, missing values were inserted according to the design (e.g., Table 1). Each condition consisted of 100 replications. The specific designs used for both the balanced and unbalanced conditions, with BIB and spiral designs, are shown in Appendix A.

To see how the total sample size was determined for the BIB condition, note that the following conditions hold for a BIB design (see Fleiss, 1986):

$$gr = nk,$$

$$g \leq n,$$

$$\lambda\,(g - 1) = r\,(k - 1), \qquad (5)$$

where *g* is the number of raters, *r* is the number of essays scored by each rater, *n* is the number of examinees (essays), *k* is the number of raters that score each essay (i.e., the block size), and λ is the number of essays scored by each pair of raters. For the design examined here, it follows from the first equation that 10 × 216 = 1,080 × 2, which shows that when each essay is scored by 2 raters out of 10, a total of 1,080 essays are scored if each rater scores 216 essays. Further, substituting 10 for *g*, 216 for *r*, and 2 for *k* in the last relation of Equation 5 shows that each rater pair scores 24 essays (which is λ). Note that if one wanted to use 1,000 essays (*n*) with each rater scoring 200 essays (*r*), then it follows from Equation 5 that λ is not a whole number (it is 22.22);

this shows that Equation 5 imposes constraints on *n* and *r* (given a choice of the number of raters, *g*, and the number of raters per essay, *k*) if λ (the number of essays scored by each rater pair) is to be a whole number.

Several software packages can be used to fit the latent class SDT model, such as LEM (Vermunt, 1997), Latent Gold (Vermunt & Magidson, 2007), and Mplus (Muthén & Muthén, 2007); Latent Gold (Version 4.5) was used here. Latent Gold uses the EM algorithm followed by the Newton-Raphson procedure to obtain maximum likelihood estimates of the parameters. Only minor modifications of the algorithms for MLE are needed for PME; Schaefer (1997) noted, for example, that any algorithm used for MLE, such as the iterative proportional fitting algorithm commonly used for log-linear models, can easily be modified to find posterior modes when Dirichlet priors are used for the conditional response probabilities and latent class probabilities (p. 307; also see Gelman et al., 1995).

A SAS macro written by the author was used to generate 100 input files for the Latent Gold analysis and also a DOS batch file, which was used to call Latent Gold repeatedly to perform the analysis. Other SAS macros stripped out information from the Latent Gold output for each replication, and the results were combined in a file for the remaining analyses. The SAS macro that stripped out and summarized the data checked and corrected for label switching, as described in a previous report (DeCarlo, 2008a). Another problem is that the solution could represent a local maximum; to decrease the likelihood of this, the number of sets of starting values was increased from the default of 10 to 20. One also has to check that the solution converged before reaching the maximum number of iterations.

## Results

**Rater parameters and latent class sizes.** Appendix B presents, for the rater parameters and latent class sizes, the population parameters, the mean parameter estimates, the bias, the percent bias (the bias divided by the population value, times 100), and the mean squared error (MSE) for fits of the model to the 100 sets of simulated data.

Table B1 shows results for the BIB condition, where all 45 possible rater pairs were used. The table shows that, for the rater discrimination parameter *d*, recovery is good, with a percent bias of less than 10% for all 10 raters; the MSE is also small, less than 0.6. The bias tends to be negative, which means that *d* tends to be underestimated. The table shows that the percent bias is largest for the largest value of *d* (5, for Rater 10). With respect to *c*, the bias and MSE tend to be larger than those for *d*, but with percent bias still generally under 10%. The percent bias tends to be large for the first criterion for each rater; however, this depends in part on the arbitrary location of the zero point (which is why the percent bias is more meaningful for slope parameters like *d* and less so for location parameters like *c*). The percent bias is larger for criteria associated with the largest values of *d*, such as for Rater 10. With respect to estimation of the latent class sizes, the percent bias is less than 10% except for the end classes, which have large positive bias; in particular, the latent class sizes of 0.08 tend to be over-estimated as 0.10.

Table B2 shows results for the unbalanced design, where all 45 rater pairs were again used, with the raters scoring different numbers of essays. With respect to *d*, the bias is less than 10% in all cases, with three notable exceptions. First, the percent bias is large for the two raters who scored the fewest number of essays (Raters 2 and 5, who scored 50 and 60 essays, respectively, as shown in Table 1). Second, as in the BIB condition, the bias is large for the largest value of *d* (5, for Rater 10). Thus the results suggest that a lack of balance leads to larger bias. With respect to the response criteria *c*, the same patterns as found in the BIB condition appear, in that the percent bias is larger for Raters 2 and 5, who graded the smallest number of essays, and Rater 10, who has the largest value of *d*. With respect to the latent class sizes, there is virtually no difference from the balanced design, in that the middle class sizes are well recovered, with a percent bias of less than 10%, whereas the end classes are overestimated (again with 0.08 estimated as 0.10).

Table B3 presents results for 10 raters each scoring 216 essays in a balanced design with "near-minimum" linkage, in that only 10 out of 45 possible rater pairs were used (the design is shown in Appendix A). The percent bias is again generally less than 10%, and the bias tends to be negative. The percent bias for raters with the largest values of *d*, Raters 4 and 5, is large and negative, and so high values of discrimination tend to be underestimated, as found above. Compared to the BIB design with 45 rater pairs, the bias for *d* is larger but only slightly so, with just two raters (Raters 8 and 10, with the largest values of *d*) having a percent bias greater than 10%. The bias tends to be larger for the response criteria, with the largest bias for the raters with the largest values of *d*. With respect to the estimates of the latent class sizes, the results are nearly identical to those for the BIB design, with good estimation of the middle class sizes and positive bias for the end classes.

Table B4 shows results for 10 raters in an unbalanced design with near-minimum linkage (i.e., 10 out of 45 possible rater pairs). The effect of a lack of balance again appears to be an increase in the bias, with the largest effects generally for the raters with the smallest sample sizes or the largest values of *d*. With respect to estimation of the latent class sizes, the results are nearly identical to those found in the other conditions.

Tables B5 and B6 show results for a spiral-like design with moderate linkage (20 out of 45 possible rater pairs). With respect to the rater parameters, the results are comparable to those obtained for the near-minimum linkage condition, except that the bias and MSE are slightly smaller. Estimation of the latent class sizes is also comparable to that found in the other conditions. Thus it appears that increasing the number of rater pairs from 10 to 20 offers at most a small improvement in recovery of the rater parameters or the latent class sizes.

**Standard errors.** Appendix C presents results for the evaluation of the estimates of the standard errors of *d* and the standard errors of the latent class sizes. The standard error estimates are computed using standard asymptotic theory (i.e., using the inverse of the observed information matrix; for details see Vermunt & Magidson, 2005). The bias is obtained by computing the standard deviation of the parameter estimates across the 100 replications (SD in the tables, which serves as the population value), and subtracting it from the mean of the estimated standard errors (i.e., across the 100 replications, which is the Mean SE shown for each parameter in the table).

Table C1 shows that the bias in the estimates of the standard errors is generally less than 10%, except for large values of *d* (4 or 5), where the SEs tend to be overestimated. The bias for the SEs of the latent class sizes is less than 10%, except for the sixth latent class. Table C2 shows that a lack of balance leads to larger bias (greater than 20%) for the SEs of *d* for the two raters with the smallest samples sizes (50 and 60); the bias is positive, and so the SEs are overestimated. For the remaining raters, however, the bias is similar to that found in Table C1, with overestimation of the SEs for large values of *d*. The bias for the SEs of the latent class sizes in Table C2 is also comparable to that found in Table C1.

Tables C3 and C4 show results for the condition with only 10 rater pairs. Table C3 shows that the percent bias for the SEs of the *d* parameters is generally large and positive, and so the SE's tend to be overestimated. The bias for the latent class sizes is smaller and similar to that found in Tables C1 and C2. Table C4 shows similar results, with about the same magnitude of bias. Thus, whereas the tables in Appendix B show that using a near-minimum number of rater pairs has little deleterious effect on estimation of the rater parameters, Tables C1 and C3 show that there is a fairly large effect on the estimates of the standard errors of  the rater parameters, which are generally overestimated. The estimates of the standard errors for the latent class sizes, on the other hand, do not appear to be heavily affected by using fewer rater pairs or by having a lack of balance, as shown in Tables C1 through C4.

Table C5 shows that, for the condition with 20 rater pairs, the bias of the SEs is generally small (i.e., under 10%) for the rater parameters and latent class sizes, except for large values of *d*. In contrast, Table C6 shows that the bias for the SEs of the rater parameters is large in the unbalanced condition. As in Tables C1 to C4, Tables C5 and C6 suggest that the SEs of the latent class sizes are adequately estimated.

Overall, the results shown in Appendices B and C suggest that the estimation of *d* and its standard error are good for values of *d* in the range of 1 to 5; estimation of *c* tends to be poorer, depending in part on the value of *d*. Tables B1 to B6 show that using an unbalanced design increases bias for the rater parameters in both complete and incomplete linkage conditions. However, estimates of the latent class sizes and their standard errors do not appear to be heavily affected by a lack of balance. With respect to linkage, decreasing the linkage by using 10 or 20 rater pairs in lieu of 45 led to perhaps a small increase in the bias, but the effect was relatively small. The effect of decreasing linkage appears to be primarily on the estimates of the standard errors of the rater parameters, which are overestimated; the latent class sizes and their standard errors, on the other hand, are reasonably well estimated in all conditions.

**Classification.** Table 2 shows results for classification accuracy (proportion correctly classified) for the six conditions discussed above. Proportion correct (PC) is the estimated proportion of cases that are correctly classified and is obtained from the posterior probabilities (for a fit of the model see, e.g., Clogg, 1995); note that PC is available for both simulated and real-world data (i.e., upon fitting the model, the PC can be estimated). In contrast, this is not the case for $\text{PC}_{\text{obt}}$, which is only available in a simulation; specifically, $\text{PC}_{\text{obt}}$ is the obtained (not estimated) proportion of cases that were actually correctly classified in the simulation (i.e., $\text{PC}_{\text{obt}}$ is computed by comparing the classifications obtained from the posterior probabilities to the true latent classes; of course the true latent class for each case is only known in a simulation). Similarly, $\text{PC}_{\text{av}}$ is the proportion of cases that were correctly classified in the simulation by using the obtained average score (rounded both up and down; the rounding that gave the largest value of $PC_{av}$ is the one that is reported). Table 2 also shows two measures of association between the classifications and the true latent classes, namely the Pearson correlation $r$ and tau-b, again for both the model-based classifications (with subscript *obt*) and the average scores (with subscript *av*).

**Table 2**

***Estimated and Obtained Proportion Correct and Correlations With True Latent Classes, Balanced and Unbalanced Designs, Varying Linkage***

| Pairs | PC | $PC_{obt}$ | $PC_{av}$ | $r_{obt}$ | $r_{av}$ | $\tau_{b\text{-}obt}$ | $\tau_{b\text{-}av}$ |
|---|---|---|---|---|---|---|---|
| 45 balanced | 0.747 | 0.730 | 0.661 | 0.927 | 0.910 | 0.874 | 0.846 |
| 45 unbalanced | 0.770 | 0.765 | 0.646 | 0.939 | 0.903 | 0.892 | 0.835 |
| 20 balanced | 0.740 | 0.719 | 0.668 | 0.922 | 0.911 | 0.868 | 0.848 |
| 20 unbalanced | 0.755 | 0.735 | 0.658 | 0.925 | 0.902 | 0.873 | 0.835 |
| 10 balanced | 0.744 | 0.733 | 0.654 | 0.931 | 0.910 | 0.880 | 0.846 |
| 10 unbalanced | 0.763 | 0.767 | 0.646 | 0.940 | 0.904 | 0.894 | 0.836 |

*Note.* PC = estimated proportion correct; $PC_{obt}$ = obtained (in the simulation) proportion correct; $PC_{av}$ = obtained (in the simulation) proportion correct using the average score; $r_{obt}$ and $\tau_{b\text{-}obt}$ are the obtained Pearson correlation and tau-b; $r_{av}$ and $\tau_{b\text{-}av}$ are the obtained correlation and tau-b for the average scores.

Several results are apparent in Table 2. First, the estimated proportion correctly classified, PC, tends to overestimate the proportion actually correctly classified in the simulation (i.e., $PC_{obt}$), although the overestimation is very small, generally around .02 or less. Overestimation of $PC_{obt}$ by PC has been noted earlier (DeCarlo, 2005, 2008a); note that a comparison of the results shown in Table 2 with those shown in Table 6 of DeCarlo (2005) shows that the overestimation is larger for smaller sample sizes (about 5% for a sample size of 300 and over 20% for a sample size of 100, compared to the 2% found here for a sample size of 1,080).

Second, Table 2 shows that the proportion correctly classified using the average score, $PC_{av,}$ is in every case 5% to 10% lower than the proportion correctly classified using the model (i.e., $PC_{obt}$). This shows that there is clearly a benefit to using the model-based classifications over the average scores, as found in other studies (DeCarlo, 2008a), assuming of course that the SDT model is appropriate. Table 2 also shows that the correlations of the average scores with the true latent classes, $r_{av}$ and $\tau_{b-av}$, also tend to be smaller than the correlations for the model-based classifications, $r_{obt}$ and $\tau_{b-obt}$.

With respect to a lack of balance, Table 2 shows that $PC_{obt}$ is actually slightly larger in the unbalanced conditions than in the balanced conditions (perhaps because some of the raters score more essays), and so a lack of balance appears to have no detrimental effect on classification accuracy. With respect to the classifications obtained by using the average rater scores, a lack of balance reduces $PC_{av}$ by only about 1%.

With respect to linkage, an interesting result shown in Table 2 is that the number of rater pairs used (i.e., 45, 20, or 10) appears to have little influence on classification accuracy. For example, correct classification ($PC_{obt}$) for the spiral design conditions with only 10 rater pairs is as high (.733 and .767 for balanced and unbalanced) as in the BIB conditions with 45 rater pairs (.730 and .765).

## Discussion

These simulations examine parameter recovery and classification in balanced and unbalanced designs with varying linkage. The results show that estimation is affected by a lack of balance and by having less than full linkage; however, the rater discrimination parameter *d* is generally well recovered, as are the latent class sizes. The largest bias is associated with the largest values of *d*, which tend to be underestimated. Note that underestimation of large values of *d* means that the raters actually are performing better than indicated by the estimate (which is not a problem, but it would be a problem if the opposite occurred — if small values of *d* were overestimated, then raters who perform poorly would appear to be better than they really were). The results also suggest that a lack of balance has a larger effect on parameter recovery than the number of rater pairs used (i.e., the degree of linkage); the effect of linkage on parameter estimation was fairly small (in terms of bias). Thus the results show that using a design with less than full linkage appears to result in little loss with respect to parameter recovery, with the main effect appearing to be larger standard errors.

The results for classification accuracy are of particular interest. First, the results in Table 2 show that there is nearly a 10% increase in classification accuracy obtained by using model-based classifications in lieu of average scores. Of course the increase in PC depends on the parameters and design (and on the validity of the SDT model). However, other studies that have used different parameters and designs also have found increases in classification accuracy in the range of 5% to 10% when model-based classifications were used in lieu of average scores (DeCarlo, 2008a). These results point to the potential advantage of using a model-based approach over simply averaging the raters' scores; the next step is to conduct validity studies with real-world data.

