

*Studies of a Latent-Class
Signal-Detection Model for
Constructed-Response Scoring*

Lawrence T. DeCarlo

December 2008

ETS RR-08-63



Studies of a Latent-Class Signal-Detection Model for Constructed-Response Scoring

Lawrence T. DeCarlo¹

Teachers College, Columbia University, NY

December 2008

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

ETS Research Reports provide preliminary and limited dissemination of ETS research prior to publication. To obtain a PDF or a print copy of a report, please visit:

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

Copyright © 2008 by Educational Testing Service. All rights reserved.

ETS, the ETS logo, GRE, and LISTENING. LEARNING. LEADING. are registered trademarks of Educational Testing Service (ETS).

PRAXIS is a trademark of ETS.

SAT is a registered trademark of the College Board.



Abstract

Rater behavior in essay grading can be viewed as a signal-detection task, in that raters attempt to discriminate between latent classes of essays, with the latent classes being defined by a scoring rubric. The present report examines basic aspects of an approach to constructed-response (CR) scoring via a latent-class signal-detection model. The model provides a psychological framework for CR scoring and includes rater parameters with a clear cognitive basis. Simulations are used to examine how well rater parameters and latent-class sizes are recovered as well as the accuracy of classification. The relation of rater parameters to agreement statistics and classification accuracy is examined. The effects of using a balanced, incomplete block design are compared to those for a fully crossed design. The model is applied to several ETS datasets.

Key words: Constructed responses, rater effects, signal detection theory, latent class models, classification, agreement, incomplete block design

Acknowledgments

The author thanks Brianna Moore for her assistance.

Table of Contents

	Page
Introduction.....	1
Simulated Data: Fully Crossed Design.....	5
Methods.....	5
Results.....	7
Discussion.....	12
Simulated Data: Balanced Incomplete Block (BIB) Design.....	13
Methods.....	13
Results.....	14
Discussion.....	17
ETS Data.....	18
Example 1: Writing Assessment.....	18
Example 2: Writing Assessment.....	22
Summary and Conclusions	24
References.....	27
Notes	29
List of Appendixes.....	30

List of Tables

	Page
Table 1. Proportion Correctly Classified and Correlations With True Latent Classes, Fully Crossed Design	10
Table 2. Average Pairwise Agreement and Weighted Kappa.....	11
Table 3. Proportion Correctly Classified and Correlations With True Latent Classes, Balanced Incomplete Block Design.....	15
Table 4. Agreement Proportions and Weighted Kappa.....	16
Table 5. Results for the Second Writing Test Treated as a Fully Crossed Design	23
Table 6. Latent Class Sizes for the Second Writing Test.....	24

List of Figures

	Page
Figure 1. A representation of signal detection theory.....	1
Figure 2. A structural equation-like representation of latent-class signal-detection theory.	3
Figure 3. Distribution of d for the writing test.....	20
Figure 4. Relative criteria locations for 44 raters.	21

Introduction

Essays and other constructed-response (CR) items must be scored by raters. The use of raters to score CR items raises questions about how raters perform the task, an understanding of which in turn is important for the choice of a model of rater behavior. One approach is to view raters as attempting to classify each essay into a latent category, where the latent categories are defined by a scoring rubric. For example, a 1–6 scoring rubric, as used in the SAT[®], GRE[®], and Praxis[™], can be viewed as defining six latent categories of essays, with the task of raters being to determine to which of the six categories each essay belongs. When viewed in this way, the task becomes one of signal detection, in that raters attempt to discriminate between latent categories of items. This suggests the use of a latent-class version of signal-detection theory (SDT) as a model of rater behavior. The approach offers a psychological framework for understanding CR scoring and includes rater parameters that have a clear cognitive basis. Up to this point, latent-class SDT models have been used primarily in medical diagnosis (see DeCarlo, 2002). However, the approach recently has been used in education and in particular as a model of rater behavior in essay scoring (DeCarlo, 2005). The present report examines this approach in more detail.

An immediate benefit of an approach to CR scoring via SDT is that it clarifies that the scores assigned by raters reflect two basic aspects of the task, a perceptual aspect and a decision aspect. This is illustrated in Figure 1.

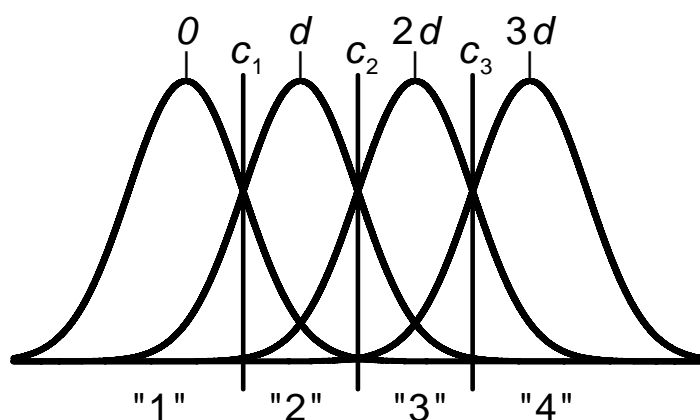


Figure 1. A representation of signal detection theory.

The perceptual aspect of the task refers to the view that, for holistic scoring, raters base their scores in part on their perception of the overall quality of an essay. A basic assumption in SDT is that the perceptions can be viewed as being realizations of a continuous, random variable with a specified probability distribution, such as the normal or logistic (other distributions can be used through the use of different link functions; see DeCarlo, 1998). In particular, it is assumed that there is a probability distribution for each latent class of essay, with a different location for each class, as shown in Figure 1. That is, Figure 1 shows that, for a 1–4 scoring rubric, it is assumed that raters attempt to discriminate between four latent classes of essays. Additionally, the perceptions of the quality of essays from a particular latent class can be represented by a probability distribution, with the result of four distributions, one for each latent class, with different locations.

Of basic interest in SDT is a rater's ability to discriminate between the latent classes, as measured by a discrimination parameter d , which is interpreted in SDT as a measure of the distance between the underlying perceptual distributions; a higher value of d indicates better discrimination and distributions that are further apart. In the version of SDT considered here, referred to as an *equal spacing* SDT model (DeCarlo, 2002, 2005), it is assumed that the raters perceive the latent classes as being equally spaced, and so the distance between perceptual distributions is the same for adjacent distributions, which gives distances of d , $2d$, $3d$, and so on, as shown in Figure 1. Note that the equal spacing is in the raters' perceptions, and not the latent classes, which are only assumed to be ordinal. As shown below, the equal-distance restriction is implemented in the model by scoring the latent classes as 0, 1, 2, and so on.

The decision aspect of the task has to do with a rater's use of the response categories, that is, what a rater considers to be a Category 4 versus a Category 3, for example. In SDT, a rater's category usage is reflected by his or her use of response criteria, c_k , which delineate the K categories, as shown in Figure 1. It is widely recognized that some raters tend to assign high scores (leniency), whereas others tend to assign low scores (strictness); in terms of SDT, this simply reflects the raters' arbitrary use of response criteria, which are lower (i.e., further to the left) for lenient raters and higher for strict raters. The locations of the response criteria also reflect any and all other peculiarities in the rater's response usage, such as avoiding end categories or spacing the categories unequally. Thus, SDT separates perceptual aspects of the

task (a rater’s ability to discriminate between the latent categories), from decision aspects (a rater’s use of response criteria).

Another way to represent the model is shown in Figure 2, which uses a diagram similar to that used in structural equation modeling (e.g., see Kline, 2005). The observed responses, Y_j , consist of ratings, such as from 1–6. As is well known in statistics and psychometrics, models with ordinal responses can be motivated by assuming a continuous underlying variable (e.g., Agresti, 2002), which is shown for each rater j as Ψ_j in Figure 2. In the SDT approach, Ψ_j represents a rater’s perception of the overall quality of an essay, as shown in Figure 1. As noted above, it is assumed that raters arrive at their observed responses by using their perceptions in conjunction with response criteria, shown as c in Figure 2. The arrows from Ψ_j to Y (actually it’s the probability of Y , and not the observed Y , but the diagram is simplified) are curved to indicate that the relation between the mean of Ψ_j and the response probabilities is nonlinear. As noted above and represented in Figure 2, the mean of the Ψ_j distribution is shifted by d_j across the latent classes, which are denoted here as $X^\#$ (i.e., $X^\#$ is used here to denote a latent categorical variable, whereas X^* is commonly used in statistics and econometrics to denote a latent continuous variable).