Second, Table 2 shows that a lack of balance has only a small effect on classification accuracy, which was reduced by less than 2%. An interesting result shown in Table 2 is that the degree of linkage appears to have virtually no effect on classification accuracy. Thus the degree of linkage appears to have little effect on either parameter estimation (in terms of bias for *d*) or on classification accuracy. This might occur because, as noted above, designs with less than full linkage are less sparse, in the sense that there are fewer possible response patterns; the fact that the number of essays for each rater pair is also larger in the spiral design than in a BIB design might also affect estimation and classification.

In sum, the current results suggest that a design with the near-minimum number of rater pairs can be used with little loss in estimation or classification; the main loss appears to be an increase in the bias of the standard errors. The finding of little or no detrimental effect on classification of using a design with near-minimum linkage has important implications, in that it suggests one can use a simpler design, such as that used above, in lieu of a BIB design. A practical consequence is that, because there are considerably fewer rater pairs in the spiral design as compared to the BIB, allocation of the essays to raters is much easier to manage. For example, for a situation with 32 raters, as in the real-world data analysis presented below, there are 496 rater pairs in a BIB design, but only 32 rater pairs in the spiral-like design used here. An advantage of having fewer rater pairs to manage is that it is easier to keep the design from getting highly unbalanced, whereas this is more difficult to do when there are hundreds of rater pairs to manage. As shown here, a lack of balance appears to have a more deleterious effect on estimation than using fewer rater pairs. Thus, whereas some large-scale assessments attempt to use nearly all (or all) of the possible rater pairs, the results presented here suggest that this may not be necessary.

The next section examines another practical problem that arises in large-scale assessments, which is that some assessments include more than one CR item. This raises questions about how to analyze multiple CR items within the framework of latent class signal detection theory.

## A Hierarchical Rater Model

Many tests, such as the SAT, include only one CR item, such as an essay. However, some tests include more than one CR item; for example, the GRE has two CR items (essays on issue and argument tasks), whereas the Praxis™ Middle School Mathematics test (MSMAT) and Reading Across the Curriculum: Elementary (RACE) test (which is also a part of the Praxis series) each include three CR items, which are problem-solving exercises. Although one can analyze each essay separately, a more comprehensive model considers the multiple essays simultaneously. In particular, as discussed above, it is important to recognize the nested structure of the data, in that raters are nested within the essays and, when there is more than one essay, essays are nested within examinees. This structure is explicitly recognized by a hierarchical rater model (HRM), which was introduced by Patz (1996) and is discussed in Patz et al. (2002). Here it is shown that the latent class SDT model can easily be used for the first level of the HRM (DeCarlo, 2008b), which offers some advantages over the SDT-like model used by Patz et al. The model also can easily be fit using MLE or PME, whereas Patz et al. used a (more computationally intense) Markov Chain Monte Carlo (MCMC) approach.

### An HRM-SDT Model

Figure 2 shows a representation of the HRM-SDT model with two essays per examinee. The first level of the model, which is simply a latent class SDT model, relates the observed scores of the raters ($Y_j$), which are ordinal, to the latent class variables ($X^{\#}{}_{l}$), which also are ordinal. As discussed above, the observed responses $Y_j$ arise from the rater's perception Ψ and their use of response criteria $c_{jk}$, in that the distance of $c_{jk}$ from the conditional mean of Ψ determines the response probability; curved arrows are used in the figure to indicate the nonlinear nature of this relation. The arrows from $X^{\#}$ to Ψ indicate that the mean of Ψ is shifted by $d_j$ as the latent category increases by one. In the second level of the model, the ordinal latent variables $X^{\#}$ serve as indicators of the examinee's ability (θ) via an item response theory (IRT) model (the generalized partial credit model is used here). The arrows from θ to Ψ are curved to indicate that the latent class variables $X^{\#}$ have a nonlinear relation to an examinees' ability, via an IRT model, with parameters *a* and *b*, for the slope (discrimination) and category steps, respectively.

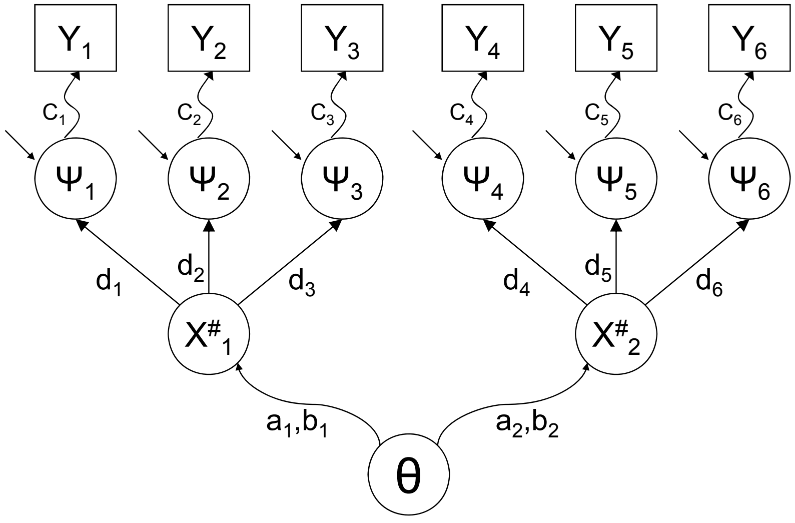

***Figure 2.* A representation of the HRM-SDT model.**

With matrices $\mathbf{Y}$ for the response patterns and $\mathbf{X}^{\#}$ for the latent class variables, and writing $p(\mathbf{y})$ for $p(\mathbf{Y} = \mathbf{y})$ and $p(\mathbf{x}^{\#})$ for $p(\mathbf{X}^{\#} = \mathbf{x}^{\#})$, the model can be written as

$$p(\mathbf{y}) = \Sigma_{x\#} \int_{\theta} p(\mathbf{x}^{\#}|\theta)\, p(\mathbf{y}|\mathbf{x}^{\#},\theta)\, p(\theta)\, d\theta, \tag{6}$$

where $p(\mathbf{y}|\mathbf{x}^{\#},\theta)$ is the rater component of the model (the first level) and $p(\mathbf{x}^{\#}|\theta)$ is the model for the CR items (the second level). As before, an assumption of independence given the latent class variables $\mathbf{X}^{\#}$ is made,

$$p(\mathbf{y}|\mathbf{x}^{\#},\theta) = \Pi_{jl}\, p(y_{jl}\,|\mathbf{x}^{\#},\theta), \tag{7}$$

where *j* indicates the rater and *l* indicates the essay. An assumption of independence of the *l* latent classes given θ is also made,

$$p(\mathbf{x}^{\#}|\theta) = \Pi_l\, p(x_l{}^{\#}\,|\theta). \tag{8}$$

As before, the first level of the model consists of a latent class SDT model,

$$p(Y_{jl} \leq y_j \mid x_l^{\#}) = F(c_{jkl} - d_{jl}\, x_l^{\#}).$$

The above is incorporated into the right side of Equation 7 by differencing the cumulative probabilities, as shown above. The second level of the model treats the latent classes as ordinal indicators of ability using an IRT model. For example, using adjacent categories logits (see Agresti, 2002) gives the generalized partial credit (GPC) model (Muraki, 1992),

$$\log\, [p(X_l^{\#} = x_l^{\#}\, |\theta)/p(X_l^{\#} = x_l^{\#}{+}1|\, \theta)] = b_{lx\#} - a_l\, \theta, \qquad (9)$$

where $X_l^{\#}$ are ordinal latent categories for the $l^{th}$ CR item, $x_l^{\#}$ are discrete values that range from zero to one minus the number of latent classes, θ is a continuous latent variable (i.e., ability), and $b_{lx\#}$ and $a_l$ are item step and discrimination parameters, respectively, for item $l$; note that the model is parameterized in a manner similar to that for the SDT model, in that lower categories are modeled (whereas in the usual version of the GPC model, higher categories are modeled). Starting with Equation 9, one can also write the model in terms of probabilities, as done by Muraki (1992) and others.

There are many possible versions of the HRM-SDT model that can be examined in simulations. Examined here is a basic version of the model with two or three constructed response items and three indicators per CR item (fully crossed). This version of the model arises in practice when the data are pooled across raters (as is commonly done), thereby giving a fully crossed design (that ignores rater effects). Also note that, although each essay in large-scale assessments is typically scored by two raters, a third rater (an adjudicator) also is used for many tests, giving a total of three raters per essay; DeCarlo and Kim (2008) showed that including the third score is feasible and provides useful information. The real-world data examined below are also of this form. The simulations provide information about parameter recovery for the different components of the HRM-SDT model as well as information about the effect of using more CR items (three instead of two).

## Methods

The hierarchical rater model previously has been fit using a fully Bayesian approach and the MCMC algorithm (e.g., Patz et al., 2002). The model as presented here, however, can easily be fit using maximum likelihood estimation (or posterior mode estimation), employing any one of several widely available software packages. For example, the models were fit here using the syntax version of Latent Gold 4.5 (Vermunt & Magidson, 2007). The model can also be fit with the freely available software Lem (Vermunt, 1997), though only with MLE and not PME.

**Data**

For the simulation, data for the hierarchical rater model, parameterized as shown above (using the GPC for the IRT portion), were generated using SAS macros written by the author. Each dataset consisted of 3,000 observations; a total of 100 datasets were generated for each of two conditions: an HRM-SDT with two CR items and an HRM-SDT with three CR items. The population values of the parameters are shown in Appendix D; the values used were chosen because they are typical of those found in other studies. SAS macros were used to strip out and summarize the results from the Latent Gold output; further details about the simulations can be found in a prior research report (DeCarlo, 2008a). As before, the SAS macro that computed the final results checked and corrected for label switching. Note that the problem of label switching is more complex for the HRM-SDT model because it can occur both for the latent categorical variable $X^{\#}$ and for the latent continuous variable θ, and so there are four possibilities that must be considered. In particular, when label switching occurs for $X^{\#}$, the sign of *d* is reversed, *c* has to be corrected by adding $K - 1$ times *d* to the obtained estimates of *c*, and the order of the latent class sizes is reversed; this is the same as above for the simple latent class SDT model. In addition, for HRM-SDT, label switching can occur for θ and not for $X^{\#}$, in which case the sign of *a* will be reversed, but the order for *b* will be correct. If label switching occurs for both θ and for $X^{\#}$, then the sign of *a* is correct, but the order of *b* is reversed. The SAS macro that combined the results for the 100 replications checked for these different possibilities and made the appropriate corrections to the parameter estimates.

Some preliminary runs showed that problems with boundary problems occasionally occurred, and so PME was used, with Bayes constants of unity for the rater responses and the latent classes (as also used in the studies above). Note that the computer time for the situation with three CR items was found to be much longer than that for two CR items. The time to complete one of the (100) replications is determined by the time it takes to complete each iteration (which is slower for Newton-Raphson steps than EM steps), the number of iterations needed for convergence, the number of sets of starting values that are used (10 were used here), and the number of nodes used for the continuous variable theta (10 nodes were used here). For the situation with two CR items, each replication took about 10 minutes to complete, however, for the situation with three CR items, each replication took about 2 hours to complete (and so the simulation took several weeks of continuous computer time).

### Results: Two CR Items

The first section of Table D1 shows the parameter estimates for the signal detection part of the model, along with the bias, percent bias, and the mean squared error. With respect to the signal detection parameters, recovery of the discrimination parameter *d* is excellent, with a bias of 1.2% or less; the MSE is also small. The bias is also small for the response criteria, with a percent bias of under 5%. Overall, recovery of the signal detection parameters is excellent.

The second section of Table D1 shows results for the IRT part of the model, that is, a GPC model that relates the six ordinal latent classes for each CR item to theta. The table shows that, for both CR items, the bias for estimates of the discrimination parameters $a_1$ and $a_2$ is fairly large, with a percent bias of 22.7% and 15.7%, respectively. The bias for the category step estimates $b_{lm}$ is also generally large (greater than 10%). Computing the Monte Carlo standard error for the *a* parameters (as the standard deviation across replications divided by the square root of the number of replications) gives 95% CIs (confidence intervals) of (1.18, 1.27) and (1.24, 1.28), neither of which contain the population values of 1.0 and 1.5, respectively. This shows that the 100 replications were enough to detect significant bias. In contrast, a 95% CI for the estimate of the first *d* gives (1.99, 2.01), which contains the population value of 2.0 and simply reflects that the bias is very small; the results are similar for the other *d* parameters. Thus there is negligible bias for estimates of the rater parameters but significant bias for the CR item parameter estimates.

It is interesting to note that the population value of *a* for the first essay (1.0) is overestimated (as 1.2), whereas the population value of 1.5 for the second essay is underestimated (as 1.3). This suggests that the estimates of *a* are shrunk towards a mean value, which would happen if the *a* parameters were random, as in multilevel models with random slope parameters (Raudenbush & Bryk, 2002). However, the *a* parameters are fixed, not random, in the HRM-SDT model, and so the reason for the shrinkage (if it is in fact occurring) is not known. It is not due to the use of posterior mode estimation, because that smooths the *a* parameters towards zero, and not towards the mean (which can be verified easily with Latent Gold by using large values for the Bayes constants). An examination of the correlation matrix of the parameter estimates showed that the correlation of the estimates of the *a* parameters for the first and second item was large and negative (around −0.95), which would account for the over- and under-estimation noted above for *a*; note that the correlation was considerably smaller (−0.50 or less) for the situation with three CR items (examined next).

Overall, the simulation suggests that parameters for the SDT part of the HRM model are well recovered, particularly the discrimination parameter *d*; recovery of the parameters for the IRT part of the model appears to be poorer. This likely occurs because there are only two indicators for the IRT model (i.e., two essays); this is examined in the next section, where an additional CR item is added. Also note that a prior simulation of the HRM-SDT with only two raters for each CR item (and two CR items; DeCarlo, 2008b) found large bias for both the SDT and IRT parts of the model, which suggests that the use of only two indicators, at either level, results in poor recovery.

### Results: Three CR Items

Table D2 shows results for fits of the HRM-SDT model for the simulation with three CR items, again with three raters per item. The top section of the table shows that the signal detection parameters are again well estimated, with a bias of generally less than 5%. The next section of the table shows that, in contrast to the results found for two CR items, the parameters of the second-level IRT model are well estimated, with bias under 5%. Thus Table D2 shows that the addition of an additional indicator at level two of the model, namely a third CR item, markedly improves estimation of the level two parameters, that is, the parameters of the IRT model for the CR items.

Overall, the results suggest that, with at least two CR items, the rater parameters are well recovered when there are three scores per essay, and so one can usefully evaluate rater performance and classification accuracy. On the other hand, the CR item parameters are well recovered primarily when there are at least three CR items per examinee. This means that if one wishes to compare different CR items in terms of item characteristics, then one should try to have at least three CR items. Another possibility (that potentially eliminates the need for more

CR items) is to include multiple-choice items in Level 2 as indicators of theta, which is being examined in current research (DeCarlo & Kim, 2009; Kim, 2009).

## ETS Data

This section applies the HRM-SDT model to the writing section of two ETS datasets. The first example is the writing portion of a large-scale language test that includes two writing tasks for each examinee, whereas the second example involves a  test with three problem-solving exercises.

### Language Test: Data

The data come from 42,608 examinees. Each essay was scored by two raters, with some essays scored by an additional rater, an adjudicator, when the first two scores differed by more than one. For the first writing task, 3.9% of the essays had a third (adjudicated) score, whereas 2.6% of the essays for the second task had a third score. Data for the third scores can be viewed as being missing at random (Rubin, 1976), in that the probability that a value is missing is determined by an observed variable – the difference between the two observed scores (i.e., the value is missing if it is less than 2). The analyses presented here include the third scores; DeCarlo and Kim (2008) showed that estimation is good for adjudicated scores, as along as a sufficient number are available, as is the case here. Thus there are three scores per essay (one with a large percentage of missing values), giving a total of six scores for the two essays.

The scoring rubric consisted of categories from 0 to 5. Note that a score of 1 to 5 indicates a judgment of the quality of an essay, whereas a score of 0 indicates that there was no essay to be judged (i.e., a blank) or that it was written in the wrong language, and so on, and so one can argue for not including essays with scores of zero (e.g., a score of zero for a blank is not a judgment of an essay's quality), which was done here; note that, for the sample used here, only 124 essays out of 42,732 essays received scores of zero.

An HRM-SDT model, as described above, was fit. The SDT component of the model, which is the first level, used the 1-to-5 essay scores from the three raters as ordinal indicators of five latent classes of essay quality for each essay. The IRT component (a GPC model), which is the second level of the model, used the 1-to-5 latent classes for the two essays as two ordinal indicators of examinee ability. In the analysis presented here, the data are treated as coming from a fully crossed design (i.e., the data are pooled across raters), in order to obtain information about the HRM-SDT model as applied to pooled data. The model was fit using Latent Gold 4.5; as in the simulations above, posterior mode estimation with Bayes constants of unity for the response variable and latent categories was used.

**Results**

With respect to the rater parameters (in this case for the pooled raters), Table 3 shows that discrimination is in the range of 3.0 to 3.5, which is good discrimination (for the logistic model; for example, for *d* = 3.5, the odds ratio is 33 to 1; also see the previous research report). For the first writing task the estimate of *d* for the adjudicated score is slightly smaller than that obtained for the other two scores, whereas for the second writing task the estimates of *d* are all about the same. The criteria estimates are also similar across the three scores; however, those for the third score are to the left of (i.e., smaller than) those for the other two scores for the first writing task; this indicates that, for the second writing task, the adjudicated scores tended to be slightly more liberal. It is also interesting to note that in all cases the response criteria estimates in Table 3 are close to the intersection point locations. For example, for the first score an estimate of *d* of 3.4 means that, if the criteria were located at the intersection points, then they would be at 1.7, 5.1, 8.5, and 11.8, and the estimates shown in Table 3 (1.7, 5.4, 9.2, and 12.7) are quite close to these values. Thus it appears that, at least for pooled data, the response criteria tend to be located close to the intersection points of the underlying logistic distributions, at least for the language test examined here (cf. DeCarlo, 2008a).

With respect to the CR item parameters, discrimination is high (4.19 for the first writing task and 2.94 for the second). The item step parameters are fairly similar across the two writing tasks. The last section of the table shows estimates of the latent class sizes. The largest latent class size is for Category 4, followed by Category 3. It is interesting to note that for both writing tasks the latent class sizes are slightly negatively skewed, which is in contrast to the approximately normal distribution of latent class sizes found for other tests (see DeCarlo, 2008a); this might be a characteristic of language tests, but more research on this is needed.

**Mathematics Test: Data**

The second example is for a large-scale mathematics test used in certification. The test includes three CR items, which are problem-solving exercises, each of which is scored by two raters. The raters score the exercises on a 4-point scale, and so a latent class model with four latent classes was used for the first level of the model (the SDT part); the generalized partial credit model was again used for the second level of the model (the IRT part).

**Table 3**

***Results for HRM-SDT Model, Two Essays, Language Test***

Rater parameters

| | First writing task | | | | | |
|---|---|---|---|---|---|---|
| | First score | | Second score | | Third score | |
| Parameter | Estimate | SE | Estimate | SE | Estimate | SE |
| $d$ | 3.42 | 0.04 | 3.46 | 0.04 | 2.94 | 0.16 |
| $c_1$ | 1.68 | 0.05 | 1.75 | 0.05 | 1.13 | 0.26 |
| $c_2$ | 5.38 | 0.08 | 5.41 | 0.08 | 4.25 | 0.31 |
| $c_3$ | 9.22 | 0.12 | 9.32 | 0.12 | 7.73 | 0.43 |
| $c_4$ | 12.70 | 0.14 | 12.83 | 0.14 | 10.80 | 0.52 |
| | Second writing task | | | | | |
| $d$ | 3.28 | 0.04 | 3.33 | 0.04 | 3.32 | 0.23 |
| $c_1$ | 0.27 | 0.08 | 0.29 | 0.08 | 0.18 | 0.54 |
| $c_2$ | 4.54 | 0.10 | 4.59 | 0.10 | 4.28 | 0.45 |
| $c_3$ | 8.94 | 0.12 | 9.06 | 0.13 | 8.85 | 0.66 |
| $c_4$ | 12.47 | 0.15 | 12.64 | 0.15 | 12.45 | 0.78 |

CR item parameters (generalized partial credit model)

| | First writing task | | Second writing task | |
|---|---|---|---|---|
| Parameter | Estimate | SE | Estimate | SE |
| $a$ | 4.19 | 0.43 | 2.95 | 0.16 |
| $b_1$ | −5.17 | 0.54 | −6.24 | 0.31 |
| $b_2$ | −2.84 | 0.26 | −3.66 | 0.18 |
| $b_3$ | 0.37 | 0.06 | −0.40 | 0.05 |
| $b_4$ | 4.21 | 0.41 | 3.03 | 0.16 |

**Latent class sizes**

| | First writing task | | Second writing task | |
|---|---|---|---|---|
| Parameter | Estimate | SE | Estimate | SE |
| $p_1$ | .12 | < .01 | .03 | < .01 |
| $p_2$ | .15 | < .01 | .11 | < .01 |
| $p_3$ | .27 | < .01 | .32 | < .01 |
| $p_4$ | .29 | < .01 | .37 | < .01 |
| $p_5$ | .17 | < .01 | .18 | < .01 |

**Results**

Table 4 shows results for a fit of the HRM-SDT model. A notable difference, compared to the results found above, is that the estimates of the discrimination parameters *d* are quite large, in the range of 8-13 across the three problems. This indicates excellent discrimination; it suggests that the mathematics problems can be classified into the scoring categories more accurately than the writing samples examined above (and for other writing tests, where rater parameter estimates in the range of 2 to 5 have been found). This has not been noted before; comparisons of discrimination across other tests might be informative in future research.

The middle section of Table 4 shows that the CR item discrimination parameter estimates (*a*) are all around unity. The bottom section of Table 4 shows estimates of the latent class sizes. In this case the class sizes are skewed to the left. For example, for the first problem, Class 3 is the largest (.44), followed by Class 4 (.37). For the second and third problems Classes 3 and 2, respectively, have the largest class sizes. Thus, most examinees receive scores of 3 or 4 for the first task, and scores of 2 or 3 for the second and third tasks. The skew might arise because the exam was a test of minimal competency; this merits closer attention in future research. It is interesting to note that the results in Table 4 suggest that the first problem was more often "passed" than the second or third problems, which could reflect a systematic difference in the type of problem-solving exercises that were used for the second and third problems, or an order effect; again, this merits closer attention in future research.

It is also interesting to note that the standard errors of the rater parameters are larger in Table 4 than in Table 3, which likely reflects that there were only two raters per essay in Table 4, whereas there were three raters per essay in Table 3. Similarly, the standard errors of the CR item parameters are larger in Table 3 than in Table 4, which reflects that there were only two CR items in Table 3, but three CR items in Table 4. Thus if one's interest is in obtaining accurate estimates of either the rater parameters or the CR item parameters, then the results suggest that a minimum of three raters or three items should be used (another option is to include multiple choice items in Level 2; DeCarlo & Kim, 2009; Kim, 2009).

**Table 4**

***Results for HRM-SDT, Three Problem-Solving Exercises, Mathematics Test***

Rater parameters

| | First problem | | | |
|---|---|---|---|---|
| | First score | | Second score | |
| Parameter | Estimate | SE | Estimate | SE |
| $d$ | 8.42 | 0.71 | 11.27 | 1.55 |
| $c_1$ | 3.28 | 0.48 | 4.88 | 1.06 |
| $c_2$ | 15.05 | 1.42 | 20.90 | 3.10 |
| $c_3$ | 24.26 | 2.13 | 32.92 | 4.65 |
| | Second problem | | | |
| $d$ | 12.46 | 1.74 | 10.74 | 1.06 |
| $c_1$ | 5.11 | 1.20 | 3.93 | 0.69 |
| $c_2$ | 13.77 | 1.74 | 12.21 | 1.06 |
| $c_3$ | 26.54 | 3.49 | 23.19 | 2.13 |
| | Third problem | | | |
| $d$ | 10.28 | 0.72 | 12.64 | 1.50 |
| $c_1$ | 4.53 | 0.53 | 4.88 | 0.82 |
| $c_2$ | 11.64 | 0.72 | 14.12 | 1.51 |
| $c_3$ | 22.07 | 1.44 | 27.08 | 3.01 |

CR item parameters

| | First problem | | Second problem | | Third problem | |
|---|---|---|---|---|---|---|
| Parameter | Estimate | SE | Estimate | SE | Estimate | SE |
| $a$ | 1.16 | 0.09 | 0.99 | 0.07 | 1.12 | 0.08 |
| $b_1$ | −3.01 | 0.18 | −3.33 | 0.13 | −2.51 | 0.11 |
| $b_2$ | −1.63 | 0.08 | −0.47 | 0.05 | −0.15 | 0.04 |
| $b_3$ | 0.47 | 0.06 | 1.63 | 0.08 | 2.22 | 0.11 |

Latent class sizes

| | First problem | | Second problem | | Third problem | |
|---|---|---|---|---|---|---|
| Parameter | Estimate | SE | Estimate | SE | Estimate | SE |
| $p_1$ | .03 | < .01 | .03 | < .01 | .08 | < .01 |
| $p_2$ | .15 | < .01 | .35 | < .01 | .44 | < .01 |
| $p_3$ | .44 | < .01 | .47 | < .01 | .38 | < .01 |
| $p_4$ | .37 | < .01 | .16 | < .01 | .10 | < .01 |

In sum, the HRM-SDT model provides information about the performance of the raters, about characteristics of the tasks, and about the constructs. For example, the scoring of the mathematics test appears to differ from other tests, such as the language test analyzed above, in that rater discrimination is quite high, which to my knowledge has not been noted before. The high discrimination likely reflects the different nature of the tasks used in the mathematics test, in that the CR items were problem-solving tasks, rather than essays from the writing section of a test such as the GRE or SAT (or the language test examined here). Note that the finding of high values of *d* for the mathematics test also has implications with respect to classification accuracy, in that it should be higher for higher values of *d*. For example, the estimates of the proportion correctly classified, PC, for the mathematics test are .96, .97, and .97, for the first through third problems, respectively (with estimates of λ of .93, .95, and .95; see DeCarlo, 2008a, or Dayton, 1998, for a discussion of lambda). In contrast, for the language test PC is .83 and .83 (with λ of .76 and .73) for the two essays, which is still good but is lower than that found for the mathematics test (in part because of the lower rater discrimination). Thus, given the high values of discrimination, classification accuracy for the mathematics test is quite high, around 97%; note that one could use this result to argue for simply using the average rater scores for the mathematics test, because classification accuracy will likely still be quite high, given the high discrimination.

## Conclusions

The latent class SDT model offers a useful approach to the analysis of constructed response data. The model can easily be incorporated into a more elaborate framework, such as in models with higher-order structures, like the HRM-SDT model. The approach provides information not only about the performance of the raters and the accuracy of the classifications, but also about characteristics of the CR items and aspects of the underlying constructs. Both the basic SDT and higher-order model, and variations, can also easily be fit using widely available software for latent class analysis or structural equation modeling, which should help to encourage applied researchers to use the models.

## References

Agresti, A. (2002). *Categorical data analysis* (2nd ed.). Hoboken, NJ: Wiley.

Clogg, C. C. (1995). Latent class models. In G. Arminger, C. C. Clogg, & M. E. Sobel (Eds.), *Handbook of statistical modeling for the social and behavioral sciences* (pp. 311-359). New York: Plenum Press.

Clogg, C. C., & Eliason, S. R. (1987). Some common problems in log-linear analysis. *Sociological Methods and Research*, *16*, 8-44.

Dayton, C. M. (1998). *Latent class scaling analysis*. Thousand Oaks, CA: Sage Publications.

DeCarlo, L. T. (1998). Signal detection theory and generalized linear models. *Psychological Methods*, *3*, 186-205.

DeCarlo, L. T. (2002). A latent class extension of signal detection theory, with applications. *Multivariate Behavioral Research*, *37*, 423-451.

DeCarlo, L. T. (2005). A model of rater behavior in essay grading based on signal detection theory. *Journal of Educational Measurement, 42*, 53-76.

DeCarlo, L. T. (2008a). *Studies of a latent-class signal-detection model for constructed response scoring* (ETS Research Rep. No. RR-08-63). Princeton, NJ: ETS.

DeCarlo, L. T. (2008b, March). *On a hierarchical rater model for essay grading: Incorporating a latent class signal detection model*. Paper presented at the annual meeting of the National Council on Measurement in Education, New York.

DeCarlo, L. T., & Kim, Y. K. (2008, March). *Score resolution in essay grading: a view from a signal detection model of rater behavior*. Paper presented at the annual meeting of the American Educational Research Association, New York.

DeCarlo, L. T., & Kim, Y. K. (2009, April). *On scoring constructed response items and multiple choice items: Incorporating signal detection and item response models into a hierarchical rater model*. Paper presented at the annual meeting of the National Council on Measurement in Education, San Diego, CA.

Eccleston, J. A., & Hedayat, A. (1974). On the theory of connected designs: Characterization and optimality. *The Annals of Statistics*, *2*, 1238-1255.

Fleiss, J. L. (1986). *The design and analysis of clinical experiments*. New York: Wiley.

Galindo-Garre, F., & Vermunt, J. K. (2006). Avoiding boundary estimates in latent class analysis by Bayesian posterior mode estimation. *Behaviormetrika*, *33*, 43-59.

Galindo-Garre, F., Vermunt, J. K., & Bergmsa, W. P. (2004). Bayesian posterior estimation of logit parameters with small samples. *Sociological Methods & Research*, *33*, 88-117.

Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (1995). *Bayesian data analysis*. New York: Chapman & Hall.

Hombo, C. M., Donoghue, J. R., & Thayer, D. T. (2001). *A simulation study of the effect of rater designs on ability estimation* (ETS Research Rep. No. RR-01-05). Princeton, NJ: ETS.

Kim, Y. K. (2009). *Combining constructed response items and multiple choice items using a hierarchical rater model*. Unpublished doctoral dissertation, Teachers College, Columbia University, New York.

Maris, E. (1999). Estimating multiple classification latent class models. *Psychometrika*, *64*, 187-212.

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

Muthén, L. K., & Muthén, B. O. (1998-2007). *Mplus user's guide* (5th ed.). Los Angeles, CA: Muthén & Muthén.

Patz, R. J. (1996). *Markov chain Monte Carlo methods for item response theory models with applications for the National Assessment of Educational Progress*. Unpublished doctoral dissertation, Carnegie Mellon University

Patz, R. J., Junker, B. W., Johnson, M. S., & Mariano, L. T. (2002). The hierarchical rater model for rated test items and its application to large-scale educational assessment data. *Journal of Educational and Behavioral Statistics*, *27*, 341-384.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (2nd ed.). Thousand Oaks, CA: Sage Publications.

Rubin, D. B. (1976). Inference and missing data. *Biometrika*, *63*, 581-592.

Schafer, J. L. (1997). *Analysis of incomplete multivariate data*. New York: Chapman & Hall.

Vermunt, J. K. (1997). LEM: A general program for the analysis of categorical data [Computer software and manual]. Retrieved December 11, 2009, from http://www.uvt.nl/faculteiten/fsw/organisatie/ departementen/mto/software2.html.

Vermunt, J. K., & Magidson, J. (2005). *Technical guide for Latent Gold 4.0: Basic and Advanced*. Belmont, MA: Statistical Innovations Inc.

Vermunt, J. K., & Magidson, J. (2007, February). *LG-Syntax$^{TM}$ user's guide: Manual for Latent Gold 4.5 Syntax Module*. Belmont, MA: Statistical Innovations Inc.

Weeks, D. L., & Williams, D. R. (1964). A note on the determination of connectedness in an N-way cross classification. *Technometrics*, *6*, 319-324.

# Appendix A

## Balanced and Unbalanced Designs Used in the Simulations

**Table A1**

***Balanced Design, 45 Rater Pairs, 24 per Pair***

| | | | | | Rater | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | Total |
| | 24 | 24 | | | | | | | | | |
| | 24 | | 24 | | | | | | | | |
| | 24 | | | 24 | | | | | | | |
| | 24 | | | | 24 | | | | | | |
| | 24 | | | | | 24 | | | | | |
| | 24 | | | | | | 24 | | | | |
| | 24 | | | | | | | 24 | | | |
| | 24 | | | | | | | | 24 | | |
| | 24 | | | | | | | | | 24 | |
| | | 24 | 24 | | | | | | | | |
| | | 24 | | 24 | | | | | | | |
| | | 24 | | | 24 | | | | | | |
| | | 24 | | | | 24 | | | | | |
| | | 24 | | | | | 24 | | | | |
| | | 24 | | | | | | 24 | | | |
| | | 24 | | | | | | | 24 | | |
| | | 24 | | | | | | | | 24 | |
| | | | 24 | 24 | | | | | | | |
| | | | 24 | | 24 | | | | | | |
| | | | 24 | | | 24 | | | | | |
| | | | 24 | | | | 24 | | | | |
| | | | 24 | | | | | 24 | | | |
| | | | 24 | | | | | | 24 | | |
| | | | 24 | | | | | | | 24 | |
| | | | | 24 | 24 | | | | | | |
| | | | | 24 | | 24 | | | | | |
| | | | | 24 | | | 24 | | | | |
| | | | | 24 | | | | 24 | | | |
| | | | | 24 | | | | | 24 | | |
| | | | | 24 | | | | | | 24 | |
| | | | | | 24 | 24 | | | | | |
| | | | | | 24 | | 24 | | | | |
| | | | | | 24 | | | 24 | | | |
| | | | | | 24 | | | | 24 | | |
| | | | | | 24 | | | | | 24 | |
| | | | | | | 24 | 24 | | | | |
| | | | | | | 24 | | 24 | | | |
| | | | | | | 24 | | | 24 | | |
| | | | | | | 24 | | | | 24 | |
| | | | | | | | 24 | 24 | | | |
| | | | | | | | 24 | | 24 | | |
| | | | | | | | 24 | | | 24 | |
| | | | | | | | | 24 | 24 | | |
| | | | | | | | | 24 | | 24 | |
| | | | | | | | | | 24 | 24 | |
| | | | | | | | | | | | 1080 |
| Total/Rater | 216 | 216 | 216 | 216 | 216 | 216 | 216 | 216 | 216 | 216 | |

**Table A2**

***Unbalanced Design, 45 Rater Pairs***

| | | | | | Rater | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | Total |
| | 5 | 5 | | | | | | | | | |
| | 15 | | 15 | | | | | | | | |
| | 20 | | | 20 | | | | | | | |
| | 5 | | | | 5 | | | | | | |
| | 20 | | | | | 20 | | | | | |
| | 5 | | | | | | 5 | | | | |
| | 100 | | | | | | | 100 | | | |
| | 15 | | | | | | | | 15 | | |
| | 185 | | | | | | | | | 185 | |
| | | 5 | 5 | | | | | | | | |
| | | 5 | | 5 | | | | | | | |
| | | 5 | | | 5 | | | | | | |
| | | 5 | | | | 5 | | | | | |
| | | 5 | | | | | 5 | | | | |
| | | 10 | | | | | | 10 | | | |
| | | 5 | | | | | | | 5 | | |
| | | 5 | | | | | | | | 5 | |
| | | | 20 | 20 | | | | | | | |
| | | | 5 | | 5 | | | | | | |
| | | | 20 | | | 20 | | | | | |
| | | | 25 | | | | 25 | | | | |
| | | | 40 | | | | | 40 | | | |
| | | | 40 | | | | | | 40 | | |
| | | | 30 | | | | | | | 30 | |
| | | | | 5 | 5 | | | | | | |
| | | | | 10 | | 10 | | | | | |
| | | | | 30 | | | 30 | | | | |
| | | | | 15 | | | | 15 | | | |
| | | | | 20 | | | | | 20 | | |
| | | | | 15 | | | | | | 15 | |
| | | | | | 5 | 5 | | | | | |
| | | | | | 5 | | 5 | | | | |
| | | | | | 20 | | | 20 | | | |
| | | | | | 5 | | | | 5 | | |
| | | | | | 5 | | | | | 5 | |
| | | | | | | 10 | 10 | | | | |
| | | | | | | 30 | | 30 | | | |
| | | | | | | 10 | | | 10 | | |
| | | | | | | 10 | | | | 10 | |
| | | | | | | | 105 | 105 | | | |
| | | | | | | | 90 | | 90 | | |
| | | | | | | | 5 | | | 5 | |
| | | | | | | | | 35 | 35 | | |
| | | | | | | | | 45 | | 45 | |
| | | | | | | | | | 10 | 10 | |
| | | | | | | | | | | | 1080 |
| Total/Rater | 370 | 50 | 200 | 140 | 60 | 120 | 280 | 400 | 230 | 310 | |

**Table A3**

***Balanced Design, 10 Rater Pairs, 108 per Pair***

| | Rater | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | 2 | 5 | 6 | 4 | 3 | 9 | 10 | 1 | 8 | 7 | Total |
| | 108 | 108 | | | | | | | | | |
| | | 108 | 108 | | | | | | | | |
| | | | 108 | 108 | | | | | | | |
| | | | | 108 | 108 | | | | | | |
| | | | | | 108 | 108 | | | | | |
| | | | | | | 108 | 108 | | | | |
| | | | | | | | 108 | 108 | | | |
| | | | | | | | | 108 | 108 | | |
| | | | | | | | | | 108 | 108 | |
| | 108 | | | | | | | | | 108 | |
| | | | | | | | | | | | 1080 |
| Total/Rater | 216 | 216 | 216 | 216 | 216 | 216 | 216 | 216 | 216 | 216 | |

**Table A4**

***Unbalanced Design, 10 Rater Pairs***

| | Rater | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | 2 | 5 | 6 | 4 | 3 | 9 | 10 | 1 | 8 | 7 | Total |
| | 20 | 20 | | | | | | | | | |
| | | 40 | 40 | | | | | | | | |
| | | | 80 | 80 | | | | | | | |
| | | | | 60 | 60 | | | | | | |
| | | | | | 140 | 140 | | | | | |
| | | | | | | 90 | 90 | | | | |
| | | | | | | | 220 | 220 | | | |
| | | | | | | | | 150 | 150 | | |
| | | | | | | | | | 250 | 250 | |
| | 30 | | | | | | | | | 30 | |
| | | | | | | | | | | | 1080 |
| Total/Rater | 50 | 60 | 120 | 140 | 200 | 230 | 310 | 370 | 400 | 280 | |

**Table A5**

***Balanced Design, 20 Rater Pairs, 54 per Pair***

| | | | | | Rater | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | Total |
| | 54 | 54 | | | | | | | | | |
| | | 54 | 54 | | | | | | | | |
| | | | 54 | 54 | | | | | | | |
| | | | | 54 | 54 | | | | | | |
| | | | | | 54 | 54 | | | | | |
| | | | | | | 54 | 54 | | | | |
| | | | | | | | 54 | 54 | | | |
| | | | | | | | | 54 | 54 | | |
| | | | | | | | | | 54 | 54 | |
| | 54 | | | | | | | | | 54 | |
| | 54 | | 54 | | | | | | | | |
| | | 54 | | 54 | | | | | | | |
| | | | 54 | | 54 | | | | | | |
| | | | | 54 | | 54 | | | | | |
| | | | | | 54 | | 54 | | | | |
| | | | | | | 54 | | 54 | | | |
| | | | | | | | 54 | | 54 | | |
| | | | | | | | | 54 | | 54 | |
| | 54 | | | | | | | | 54 | | |
| | | 54 | | | | | | | | 54 | |
| | | | | | | | | | | | 1080 |
| Total/Rater | 216 | 216 | 216 | 216 | 216 | 216 | 216 | 216 | 216 | 216 | |

**Table A6**

***Unbalanced Design, 20 Rater Pairs***

| | | | | | Rater | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | Total |
| | 35 | 35 | | | | | | | | | |
| | | 5 | 5 | | | | | | | | |
| | | | 50 | 50 | | | | | | | |
| | | | | 15 | 15 | | | | | | |
| | | | | | 5 | 5 | | | | | |
| | | | | | | 40 | 40 | | | | |
| | | | | | | | 185 | 185 | | | |
| | | | | | | | | 35 | 35 | | |
| | | | | | | | | | 35 | 35 | |
| | 95 | | | | | | | | | 95 | |
| | 120 | | 120 | | | | | | | | |
| | | 5 | | 5 | | | | | | | |
| | | | 25 | | 25 | | | | | | |
| | | | | 70 | | 70 | | | | | |
| | | | | | 15 | | 15 | | | | |
| | | | | | | 5 | | 5 | | | |
| | | | | | | | 40 | | 40 | | |
| | | | | | | | | 175 | | 175 | |
| | 120 | | | | | | | | 120 | | |
| | | 5 | | | | | | | | 5 | |
| | | | | | | | | | | | 1080 |
| Total/Rater | 370 | 50 | 200 | 140 | 60 | 120 | 280 | 400 | 230 | 310 | |

## Appendix B

## Parameter Estimates, Bias, Percent Bias, and MSE

**Table B1**

***BIB Design, Normal d, N = 1,080, 45 Rater Pairs, 24 per Pair***

| Size | Parameter | Value | Estimate | Bias | %Bias | MSE |
|---|---|---|---|---|---|---|
| | | | Rater parameters | | | |
| 216 | $d_1$ | 1.0 | 0.954 | −0.0462 | −4.620 | 0.023 |
| 216 | $d_2$ | 2.0 | 1.928 | −0.0719 | −3.595 | 0.061 |
| 216 | $d_3$ | 2.0 | 1.928 | −0.0722 | −3.610 | 0.081 |
| 216 | $d_4$ | 3.0 | 2.971 | −0.0290 | −0.967 | 0.252 |
| 216 | $d_5$ | 3.0 | 2.908 | −0.0916 | −3.053 | 0.222 |
| 216 | $d_6$ | 3.0 | 2.941 | −0.0591 | −1.970 | 0.194 |
| 216 | $d_7$ | 3.0 | 2.969 | −0.0310 | −1.030 | 0.167 |
| 216 | $d_8$ | 4.0 | 3.844 | −0.1560 | −3.900 | 0.370 |
| 216 | $d_9$ | 4.0 | 3.828 | −0.1719 | −4.298 | 0.402 |
| 216 | $d_{10}$ | 5.0 | 4.553 | −0.4475 | −8.950 | 0.575 |
| | $c_{11}$ | 0.5 | 0.362 | −0.1384 | −27.680 | 0.143 |
| | $c_{12}$ | 1.5 | 1.372 | −0.1277 | −8.513 | 0.166 |
| | $c_{13}$ | 2.5 | 2.391 | −0.1094 | −4.376 | 0.193 |
| | $c_{14}$ | 3.5 | 3.398 | −0.1024 | −2.926 | 0.239 |
| | $c_{15}$ | 4.5 | 4.399 | −0.1010 | −2.244 | 0.326 |
| | $c_{21}$ | 1.0 | 0.728 | −0.2723 | −27.230 | 0.217 |
| | $c_{22}$ | 3.0 | 2.788 | −0.2124 | −7.080 | 0.332 |
| | $c_{23}$ | 5.0 | 4.785 | −0.2149 | −4.298 | 0.429 |
| | $c_{24}$ | 7.0 | 6.808 | −0.1917 | −2.739 | 0.671 |
| | $c_{25}$ | 9.0 | 8.918 | −0.0825 | −0.917 | 0.982 |
| | $c_{31}$ | 1.0 | 0.763 | −0.2370 | −23.700 | 0.268 |
| | $c_{32}$ | 3.0 | 2.817 | −0.1827 | −6.090 | 0.373 |
| | $c_{33}$ | 5.0 | 4.822 | −0.1778 | −3.556 | 0.591 |
| | $c_{34}$ | 7.0 | 6.850 | −0.1502 | −2.146 | 0.854 |

| Size | Parameter | Value | Estimate | Bias | %Bias | MSE |
|---|---|---|---|---|---|---|
| | $c_{35}$ | 9.0 | 8.894 | −0.1065 | −1.183 | 1.326 |
| | $c_{41}$ | 1.5 | 1.149 | −0.3506 | −23.373 | 0.562 |
| | $c_{42}$ | 4.5 | 4.284 | −0.2160 | −4.800 | 0.992 |
| | $c_{43}$ | 7.5 | 7.426 | −0.0742 | −0.989 | 1.809 |
| | $c_{44}$ | 10.5 | 10.505 | 0.0046 | 0.044 | 3.185 |
| | $c_{45}$ | 13.5 | 13.777 | 0.2768 | 2.050 | 5.423 |
| | $c_{51}$ | 1.5 | 1.033 | −0.4671 | −31.140 | 0.521 |
| | $c_{52}$ | 4.5 | 4.143 | −0.3569 | −7.931 | 0.846 |
| | $c_{53}$ | 7.5 | 7.255 | −0.2449 | −3.265 | 1.519 |
| | $c_{54}$ | 10.5 | 10.293 | −0.2073 | −1.974 | 2.429 |
| | $c_{55}$ | 13.5 | 13.420 | −0.0797 | −0.590 | 4.213 |
| | $c_{61}$ | 1.5 | 1.103 | −0.3970 | −26.467 | 0.499 |
| | $c_{62}$ | 4.5 | 4.304 | −0.1962 | −4.360 | 0.641 |
| | $c_{63}$ | 7.5 | 7.330 | −0.1703 | −2.271 | 1.238 |
| | $c_{64}$ | 10.5 | 10.417 | −0.0831 | −0.791 | 2.187 |
| | $c_{65}$ | 13.5 | 13.539 | 0.0390 | 0.289 | 3.686 |
| | $c_{71}$ | 1.5 | 1.112 | −0.3878 | −25.853 | 0.450 |
| | $c_{72}$ | 4.5 | 4.308 | −0.1923 | −4.273 | 0.672 |
| | $c_{73}$ | 7.5 | 7.446 | −0.0538 | −0.717 | 1.475 |
| | $c_{74}$ | 10.5 | 10.526 | 0.0261 | 0.249 | 2.128 |
| | $c_{75}$ | 13.5 | 13.695 | 0.1954 | 1.447 | 3.155 |
| | $c_{81}$ | 2.0 | 1.285 | −0.7152 | −35.760 | 0.993 |
| | $c_{82}$ | 6.0 | 5.469 | −0.5309 | −8.848 | 1.629 |
| | $c_{83}$ | 10.0 | 9.658 | −0.3423 | −3.423 | 2.888 |
| | $c_{84}$ | 14.0 | 13.643 | −0.3568 | −2.549 | 4.633 |
| | $c_{85}$ | 18.0 | 17.778 | −0.2225 | −1.236 | 6.724 |
| | $c_{91}$ | 2.0 | 1.307 | −0.6927 | −34.635 | 0.976 |
| | $c_{92}$ | 6.0 | 5.486 | −0.5143 | −8.572 | 1.529 |
| | $c_{93}$ | 10.0 | 9.556 | −0.4442 | −4.442 | 2.959 |

| Size | Parameter | Value | Estimate | Bias | %Bias | MSE |
|---|---|---|---|---|---|---|
| | $c_{94}$ | 14.0 | 13.591 | −0.4089 | −2.921 | 4.748 |
| | $c_{95}$ | 18.0 | 17.748 | −0.2518 | −1.399 | 7.103 |
| | $c_{101}$ | 2.5 | 1.493 | −1.0072 | −40.288 | 1.747 |
| | $c_{102}$ | 7.5 | 6.438 | −1.0620 | −14.160 | 2.583 |
| | $c_{103}$ | 12.5 | 11.493 | −1.0069 | −8.055 | 4.414 |
| | $c_{104}$ | 17.5 | 16.226 | −1.2741 | −7.281 | 6.645 |
| | $c_{105}$ | 22.5 | 21.266 | −1.2341 | −5.485 | 9.175 |
| Latent class sizes | | | | | | |
| | Class 1 | 0.080 | 0.101 | 0.0210 | 26.250 | |
| | Class 2 | 0.170 | 0.160 | −0.0100 | −5.882 | |
| | Class 3 | 0.250 | 0.239 | −0.0110 | −4.400 | |
| | Class 4 | 0.250 | 0.244 | −0.0060 | −2.400 | |
| | Class 5 | 0.170 | 0.157 | −0.0130 | −7.647 | |
| | Class 6 | 0.080 | 0.099 | 0.0190 | 23.750 | |

*Note*. Size is the number of essays scored by each rater.

**Table B2**

***Unbalanced Design, Normal d, N = 1,080, 45 Rater Pairs***

| Size | Parameter | Value | Estimate | Bias | %Bias | MSE |
|---|---|---|---|---|---|---|
| Rater parameters | | | | | | |
| 370 | $d_1$ | 1.0 | 0.970 | −0.0302 | −3.020 | 0.013 |
| 50 | $d_2$ | 2.0 | 1.771 | −0.2292 | −11.460 | 0.194 |
| 200 | $d_3$ | 2.0 | 1.967 | −0.0334 | −1.670 | 0.081 |
| 140 | $d_4$ | 3.0 | 2.903 | −0.0974 | −3.247 | 0.266 |
| 60 | $d_5$ | 3.0 | 2.542 | −0.4576 | −15.253 | 0.449 |
| 120 | $d_6$ | 3.0 | 2.847 | −0.1534 | −5.113 | 0.363 |
| 280 | $d_7$ | 3.0 | 2.914 | −0.0865 | −2.883 | 0.134 |
| 400 | $d_8$ | 4.0 | 4.208 | 0.2084 | 5.210 | 0.371 |
| 230 | $d_9$ | 4.0 | 3.828 | −0.1719 | −4.298 | 0.391 |
| 310 | $d_{10}$ | 5.0 | 4.137 | −0.8635 | −17.270 | 1.103 |
| | $c_{11}$ | 0.5 | 0.385 | −0.1153 | −23.060 | 0.064 |
| | $c_{12}$ | 1.5 | 1.407 | −0.0928 | −6.187 | 0.080 |
| | $c_{13}$ | 2.5 | 2.424 | −0.0760 | −3.040 | 0.091 |
| | $c_{14}$ | 3.5 | 3.448 | −0.0525 | −1.500 | 0.107 |
| | $c_{15}$ | 4.5 | 4.484 | −0.0157 | −0.349 | 0.141 |
| | $c_{21}$ | 1.0 | 0.510 | −0.4904 | −49.040 | 0.792 |
| | $c_{22}$ | 3.0 | 2.564 | −0.4358 | −14.527 | 0.787 |
| | $c_{23}$ | 5.0 | 4.469 | −0.5313 | −10.626 | 1.342 |
| | $c_{24}$ | 7.0 | 6.431 | −0.5690 | −8.129 | 1.827 |
| | $c_{25}$ | 9.0 | 8.467 | −0.5328 | −5.920 | 2.885 |
| | $c_{31}$ | 1.0 | 0.748 | −0.2516 | −25.160 | 0.241 |
| | $c_{32}$ | 3.0 | 2.844 | −0.1557 | −5.190 | 0.352 |
| | $c_{33}$ | 5.0 | 4.933 | −0.0667 | −1.334 | 0.597 |
| | $c_{34}$ | 7.0 | 6.985 | −0.0150 | −0.214 | 0.912 |
| | $c_{35}$ | 9.0 | 9.105 | 0.1045 | 1.161 | 1.410 |
| | $c_{41}$ | 1.5 | 1.084 | −0.4162 | −27.747 | 0.573 |

| | | | | | |
|---|---|---|---|---|---|
| $c_{42}$ | 4.5 | 4.187 | −0.3126 | −6.947 | 0.849 |
| $c_{43}$ | 7.5 | 7.277 | −0.2234 | −2.979 | 1.918 |
| $c_{44}$ | 10.5 | 10.360 | −0.1396 | −1.330 | 3.110 |
| $c_{45}$ | 13.5 | 13.341 | −0.1587 | −1.176 | 4.412 |
| $c_{51}$ | 1.5 | 0.809 | −0.6908 | −46.053 | 1.133 |
| $c_{52}$ | 4.5 | 3.587 | −0.9129 | −20.287 | 1.785 |
| $c_{53}$ | 7.5 | 6.464 | −1.0357 | −13.809 | 2.839 |
| $c_{54}$ | 10.5 | 9.250 | −1.2501 | −11.906 | 4.443 |
| $c_{55}$ | 13.5 | 11.863 | −1.6367 | −12.124 | 7.426 |
| $c_{61}$ | 1.5 | 0.945 | −0.5554 | −37.027 | 0.776 |
| $c_{62}$ | 4.5 | 4.097 | −0.4034 | −8.964 | 1.380 |
| $c_{63}$ | 7.5 | 7.162 | −0.3383 | −4.511 | 2.433 |
| $c_{64}$ | 10.5 | 10.131 | −0.3693 | −3.517 | 4.131 |
| $c_{65}$ | 13.5 | 13.304 | −0.1959 | −1.451 | 6.671 |

| Parameter | Value | Estimate | Bias | %Bias | MSE |
|---|---|---|---|---|---|
| $c_{71}$ | 1.5 | 1.027 | −0.4730 | −31.533 | 0.469 |
| $c_{72}$ | 4.5 | 4.184 | −0.3164 | −7.031 | 0.647 |
| $c_{73}$ | 7.5 | 7.293 | −0.2072 | −2.763 | 0.878 |
| $c_{74}$ | 10.5 | 10.402 | −0.0981 | −0.934 | 1.491 |
| $c_{75}$ | 13.5 | 13.461 | −0.0386 | −0.286 | 2.237 |
| $c_{81}$ | 2.0 | 1.407 | −0.5927 | −29.635 | 0.766 |
| $c_{82}$ | 6.0 | 5.995 | −0.0051 | −0.085 | 0.970 |
| $c_{83}$ | 10.0 | 10.572 | 0.5717 | 5.717 | 2.702 |
| $c_{84}$ | 14.0 | 15.047 | 1.0466 | 7.476 | 5.501 |
| $c_{85}$ | 18.0 | 19.582 | 1.5821 | 8.789 | 9.741 |
| $c_{91}$ | 2.0 | 1.342 | −0.6577 | −32.885 | 0.888 |
| $c_{92}$ | 6.0 | 5.515 | −0.4850 | −8.083 | 1.362 |
| $c_{93}$ | 10.0 | 9.620 | −0.3805 | −3.805 | 2.637 |
| $c_{94}$ | 14.0 | 13.782 | −0.2181 | −1.558 | 4.827 |
| $c_{95}$ | 18.0 | 17.828 | −0.1723 | −0.957 | 7.126 |
| $c_{101}$ | 2.5 | 1.240 | −1.2604 | −50.416 | 2.021 |

| | | | | | |
|---|---|---|---|---|---|
| $c_{102}$ | 7.5 | 5.803 | −1.6968 | −22.624 | 3.913 |
| $c_{103}$ | 12.5 | 10.344 | −2.1562 | −17.250 | 6.891 |
| $c_{104}$ | 17.5 | 14.833 | −2.6666 | −15.238 | 11.576 |
| $c_{105}$ | 22.5 | 19.315 | −3.1847 | −14.154 | 18.167 |
| Latent class sizes | | | | | |
| Class 1 | 0.080 | 0.102 | 0.0220 | 27.500 | |
| Class 2 | 0.170 | 0.162 | −0.0080 | −4.706 | |
| Class 3 | 0.250 | 0.235 | −0.0150 | −6.000 | |
| Class 4 | 0.250 | 0.234 | −0.0160 | −6.400 | |
| Class 5 | 0.170 | 0.165 | −0.0050 | −2.941 | |
| Class 6 | 0.080 | 0.101 | 0.0210 | 26.250 | |

**Table B3**

***Balanced Design, Normal d, N = 1,080, 10 Rater Pairs, 108 per Pair***

| Size | Parameter | Value | Estimate | Bias | %Bias | MSE |
|---|---|---|---|---|---|---|
| | | | Rater parameters | | | |
| 216 | $d_1$ | 1.0 | 0.993 | −0.0072 | −0.720 | 0.019 |
| 216 | $d_2$ | 2.0 | 1.932 | −0.0685 | −3.425 | 0.077 |
| 216 | $d_3$ | 2.0 | 1.882 | −0.1177 | −5.885 | 0.058 |
| 216 | $d_4$ | 3.0 | 2.786 | −0.2140 | −7.133 | 0.220 |
| 216 | $d_5$ | 3.0 | 2.758 | −0.2422 | −8.073 | 0.215 |
| 216 | $d_6$ | 3.0 | 2.991 | −0.0090 | −0.300 | 0.167 |
| 216 | $d_7$ | 3.0 | 3.095 | 0.0954 | 3.180 | 0.193 |
| 216 | $d_8$ | 4.0 | 3.195 | −0.8052 | −20.130 | 0.851 |
| 216 | $d_9$ | 4.0 | 4.203 | 0.2034 | 5.085 | 0.249 |
| 216 | $d_{10}$ | 5.0 | 4.210 | −0.7904 | −15.808 | 0.890 |
| | $c_{11}$ | 0.5 | 0.405 | −0.0952 | −19.040 | 0.089 |
| | $c_{12}$ | 1.5 | 1.453 | −0.0466 | −3.107 | 0.101 |
| | $c_{13}$ | 2.5 | 2.460 | −0.0405 | −1.620 | 0.139 |
| | $c_{14}$ | 3.5 | 3.488 | −0.0125 | −0.357 | 0.217 |
| | $c_{15}$ | 4.5 | 4.528 | 0.0280 | 0.622 | 0.231 |
| | $c_{21}$ | 1.0 | 0.783 | −0.2170 | −21.700 | 0.219 |
| | $c_{22}$ | 3.0 | 2.845 | −0.1554 | −5.180 | 0.385 |
| | $c_{23}$ | 5.0 | 4.858 | −0.1419 | −2.838 | 0.595 |
| | $c_{24}$ | 7.0 | 6.937 | −0.0633 | −0.904 | 0.873 |
| | $c_{25}$ | 9.0 | 8.960 | −0.0405 | −0.450 | 1.227 |
| | $c_{31}$ | 1.0 | 0.692 | −0.3085 | −30.850 | 0.217 |
| | $c_{32}$ | 3.0 | 2.723 | −0.2774 | −9.247 | 0.274 |
| | $c_{33}$ | 5.0 | 4.750 | −0.2504 | −5.008 | 0.385 |
| | $c_{34}$ | 7.0 | 6.749 | −0.2508 | −3.583 | 0.546 |

| Size | Parameter | Value | Estimate | Bias | %Bias | MSE |
|---|---|---|---|---|---|---|
|  | $c_{35}$ | 9.0 | 8.775 | −0.2251 | −2.501 | 0.808 |
|  | $c_{41}$ | 1.5 | 0.978 | −0.5223 | −34.820 | 0.656 |
|  | $c_{42}$ | 4.5 | 3.970 | −0.5300 | −11.778 | 0.934 |
|  | $c_{43}$ | 7.5 | 6.969 | −0.5310 | −7.080 | 1.639 |
|  | $c_{44}$ | 10.5 | 9.962 | −0.5378 | −5.122 | 2.377 |
|  | $c_{45}$ | 13.5 | 12.947 | −0.5526 | −4.093 | 3.487 |
|  | $c_{51}$ | 1.5 | 1.001 | −0.4992 | −33.280 | 0.602 |
|  | $c_{52}$ | 4.5 | 3.904 | −0.5965 | −13.256 | 1.048 |
|  | $c_{53}$ | 7.5 | 6.970 | −0.5303 | −7.071 | 1.405 |
|  | $c_{54}$ | 10.5 | 9.926 | −0.5736 | −5.463 | 2.239 |
|  | $c_{55}$ | 13.5 | 12.917 | −0.5833 | −4.321 | 3.421 |
|  | $c_{61}$ | 1.5 | 1.145 | −0.3548 | −23.653 | 0.522 |
|  | $c_{62}$ | 4.5 | 4.328 | −0.1720 | −3.822 | 0.670 |
|  | $c_{63}$ | 7.5 | 7.529 | 0.0289 | 0.385 | 1.184 |
|  | $c_{64}$ | 10.5 | 10.637 | 0.1365 | 1.300 | 2.009 |
|  | $c_{65}$ | 13.5 | 13.920 | 0.4197 | 3.109 | 3.626 |
|  | $c_{71}$ | 1.5 | 1.190 | −0.3106 | −20.707 | 0.484 |
|  | $c_{72}$ | 4.5 | 4.530 | 0.0302 | 0.671 | 0.752 |
|  | $c_{73}$ | 7.5 | 7.808 | 0.3075 | 4.100 | 1.419 |
|  | $c_{74}$ | 10.5 | 11.053 | 0.5528 | 5.265 | 2.566 |
|  | $c_{75}$ | 13.5 | 14.366 | 0.8656 | 6.412 | 4.403 |
|  | $c_{81}$ | 2.0 | 0.874 | −1.1260 | −56.300 | 1.627 |
|  | $c_{82}$ | 6.0 | 4.497 | −1.5035 | −25.058 | 3.112 |
|  | $c_{83}$ | 10.0 | 8.018 | −1.9816 | −19.816 | 5.520 |
|  | $c_{84}$ | 14.0 | 11.539 | −2.4606 | −17.576 | 8.459 |
|  | $c_{85}$ | 18.0 | 15.059 | −2.9406 | −16.337 | 12.715 |
|  | $c_{91}$ | 2.0 | 1.513 | −0.4867 | −24.335 | 0.790 |
|  | $c_{92}$ | 6.0 | 6.099 | 0.0987 | 1.645 | 0.788 |

| Size | Parameter | Value | Estimate | Bias | %Bias | MSE |
|---|---|---|---|---|---|---|
| | $c_{93}$ | 10.0 | 10.515 | 0.5153 | 5.153 | 1.628 |
| | $c_{94}$ | 14.0 | 14.915 | 0.9151 | 6.536 | 3.232 |
| | $c_{95}$ | 18.0 | 19.590 | 1.5904 | 8.836 | 6.603 |
| | $c_{101}$ | 2.5 | 1.451 | −1.0486 | −41.944 | 1.537 |
| | $c_{102}$ | 7.5 | 5.992 | −1.5076 | −20.101 | 3.252 |
| | $c_{103}$ | 12.5 | 10.590 | −1.9099 | −15.279 | 5.647 |
| | $c_{104}$ | 17.5 | 15.019 | −2.4808 | −14.176 | 9.425 |
| | $c_{105}$ | 22.5 | 19.753 | −2.7471 | −12.209 | 13.152 |
| | | | Latent class sizes | | | |
| | Class 1 | 0.080 | 0.103 | 0.0230 | 28.750 | |
| | Class 2 | 0.170 | 0.159 | −0.0110 | −6.471 | |
| | Class 3 | 0.250 | 0.234 | −0.0160 | −6.400 | |
| | Class 4 | 0.250 | 0.241 | −0.0090 | −3.600 | |
| | Class 5 | 0.170 | 0.159 | −0.0110 | −6.471 | |
| | Class 6 | 0.080 | 0.105 | 0.0250 | 31.250 | |

**Table B4**

***Unbalanced Design, Normal d, N = 1,080, 10 Rater Pairs***

| | Parameter | Value | Estimate | Bias | %Bias | MSE |
|---|---|---|---|---|---|---|
| | | | Rater parameters | | | |
| 370 | $d_1$ | 1.0 | 0.986 | −0.0139 | −1.390 | 0.009 |
| 50 | $d_2$ | 2.0 | 1.883 | −0.1175 | −8.780 | 0.194 |
| 200 | $d_3$ | 2.0 | 1.960 | −0.0404 | −2.515 | 0.070 |
| 140 | $d_4$ | 3.0 | 2.696 | −0.3042 | −13.607 | 0.275 |
| 60 | $d_5$ | 3.0 | 2.435 | −0.5646 | −17.430 | 0.484 |
| 120 | $d_6$ | 3.0 | 2.996 | −0.0043 | −13.790 | 0.201 |
| 280 | $d_7$ | 3.0 | 3.410 | 0.4100 | 2.167 | 0.350 |
| 400 | $d_8$ | 4.0 | 3.249 | −0.7510 | −19.070 | 0.693 |
| 230 | $d_9$ | 4.0 | 4.259 | 0.2592 | 11.953 | 0.291 |
| 310 | $d_{10}$ | 5.0 | 4.115 | −0.8850 | −33.388 | 1.034 |
| | $c_{11}$ | 0.5 | 0.447 | −0.0533 | −14.660 | 0.060 |
| | $c_{12}$ | 1.5 | 1.460 | −0.0398 | −1.313 | 0.056 |
| | $c_{13}$ | 2.5 | 2.501 | 0.0010 | 1.552 | 0.069 |
| | $c_{14}$ | 3.5 | 3.517 | 0.0167 | 2.417 | 0.094 |
| | $c_{15}$ | 4.5 | 4.533 | 0.0331 | 2.660 | 0.113 |
| | $c_{21}$ | 1.0 | 0.700 | −0.3005 | −41.090 | 0.674 |
| | $c_{22}$ | 3.0 | 2.721 | −0.2791 | −14.237 | 0.866 |
| | $c_{23}$ | 5.0 | 4.724 | −0.2757 | −8.862 | 1.393 |
| | $c_{24}$ | 7.0 | 6.788 | −0.2121 | −6.306 | 2.238 |
| | $c_{25}$ | 9.0 | 8.904 | −0.0959 | −4.751 | 3.150 |
| | $c_{31}$ | 1.0 | 0.792 | −0.2082 | −23.450 | 0.327 |
| | $c_{32}$ | 3.0 | 2.830 | −0.1696 | −1.973 | 0.383 |
| | $c_{33}$ | 5.0 | 4.917 | −0.0833 | −2.956 | 0.573 |
| | $c_{34}$ | 7.0 | 6.993 | −0.0075 | −4.414 | 0.922 |
| | $c_{35}$ | 9.0 | 9.064 | 0.0639 | 5.504 | 1.326 |
| | $c_{41}$ | 1.5 | 0.902 | −0.5982 | −51.147 | 0.800 |

| Parameter | Value | Estimate | Bias | %Bias | MSE |
|---|---|---|---|---|---|
| $c_{42}$ | 4.5 | 3.828 | −0.6719 | −18.473 | 1.271 |
| $c_{43}$ | 7.5 | 6.749 | −0.7513 | −14.465 | 1.885 |
| $c_{44}$ | 10.5 | 9.646 | −0.8541 | −10.626 | 2.978 |
| $c_{45}$ | 13.5 | 12.571 | −0.9291 | −8.887 | 4.293 |
| $c_{51}$ | 1.5 | 0.788 | −0.7119 | −60.727 | 1.198 |
| $c_{52}$ | 4.5 | 3.466 | −1.0336 | −24.851 | 2.022 |
| $c_{53}$ | 7.5 | 6.089 | −1.4109 | −17.681 | 3.307 |
| $c_{54}$ | 10.5 | 8.849 | −1.6511 | −15.172 | 4.928 |
| $c_{55}$ | 13.5 | 11.557 | −1.9428 | −13.393 | 7.349 |
| $c_{61}$ | 1.5 | 1.125 | −0.3752 | −50.807 | 0.608 |
| $c_{62}$ | 4.5 | 4.256 | −0.2439 | −19.693 | 0.993 |
| $c_{63}$ | 7.5 | 7.514 | 0.0135 | 15.033 | 1.683 |
| $c_{64}$ | 10.5 | 10.716 | 0.2161 | 11.291 | 2.471 |
| $c_{65}$ | 13.5 | 13.955 | 0.4548 | 9.986 | 4.016 |
| $c_{71}$ | 1.5 | 1.468 | −0.0318 | −36.427 | 0.426 |
| $c_{72}$ | 4.5 | 5.006 | 0.5064 | 3.853 | 1.039 |
| $c_{73}$ | 7.5 | 8.464 | 0.9640 | 1.711 | 2.085 |
| $c_{74}$ | 10.5 | 12.052 | 1.5516 | 4.038 | 4.337 |
| $c_{75}$ | 13.5 | 15.723 | 2.2230 | 6.173 | 8.508 |
| $c_{81}$ | 2.0 | 1.110 | −0.8902 | −58.085 | 1.063 |
| $c_{82}$ | 6.0 | 4.664 | −1.3365 | −26.143 | 2.298 |
| $c_{83}$ | 10.0 | 8.134 | −1.8665 | −18.622 | 4.402 |
| $c_{84}$ | 14.0 | 11.621 | −2.3786 | −16.774 | 7.406 |
| $c_{85}$ | 18.0 | 15.160 | −2.8398 | −15.198 | 10.747 |
| $c_{91}$ | 2.0 | 1.606 | −0.3945 | −50.550 | 0.771 |
| $c_{92}$ | 6.0 | 6.219 | 0.2185 | 18.187 | 1.068 |
| $c_{93}$ | 10.0 | 10.683 | 0.6832 | 11.929 | 2.007 |
| $c_{94}$ | 14.0 | 15.098 | 1.0981 | 9.118 | 3.965 |
| $c_{95}$ | 18.0 | 19.658 | 1.6576 | 7.248 | 6.944 |

| Parameter | Value | Estimate | Bias | %Bias | MSE |
|---|---|---|---|---|---|
| $c_{101}$ | 2.5 | 1.387 | −1.1130 | −65.588 | 1.681 |
| $c_{102}$ | 7.5 | 5.943 | −1.5566 | −38.940 | 3.385 |
| $c_{103}$ | 12.5 | 10.387 | −2.1126 | −33.582 | 6.402 |
| $c_{104}$ | 17.5 | 14.805 | −2.6952 | −30.991 | 10.584 |
| $c_{105}$ | 22.5 | 19.262 | −3.2382 | −29.630 | 15.763 |
| | | Latent class sizes | | | |
| Class 1 | 0.080 | 0.100 | 0.0200 | 25.000 | |
| Class 2 | 0.170 | 0.161 | −0.0090 | −5.294 | |
| Class 3 | 0.250 | 0.238 | −0.0120 | −4.800 | |
| Class 4 | 0.250 | 0.239 | −0.0110 | −4.400 | |
| Class 5 | 0.170 | 0.159 | −0.0110 | −6.471 | |
| Class 6 | 0.080 | 0.102 | 0.0220 | 27.500 | |

**Table B5**

***Balanced Design, Normal d, N = 1,080, 20 Rater Pairs, 54 per Pair***

| Size | Parameter | Value | Estimate | Bias | %Bias | MSE |
|---|---|---|---|---|---|---|
| | | | Rater parameters | | | |
| 216 | $d_1$ | 1.0 | 0.969 | −0.0312 | −3.120 | 0.019 |
| 216 | $d_2$ | 2.0 | 1.995 | −0.0054 | −0.270 | 0.093 |
| 216 | $d_3$ | 2.0 | 1.943 | −0.0572 | −2.860 | 0.100 |
| 216 | $d_4$ | 3.0 | 2.728 | −0.2720 | −9.067 | 0.254 |
| 216 | $d_5$ | 3.0 | 2.823 | −0.1768 | −5.893 | 0.289 |
| 216 | $d_6$ | 3.0 | 2.954 | −0.0459 | −1.530 | 0.211 |
| 216 | $d_7$ | 3.0 | 2.914 | −0.0858 | −2.860 | 0.160 |
| 216 | $d_8$ | 4.0 | 4.043 | 0.0427 | 1.068 | 0.313 |
| 216 | $d_9$ | 4.0 | 3.996 | −0.0044 | −0.110 | 0.352 |
| 216 | $d_{10}$ | 5.0 | 4.511 | −0.4889 | −9.778 | 0.625 |
| | $c_{11}$ | 0.5 | 0.383 | −0.1168 | −23.360 | 0.119 |
| | $c_{12}$ | 1.5 | 1.427 | −0.0728 | −4.853 | 0.101 |
| | $c_{13}$ | 2.5 | 2.459 | −0.0411 | −1.644 | 0.121 |
| | $c_{14}$ | 3.5 | 3.466 | −0.0345 | −0.986 | 0.163 |
| | $c_{15}$ | 4.5 | 4.463 | −0.0375 | −0.833 | 0.215 |
| | $c_{21}$ | 1.0 | 0.776 | −0.2245 | −22.450 | 0.278 |
| | $c_{22}$ | 3.0 | 2.870 | −0.1299 | −4.330 | 0.432 |
| | $c_{23}$ | 5.0 | 4.966 | −0.0344 | −0.688 | 0.589 |
| | $c_{24}$ | 7.0 | 7.062 | 0.0624 | 0.891 | 1.011 |
| | $c_{25}$ | 9.0 | 9.211 | 0.2110 | 2.344 | 1.621 |
| | $c_{31}$ | 1.0 | 0.762 | −0.2378 | −23.780 | 0.267 |
| | $c_{32}$ | 3.0 | 2.794 | −0.2060 | −6.867 | 0.452 |
| | $c_{33}$ | 5.0 | 4.849 | −0.1506 | −3.012 | 0.649 |
| | $c_{34}$ | 7.0 | 6.912 | −0.0880 | −1.257 | 1.100 |

| Size | Parameter | Value | Estimate | Bias | %Bias | MSE |
|---|---|---|---|---|---|---|
| | $c_{35}$ | 9.0 | 8.910 | −0.0896 | −0.996 | 1.673 |
| | $c_{41}$ | 1.5 | 0.939 | −0.5607 | −37.380 | 0.645 |
| | $c_{42}$ | 4.5 | 3.895 | −0.6048 | −13.440 | 0.975 |
| | $c_{43}$ | 7.5 | 6.818 | −0.6816 | −9.088 | 1.695 |
| | $c_{44}$ | 10.5 | 9.720 | −0.7801 | −7.430 | 2.697 |
| | $c_{45}$ | 13.5 | 12.652 | −0.8477 | −6.279 | 4.485 |
| | $c_{51}$ | 1.5 | 1.035 | −0.4647 | −30.980 | 0.567 |
| | $c_{52}$ | 4.5 | 4.057 | −0.4434 | −9.853 | 0.965 |
| | $c_{53}$ | 7.5 | 7.048 | −0.4517 | −6.023 | 1.964 |
| | $c_{54}$ | 10.5 | 10.030 | −0.4704 | −4.480 | 3.356 |
| | $c_{55}$ | 13.5 | 13.045 | −0.4554 | −3.373 | 4.957 |
| | $c_{61}$ | 1.5 | 1.124 | −0.3759 | −25.060 | 0.400 |
| | $c_{62}$ | 4.5 | 4.208 | −0.2920 | −6.489 | 0.687 |
| | $c_{63}$ | 7.5 | 7.373 | −0.1266 | −1.688 | 1.599 |
| | $c_{64}$ | 10.5 | 10.433 | −0.0671 | −0.639 | 2.621 |
| | $c_{65}$ | 13.5 | 13.615 | 0.1148 | 0.850 | 4.137 |
| | $c_{71}$ | 1.5 | 1.085 | −0.4153 | −27.687 | 0.506 |
| | $c_{72}$ | 4.5 | 4.155 | −0.3451 | −7.669 | 0.630 |
| | $c_{73}$ | 7.5 | 7.303 | −0.1966 | −2.621 | 1.072 |
| | $c_{74}$ | 10.5 | 10.350 | −0.1497 | −1.426 | 1.731 |
| | $c_{75}$ | 13.5 | 13.537 | 0.0370 | 0.274 | 2.739 |
| | $c_{81}$ | 2.0 | 1.418 | −0.5818 | −29.090 | 0.683 |
| | $c_{82}$ | 6.0 | 5.728 | −0.2718 | −4.530 | 0.984 |
| | $c_{83}$ | 10.0 | 10.103 | 0.1030 | 1.030 | 1.916 |
| | $c_{84}$ | 14.0 | 14.423 | 0.4230 | 3.021 | 4.064 |
| | $c_{85}$ | 18.0 | 18.816 | 0.8163 | 4.535 | 7.243 |
| | $c_{91}$ | 2.0 | 1.370 | −0.6303 | −31.513 | 1.030 |
| | $c_{92}$ | 6.0 | 5.707 | −0.2928 | −4.880 | 1.579 |

| Size | Parameter | Value | Estimate | Bias | %Bias | MSE |
|---|---|---|---|---|---|---|
| | $c_{93}$ | 10.0 | 10.050 | 0.0499 | 0.499 | 2.716 |
| | $c_{94}$ | 14.0 | 14.180 | 0.1798 | 1.284 | 4.435 |
| | $c_{95}$ | 18.0 | 18.556 | 0.5560 | 3.089 | 6.848 |
| | $c_{101}$ | 2.5 | 1.395 | −1.1047 | −44.188 | 1.957 |
| | $c_{102}$ | 7.5 | 6.401 | −1.0988 | −14.650 | 2.676 |
| | $c_{103}$ | 12.5 | 11.337 | −1.1629 | −9.303 | 4.070 |
| | $c_{104}$ | 17.5 | 16.158 | −1.3417 | −7.667 | 6.416 |
| | $c_{105}$ | 22.5 | 20.990 | −1.5104 | −6.713 | 9.830 |
| | | | Latent class sizes | | | |
| | Class 1 | 0.080 | 0.101 | 0.0210 | 26.250 | |
| | Class 2 | 0.170 | 0.161 | −0.0090 | −5.294 | |
| | Class 3 | 0.250 | 0.237 | −0.0130 | −5.200 | |
| | Class 4 | 0.250 | 0.242 | −0.0080 | −3.200 | |
| | Class 5 | 0.170 | 0.159 | −0.0110 | −6.471 | |
| | Class 6 | 0.080 | 0.100 | 0.0200 | 25.000 | |

**Table B6**

***Unbalanced Design, Normal d, N = 1,080, 20 Rater Pairs***

| Size | Parameter | Value | Estimate | Bias | %Bias | MSE |
|---|---|---|---|---|---|---|
| | | | Rater parameters | | | |
| 370 | $d_1$ | 1.0 | 0.973 | −0.0273 | −2.730 | 0.011 |
| 50 | $d_2$ | 2.0 | 1.582 | −0.4178 | −20.890 | 0.369 |
| 200 | $d_3$ | 2.0 | 1.875 | −0.1250 | −6.250 | 0.145 |
| 140 | $d_4$ | 3.0 | 2.935 | −0.0651 | −2.170 | 0.252 |
| 60 | $d_5$ | 3.0 | 2.546 | −0.4541 | −15.137 | 0.464 |
| 120 | $d_6$ | 3.0 | 2.816 | −0.1844 | −6.147 | 0.282 |
| 280 | $d_7$ | 3.0 | 2.869 | −0.1312 | −4.373 | 0.147 |
| 400 | $d_8$ | 4.0 | 4.282 | 0.2822 | 7.055 | 0.344 |
| 230 | $d_9$ | 4.0 | 3.754 | −0.2459 | −6.148 | 0.423 |
| 310 | $d_{10}$ | 5.0 | 4.640 | −0.3596 | −7.192 | 0.446 |
| | $c_{11}$ | 0.5 | 0.423 | −0.0774 | −15.480 | 0.072 |
| | $c_{12}$ | 1.5 | 1.441 | −0.0586 | −3.907 | 0.072 |
| | $c_{13}$ | 2.5 | 2.446 | −0.0545 | −2.180 | 0.086 |
| | $c_{14}$ | 3.5 | 3.455 | −0.0449 | −1.283 | 0.120 |
| | $c_{15}$ | 4.5 | 4.450 | −0.0498 | −1.107 | 0.146 |
| | $c_{21}$ | 1.0 | 0.345 | −0.6551 | −65.510 | 1.133 |
| | $c_{22}$ | 3.0 | 2.152 | −0.8478 | −28.260 | 1.611 |
| | $c_{23}$ | 5.0 | 3.994 | −1.0056 | −20.112 | 2.506 |
| | $c_{24}$ | 7.0 | 5.777 | −1.2232 | −17.474 | 3.937 |
| | $c_{25}$ | 9.0 | 7.533 | −1.4671 | −16.301 | 5.552 |
| | $c_{31}$ | 1.0 | 0.747 | −0.2532 | −25.320 | 0.348 |
| | $c_{32}$ | 3.0 | 2.705 | −0.2948 | −9.827 | 0.583 |
| | $c_{33}$ | 5.0 | 4.715 | −0.2849 | −5.698 | 0.999 |
| | $c_{34}$ | 7.0 | 6.700 | −0.3001 | −4.287 | 1.624 |

| Size | Parameter | Value | Estimate | Bias | %Bias | MSE |
|---|---|---|---|---|---|---|
| | $c_{35}$ | 9.0 | 8.726 | −0.2741 | −3.046 | 2.412 |
| | $c_{41}$ | 1.5 | 1.036 | −0.4639 | −30.927 | 0.675 |
| | $c_{42}$ | 4.5 | 4.083 | −0.4166 | −9.258 | 1.072 |
| | $c_{43}$ | 7.5 | 7.300 | −0.1997 | −2.663 | 1.911 |
| | $c_{44}$ | 10.5 | 10.501 | 0.0011 | 0.010 | 2.994 |
| | $c_{45}$ | 13.5 | 13.634 | 0.1340 | 0.993 | 5.074 |
| | $c_{51}$ | 1.5 | 0.821 | −0.6795 | −45.300 | 1.236 |
| | $c_{52}$ | 4.5 | 3.601 | −0.8994 | −19.987 | 1.987 |
| | $c_{53}$ | 7.5 | 6.419 | −1.0809 | −14.412 | 3.037 |
| | $c_{54}$ | 10.5 | 9.148 | −1.3523 | −12.879 | 4.943 |
| | $c_{55}$ | 13.5 | 12.045 | −1.4547 | −10.776 | 7.354 |
| | $c_{61}$ | 1.5 | 0.985 | −0.5154 | −34.360 | 0.762 |
| | $c_{62}$ | 4.5 | 3.987 | −0.5132 | −11.404 | 1.243 |
| | $c_{63}$ | 7.5 | 7.065 | −0.4349 | −5.799 | 1.964 |
| | $c_{64}$ | 10.5 | 10.138 | −0.3618 | −3.446 | 3.316 |
| | $c_{65}$ | 13.5 | 13.014 | −0.4856 | −3.597 | 4.985 |
| | $c_{71}$ | 1.5 | 1.073 | −0.4275 | −28.500 | 0.411 |
| | $c_{72}$ | 4.5 | 4.126 | −0.3741 | −8.313 | 0.581 |
| | $c_{73}$ | 7.5 | 7.201 | −0.2994 | −3.992 | 0.882 |
| | $c_{74}$ | 10.5 | 10.209 | −0.2911 | −2.772 | 1.513 |
| | $c_{75}$ | 13.5 | 13.254 | −0.2457 | −1.820 | 2.390 |
| | $c_{81}$ | 2.0 | 1.624 | −0.3764 | −18.820 | 0.513 |
| | $c_{82}$ | 6.0 | 6.216 | 0.2163 | 3.605 | 0.960 |
| | $c_{83}$ | 10.0 | 10.714 | 0.7136 | 7.136 | 2.414 |
| | $c_{84}$ | 14.0 | 15.253 | 1.2526 | 8.947 | 5.108 |
| | $c_{85}$ | 18.0 | 19.874 | 1.8735 | 10.408 | 9.231 |
| | $c_{91}$ | 2.0 | 1.372 | −0.6277 | −31.384 | 0.800 |
| | $c_{92}$ | 6.0 | 5.509 | −0.4910 | −8.184 | 1.497 |

| Size | Parameter | Value | Estimate | Bias | %Bias | MSE |
|---|---|---|---|---|---|---|
| | $c_{93}$ | 10.0 | 9.479 | −0.5215 | −5.215 | 2.769 |
| | $c_{94}$ | 14.0 | 13.461 | −0.5391 | −3.851 | 4.650 |
| | $c_{95}$ | 18.0 | 17.447 | −0.5534 | −3.074 | 7.740 |
| | $c_{101}$ | 2.5 | 1.607 | −0.8930 | −35.722 | 1.363 |
| | $c_{102}$ | 7.5 | 6.727 | −0.7730 | −10.307 | 1.834 |
| | $c_{103}$ | 12.5 | 11.564 | −0.9365 | −7.492 | 3.177 |
| | $c_{104}$ | 17.5 | 16.724 | −0.7757 | −4.432 | 5.381 |
| | $c_{105}$ | 22.5 | 21.717 | −0.7833 | −3.481 | 7.844 |
| | | | Latent class sizes | | | |
| | Class 1 | 0.080 | 0.098 | 0.0180 | 22.500 | |
| | Class 2 | 0.170 | 0.160 | −0.0100 | −5.882 | |
| | Class 3 | 0.250 | 0.241 | −0.0090 | −3.600 | |
| | Class 4 | 0.250 | 0.234 | −0.0160 | −6.400 | |
| | Class 5 | 0.170 | 0.167 | −0.0030 | −1.765 | |
| | Class 6 | 0.080 | 0.100 | 0.0200 | 25.000 | |

## Appendix C

## Evaluation of the Estimated Standard Errors for *d* and the Latent Class Sizes

**Table C1**

***BIB Design, Normal d, N = 1,080, 45 Rater Pairs, 24 per Pair***

| Size | Parameter | SD | Mean SE | Bias | %Bias |
|---|---|---|---|---|---|
| 216 | $d_1$ | 0.145 | 0.136 | −0.009 | −6.071 |
| 216 | $d_2$ | 0.238 | 0.258 | 0.020 | 8.476 |
| 216 | $d_3$ | 0.276 | 0.256 | −0.020 | −7.340 |
| 216 | $d_4$ | 0.504 | 0.469 | −0.035 | −6.933 |
| 216 | $d_5$ | 0.464 | 0.453 | −0.011 | −2.453 |
| 216 | $d_6$ | 0.439 | 0.460 | 0.021 | 4.748 |
| 216 | $d_7$ | 0.410 | 0.464 | 0.053 | 13.018 |
| 216 | $d_8$ | 0.591 | 0.684 | 0.094 | 15.834 |
| 216 | $d_9$ | 0.613 | 0.678 | 0.065 | 10.580 |
| 216 | $d_{10}$ | 0.615 | 0.856 | 0.241 | 39.150 |
|  | *Class Size 1* | 0.020 | 0.020 | 0.000 | 1.574 |
|  | *Class Size 2* | 0.026 | 0.026 | 0.000 | 0.990 |
|  | *Class Size 3* | 0.027 | 0.030 | 0.003 | 9.971 |
|  | *Class Size 4* | 0.029 | 0.030 | 0.001 | 2.634 |
|  | *Class Size 5* | 0.024 | 0.026 | 0.002 | 6.952 |
|  | *Class Size 6* | 0.017 | 0.020 | 0.003 | 20.627 |

**Table C2**

***Unbalanced Design, Normal d, N = 1,080, 45 Rater Pairs, 24 per Pair***

| Size | Parameter | SD | Mean SE | Bias | %Bias |
|---|---|---|---|---|---|
| 370 | $d_1$ | 0.110 | 0.101 | −0.009 | −8.514 |
| 50 | $d_2$ | 0.377 | 0.455 | 0.078 | 20.562 |
| 200 | $d_3$ | 0.284 | 0.258 | −0.026 | −9.251 |
| 140 | $d_4$ | 0.509 | 0.544 | 0.035 | 6.834 |
| 60 | $d_5$ | 0.492 | 0.619 | 0.127 | 25.864 |
| 120 | $d_6$ | 0.585 | 0.558 | −0.027 | −4.681 |
| 280 | $d_7$ | 0.358 | 0.379 | 0.021 | 5.951 |
| 400 | $d_8$ | 0.575 | 0.694 | 0.119 | 20.699 |
| 230 | $d_9$ | 0.604 | 0.671 | 0.067 | 11.060 |
| 310 | $d_{10}$ | 0.601 | 0.816 | 0.215 | 35.730 |
|  | *Class Size 1* | 0.019 | 0.021 | 0.002 | 12.903 |
|  | *Class Size 2* | 0.027 | 0.028 | 0.001 | 2.941 |
|  | *Class Size 3* | 0.030 | 0.032 | 0.002 | 8.108 |
|  | *Class Size 4* | 0.030 | 0.031 | 0.001 | 1.974 |
|  | *Class Size 5* | 0.025 | 0.027 | 0.002 | 7.143 |
|  | *Class Size 6* | 0.019 | 0.020 | 0.001 | 5.263 |

**Table C3**
***BIB Design, Normal d, N = 1,080, 10 Rater Pairs, 108 per Pair***

| Size | Parameter | SD | Mean SE | Bias | %Bias |
|---|---|---|---|---|---|
| 216 | $d_1$ | 0.137 | 0.134 | −0.003 | −2.047 |
| 216 | $d_2$ | 0.270 | 0.301 | 0.031 | 11.481 |
| 216 | $d_3$ | 0.211 | 0.250 | 0.039 | 18.483 |
| 216 | $d_4$ | 0.420 | 0.569 | 0.149 | 35.573 |
| 216 | $d_5$ | 0.398 | 0.569 | 0.171 | 42.965 |
| 216 | $d_6$ | 0.411 | 0.631 | 0.220 | 53.678 |
| 216 | $d_7$ | 0.431 | 0.681 | 0.250 | 58.035 |
| 216 | $d_8$ | 0.453 | 0.747 | 0.294 | 64.863 |
| 216 | $d_9$ | 0.458 | 0.855 | 0.396 | 86.491 |
| 216 | $d_{10}$ | 0.518 | 0.871 | 0.353 | 68.213 |
|  | *Class Size 1* | 0.019 | 0.021 | 0.002 | 9.948 |
|  | *Class Size 2* | 0.030 | 0.027 | −0.003 | −11.184 |
|  | *Class Size 3* | 0.031 | 0.031 | 0.000 | 0.977 |
|  | *Class Size 4* | 0.033 | 0.031 | −0.002 | −6.344 |
|  | *Class Size 5* | 0.028 | 0.028 | 0.001 | 1.818 |
|  | *Class Size 6* | 0.017 | 0.022 | 0.005 | 27.168 |

**Table C4**

***Unbalanced Design, Normal d, N = 1,080, 10 Rater Pairs***

| Size | Parameter | SD | Mean SE | Bias | %Bias |
|---|---|---|---|---|---|
| 370 | $d_1$ | 0.097 | 0.106 | 0.009 | 9.731 |
| 50 | $d_2$ | 0.427 | 0.489 | 0.062 | 14.627 |
| 200 | $d_3$ | 0.262 | 0.268 | 0.006 | 2.290 |
| 140 | $d_4$ | 0.429 | 0.605 | 0.176 | 40.894 |
| 60 | $d_5$ | 0.409 | 0.654 | 0.245 | 59.980 |
| 120 | $d_6$ | 0.450 | 0.731 | 0.281 | 62.300 |
| 280 | $d_7$ | 0.429 | 0.763 | 0.334 | 77.951 |
| 400 | $d_8$ | 0.361 | 0.695 | 0.334 | 92.463 |
| 230 | $d_9$ | 0.476 | 0.902 | 0.426 | 89.613 |
| 310 | $d_{10}$ | 0.504 | 0.859 | 0.355 | 70.484 |
| | *Class Size 1* | 0.018 | 0.021 | 0.003 | 19.318 |
| | *Class Size 2* | 0.029 | 0.028 | −0.001 | −2.778 |
| | *Class Size 3* | 0.033 | 0.031 | −0.002 | −5.775 |
| | *Class Size 4* | 0.033 | 0.031 | −0.002 | −6.907 |
| | *Class Size 5* | 0.027 | 0.028 | 0.001 | 2.190 |
| | *Class Size 6* | 0.017 | 0.020 | 0.003 | 14.943 |

**Table C5**

***BIB Design, Normal d, N = 1,080, 20 Rater Pairs, 54 per Pair***

| Size | Parameter | SD | Mean SE | Bias | %Bias |
|---|---|---|---|---|---|
| 216 | $d_1$ | 0.133 | 0.140 | 0.007 | 5.042 |
| 216 | $d_2$ | 0.307 | 0.311 | 0.004 | 1.270 |
| 216 | $d_3$ | 0.313 | 0.327 | 0.014 | 4.473 |
| 216 | $d_4$ | 0.426 | 0.481 | 0.055 | 12.913 |
| 216 | $d_5$ | 0.510 | 0.486 | −0.024 | −4.689 |
| 216 | $d_6$ | 0.459 | 0.462 | 0.003 | 0.570 |
| 216 | $d_7$ | 0.392 | 0.423 | 0.031 | 7.779 |
| 216 | $d_8$ | 0.561 | 0.658 | 0.097 | 17.278 |
| 216 | $d_9$ | 0.596 | 0.700 | 0.104 | 17.390 |
| 216 | $d_{10}$ | 0.624 | 0.852 | 0.228 | 36.447 |
| | *Class Size 1* | 0.017 | 0.020 | 0.003 | 14.416 |
| | *Class Size 2* | 0.028 | 0.026 | −0.002 | −6.609 |
| | *Class Size 3* | 0.031 | 0.030 | −0.001 | −3.007 |
| | *Class Size 4* | 0.028 | 0.030 | 0.002 | 7.720 |
| | *Class Size 5* | 0.024 | 0.025 | 0.001 | 3.263 |
| | *Class Size 6* | 0.018 | 0.019 | 0.001 | 4.110 |

**Table C6**

***Unbalanced Design, Normal d, N = 1,080, 20 Rater Pairs***

| Size | Parameter | SD | Mean SE | Bias | %Bias |
|---|---|---|---|---|---|
| 370 | $d_1$ | 0.103 | 0.111 | 0.008 | 7.725 |
| 50 | $d_2$ | 0.444 | 0.564 | 0.120 | 27.122 |
| 200 | $d_3$ | 0.361 | 0.396 | 0.035 | 9.547 |
| 140 | $d_4$ | 0.500 | 0.694 | 0.194 | 38.828 |
| 60 | $d_5$ | 0.510 | 0.683 | 0.173 | 33.843 |
| 120 | $d_6$ | 0.500 | 0.582 | 0.082 | 16.286 |
| 280 | $d_7$ | 0.362 | 0.369 | 0.007 | 2.032 |
| 400 | $d_8$ | 0.516 | 0.674 | 0.158 | 30.576 |
| 230 | $d_9$ | 0.605 | 0.681 | 0.076 | 12.489 |
| 310 | $d_{10}$ | 0.566 | 0.818 | 0.252 | 44.623 |
|  | *Class Size 1* | 0.017 | 0.018 | 0.001 | 3.926 |
|  | *Class Size 2* | 0.027 | 0.025 | −0.002 | −6.472 |
|  | *Class Size 3* | 0.029 | 0.029 | 0.000 | 0.694 |
|  | *Class Size 4* | 0.035 | 0.029 | −0.006 | −16.378 |
|  | *Class Size 5* | 0.024 | 0.025 | 0.001 | 2.375 |
|  | *Class Size 6* | 0.017 | 0.018 | 0.001 | 2.975 |

## Appendix D

## Evaluation of Parameter Estimates: HRM-SDT Model

**Table D1**

***Two CR Items, Three Scores per CR Item, Fully Crossed, N = 3,000***

| Parameter | Value | Estimate | Bias | %Bias | MSE |
|---|---|---|---|---|---|
| | | Signal detection parameters | | | |
| $d_1$ | 2 | 2.003 | 0.003 | 0.150 | 0.004 |
| $d_2$ | 3 | 3.037 | 0.037 | 1.233 | 0.011 |
| $d_3$ | 4 | 4.015 | 0.015 | 0.375 | 0.028 |
| $d_4$ | 2 | 1.998 | −0.002 | −0.100 | 0.003 |
| $d_5$ | 3 | 3.002 | 0.002 | 0.067 | 0.010 |
| $d_6$ | 4 | 3.983 | −0.017 | −0.425 | 0.023 |
| $c_{11}$ | 1 | 1.022 | 0.022 | 2.200 | 0.019 |
| $c_{12}$ | 3 | 3.016 | 0.016 | 0.533 | 0.020 |
| $c_{13}$ | 5 | 5.013 | 0.013 | 0.260 | 0.030 |
| $c_{14}$ | 7 | 7.009 | 0.009 | 0.129 | 0.043 |
| $c_{15}$ | 9 | 9.008 | 0.008 | 0.089 | 0.063 |
| $c_{21}$ | 1.5 | 1.562 | 0.062 | 4.133 | 0.046 |
| $c_{22}$ | 4.5 | 4.573 | 0.073 | 1.622 | 0.057 |
| $c_{23}$ | 7.5 | 7.596 | 0.096 | 1.280 | 0.085 |
| $c_{24}$ | 10.5 | 10.630 | 0.127 | 1.210 | 0.140 |
| $c_{25}$ | 13.5 | 13.650 | 0.154 | 1.141 | 0.200 |
| $c_{31}$ | 2 | 2.016 | 0.016 | 0.800 | 0.076 |
| $c_{32}$ | 6 | 6.048 | 0.048 | 0.800 | 0.114 |
| $c_{33}$ | 10 | 10.036 | 0.036 | 0.360 | 0.193 |
| $c_{34}$ | 14 | 14.052 | 0.052 | 0.371 | 0.364 |
| $c_{35}$ | 18 | 18.056 | 0.056 | 0.311 | 0.518 |
| $c_{41}$ | 1 | 0.978 | −0.022 | −2.200 | 0.017 |

| | | | | | |
|---|---|---|---|---|---|
| $c_{42}$ | 3 | 2.975 | −0.025 | −0.833 | 0.020 |
| $c_{43}$ | 5 | 4.984 | −0.016 | −0.320 | 0.025 |
| $c_{44}$ | 7 | 6.990 | −0.010 | −0.143 | 0.033 |
| $c_{45}$ | 9 | 9.004 | 0.004 | 0.044 | 0.040 |
| $c_{51}$ | 1.5 | 1.477 | −0.023 | −1.533 | 0.037 |
| $c_{52}$ | 4.5 | 4.483 | −0.017 | −0.378 | 0.044 |
| $c_{53}$ | 7.5 | 7.482 | −0.018 | −0.240 | 0.067 |
| $c_{54}$ | 10.5 | 10.52 | 0.017 | 0.162 | 0.109 |
| $c_{55}$ | 13.5 | 13.53 | 0.034 | 0.252 | 0.162 |
| $c_{61}$ | 2 | 1.921 | −0.079 | −3.950 | 0.101 |
| $c_{62}$ | 6 | 5.936 | −0.064 | −1.067 | 0.107 |
| $c_{63}$ | 10 | 9.940 | −0.060 | −0.600 | 0.161 |
| $c_{64}$ | 14 | 13.954 | −0.046 | −0.329 | 0.227 |
| $c_{65}$ | 18 | 17.974 | −0.026 | −0.144 | 0.370 |
| Parameter | Value | Estimate | Bias | %Bias | MSE |
| | | CR Item parameters | | | |
| $a_{1}$ | 1 | 1.227 | 0.227 | 22.700 | 0.105 |
| $a_{2}$ | 1.5 | 1.265 | −0.235 | −15.667 | 0.091 |
| $b_{11}$ | −2 | −2.344 | −0.344 | −17.200 | 0.290 |
| $b_{12}$ | −1 | −1.177 | −0.177 | −17.700 | 0.071 |
| $b_{13}$ | 0 | −0.045 | −0.045 | — | 0.011 |
| $b_{14}$ | 0.5 | 0.604 | 0.104 | 20.800 | 0.029 |
| $b_{15}$ | 1 | 1.232 | 0.232 | 23.200 | 0.133 |
| $b_{21}$ | −3 | −2.599 | −0.401 | −13.367 | 0.283 |
| $b_{22}$ | −1.5 | −1.326 | −0.174 | −11.600 | 0.059 |
| $b_{23}$ | 0 | 0.006 | 0.006 | — | 0.007 |
| $b_{24}$ | 1.5 | 1.315 | −0.185 | −12.333 | 0.067 |
| $b_{25}$ | 3 | 2.616 | −0.384 | −12.800 | 0.243 |

*Note*. %Bias is not defined when the population parameter is zero.

**Table D2**

***Three CR Items, Three Scores per CR Item, Fully Crossed, N = 3,000***

| Parameter | Value | Estimate | Bias | %Bias | MSE |
|---|---|---|---|---|---|
| | | Rater parameters | | | |
| $d_1$ | 2 | 2.008 | 0.008 | 0.400 | 0.007 |
| $d_2$ | 3 | 3.044 | 0.044 | 1.467 | 0.016 |
| $d_3$ | 4 | 4.02 | 0.020 | 0.500 | 0.029 |
| $d_4$ | 2 | 2.001 | 0.001 | 0.050 | 0.004 |
| $d_5$ | 3 | 3.006 | 0.006 | 0.200 | 0.011 |
| $d_6$ | 4 | 3.987 | −0.013 | −0.325 | 0.023 |
| $d_7$ | 2 | 2.001 | 0.001 | 0.050 | 0.003 |
| $d_8$ | 3 | 3.012 | 0.012 | 0.400 | 0.006 |
| $d_9$ | 4 | 4.007 | 0.007 | 0.175 | 0.023 |
| $c_{11}$ | 1 | 1.031 | 0.031 | 3.100 | 0.123 |
| $c_{12}$ | 3 | 3.024 | 0.024 | 0.800 | 0.142 |
| $c_{13}$ | 5 | 5.022 | 0.022 | 0.440 | 0.149 |
| $c_{14}$ | 7 | 7.018 | 0.018 | 0.257 | 0.163 |
| $c_{15}$ | 9 | 9.016 | 0.016 | 0.178 | 0.175 |
| $c_{21}$ | 1.5 | 1.575 | 0.075 | 5.000 | 0.269 |
| $c_{22}$ | 4.5 | 4.584 | 0.084 | 1.867 | 0.299 |
| $c_{23}$ | 7.5 | 7.607 | 0.107 | 1.427 | 0.333 |
| $c_{24}$ | 10.5 | 10.639 | 0.139 | 1.324 | 0.375 |
| $c_{25}$ | 13.5 | 13.662 | 0.162 | 1.200 | 0.427 |
| $c_{31}$ | 2 | 2.025 | 0.024 | 1.199 | 0.391 |
| $c_{32}$ | 6 | 6.050 | 0.050 | 0.833 | 0.423 |
| $c_{33}$ | 10 | 10.037 | 0.037 | 0.370 | 0.444 |
| $c_{34}$ | 14 | 14.052 | 0.052 | 0.371 | 0.570 |
| $c_{35}$ | 18 | 18.049 | 0.049 | 0.272 | 0.695 |

| | | | | | |
|---|---|---|---|---|---|
| $c_{41}$ | 1 | 0.967 | −0.033 | −3.300 | 0.034 |
| $c_{42}$ | 3 | 2.963 | −0.037 | −1.233 | 0.041 |
| $c_{43}$ | 5 | 4.971 | −0.029 | −0.580 | 0.048 |
| $c_{44}$ | 7 | 6.978 | −0.022 | −0.314 | 0.057 |
| $c_{45}$ | 9 | 8.992 | −0.008 | −0.089 | 0.061 |
| $c_{51}$ | 1.5 | 1.459 | −0.041 | −2.733 | 0.069 |
| $c_{52}$ | 4.5 | 4.463 | −0.037 | −0.822 | 0.087 |
| $c_{53}$ | 7.5 | 7.460 | −0.040 | −0.533 | 0.111 |
| $c_{54}$ | 10.5 | 10.493 | −0.007 | −0.067 | 0.151 |
| $c_{55}$ | 13.5 | 13.510 | 0.010 | 0.074 | 0.204 |
| $c_{61}$ | 2 | 1.896 | −0.104 | −5.200 | 0.161 |
| $c_{62}$ | 6 | 5.912 | −0.088 | −1.467 | 0.189 |
| $c_{63}$ | 10 | 9.912 | −0.088 | −0.880 | 0.258 |
| $c_{64}$ | 14 | 13.922 | −0.078 | −0.557 | 0.327 |
| $c_{65}$ | 18 | 17.941 | −0.059 | −0.328 | 0.458 |
| $c_{71}$ | 1 | 1.009 | 0.009 | 0.900 | 0.015 |
| Parameter | Value | Estimate | Bias | %Bias | MSE |
| $c_{72}$ | 3 | 2.995 | −0.005 | −0.167 | 0.016 |
| $c_{73}$ | 5 | 5.002 | 0.002 | 0.040 | 0.023 |
| $c_{74}$ | 7 | 6.997 | −0.003 | −0.430 | 0.032 |
| $c_{75}$ | 9 | 9.020 | 0.020 | 0.222 | 0.045 |
| $c_{81}$ | 1.5 | 1.503 | 0.003 | 0.200 | 0.026 |
| $c_{82}$ | 4.5 | 4.527 | 0.027 | 0.600 | 0.035 |
| $c_{83}$ | 7.5 | 7.527 | 0.027 | 0.360 | 0.052 |
| $c_{84}$ | 10.5 | 10.550 | 0.050 | 0.476 | 0.082 |
| $c_{85}$ | 13.5 | 13.580 | 0.080 | 0.593 | 0.127 |
| $c_{91}$ | 2 | 2.014 | 0.014 | 0.700 | 0.067 |
| $c_{92}$ | 6 | 6.018 | 0.018 | 0.300 | 0.118 |
| $c_{93}$ | 10 | 10.014 | 0.014 | 0.140 | 0.191 |
| $c_{94}$ | 14 | 14.042 | 0.042 | 0.300 | 0.278 |

| | | | | | |
|---|---|---|---|---|---|
| $c_{95}$ | 18 | 18.047 | 0.047 | 0.261 | 0.454 |

| Parameter | Value | Estimate | Bias | %Bias | MSE |
|---|---|---|---|---|---|
| | | CR Item parameters | | | |
| $a_1$ | 1 | 1.016 | 0.016 | 1.600 | 0.014 |
| $a_2$ | 1.5 | 1.572 | 0.072 | 4.800 | 0.836 |
| $a_3$ | 0.5 | 0.503 | 0.003 | 0.600 | 0.001 |
| $b_{11}$ | −2 | −2.058 | −0.058 | −2.900 | 0.291 |
| $b_{12}$ | −1 | −1.022 | −0.022 | −2.200 | 0.053 |
| $b_{13}$ | 0 | −0.016 | −0.016 | — | 0.022 |
| $b_{14}$ | 0.5 | 0.510 | 0.010 | 2.080 | 0.022 |
| $b_{15}$ | 1 | 1.054 | 0.054 | 5.400 | 0.326 |
| $b_{21}$ | −3 | −3.059 | −0.059 | −1.967 | 2.080 |
| $b_{22}$ | −1.5 | −1.525 | −0.025 | −1.667 | 0.290 |
| $b_{23}$ | 0 | 0.003 | 0.003 | — | 0.075 |
| $b_{24}$ | 1.5 | 1.550 | 0.050 | 3.333 | 0.340 |
| $b_{25}$ | 3 | 3.135 | 0.135 | 4.500 | 2.214 |
| $b_{31}$ | −1 | −0.999 | 0.001 | 0.100 | 0.016 |
| $b_{32}$ | −0.5 | −0.504 | −0.004 | −0.800 | 0.008 |
| $b_{33}$ | 0 | 0.001 | 0.001 | — | 0.006 |
| $b_{34}$ | 0.5 | 0.502 | 0.002 | 0.400 | 0.007 |
| $b_{35}$ | 1 | 0.996 | −0.004 | −0.400 | 0.017 |

*Note*. %Bias is not defined when the population parameter is zero.