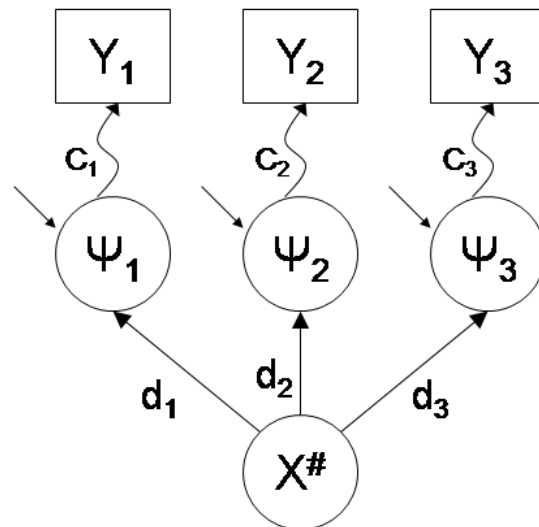


Figure 2. A structural equation-like representation of latent-class signal-detection theory.

The latent-class SDT model can be written as follows. Consider the situation where J raters examine N cases (e.g., essays) and assign a discrete score k to each case, where $1 \leq k \leq K$ and K is the number of response categories. For the equal-distance version of the SDT model, the model is

$$p(Y_j \leq k | X^\# = x^\#) = F(c_{jk} - d_j x^\#) \quad (1)$$

where Y_j is the response variable for rater j (e.g., a 1–6 response), $X^\#$ is a latent categorical variable with values of $x^\#$ from 0 to $K-1$ (note that this particular scoring implements the equal distance restriction), c_{jk} are $K-1$ strictly ordered response criteria for the j th rater and k th response category, d_j is the discrimination parameter for the j th rater, and F is a cumulative distribution function; the logistic cumulative distribution function is used here.

To complete the model, Equation 1 is incorporated into a restricted latent class model, as shown in DeCarlo (2002, 2005). A latent-class model is a model for the probability of the response patterns (k_1, k_2, \dots, k_J) for the J raters and can be written as

$$p(Y_1 = k_1, \dots, Y_J = k_J) = \sum_{x^\#} p(X^\# = x^\#) p(Y_1 = k_1, \dots, Y_J = k_J | X^\# = x^\#), \quad (2)$$

where the summation is over the latent classes $X^\#$. With an assumption of local independence, the second term on the right becomes

$$p(Y_1 = k_1, \dots, Y_J = k_J | X^\# = x^\#) = \prod_j p(Y_j = k_j | X^\# = x^\#), \quad (3)$$

where the product is over the J raters. The latent-class SDT model of Equation 1 is then used for the product on the right in Equation 3 by differencing the cumulative probabilities to get response probabilities, as done for item-response theory models such as the graded-response model (Samejima, 1969). Equations 1 and 3 are then incorporated into Equation 2 to complete the model. Note that, for the version of the latent-class SDT model considered here, which has K ordinal response categories for K latent classes, a minimum of three raters is generally needed in order for the model to be identified; this is not an issue for the large-scale assessments examined here, because many raters are used; yet, it is relevant when pooled data are analyzed (see the section on the second writing test below).

As has been noted previously (DeCarlo, 2002, 2005), from a statistical perspective, the latent-class SDT model is closely related to several other models discussed in psychometrics. For

example, it can be viewed as a discretized version of the graded-response model (see Heinen, 1996) and is also related to the D-factor models of Vermunt and Magidson (2007). The difference is that D-factor models use adjacent category logits, whereas SDT uses cumulative links, because of the motivation in terms of underlying distributions. Indeed, it should be clear from Figure 2 that the latent-class SDT model is a type of factor analysis model, albeit with a discrete factor. Previous research (DeCarlo, 2002, 2005) has compared the latent-class SDT model with D-factor models and item-response theory models.

The present report examines an approach to CR scoring via latent-class SDT. A basic goal is to obtain information, through simulations, about how well the parameters are recovered and how accurate the classifications are. We have little or no information about this at this time. Also investigated are the relation of rater parameters to agreement statistics and the relation of rater discrimination to classification accuracy.

For large-scale assessments, incomplete designs are a necessity, because there are a large number of essays; thus, not all of the raters can score all of the essays. Instead, each essay is graded by a subset of raters, typically 2. This makes it possible for a relatively small number of raters to score a relatively large number of essays. For example, if a balanced, incomplete block (BIB) design is used, with each essay scored by 2 raters, a total of 1,080 essays can be scored by 10 raters, with each rater scoring 216 essays. The present report examines incomplete designs and compares the results to those obtained with complete (fully crossed) designs. Applications to real-world data are also presented.

Simulated Data: Fully Crossed Design

First examined are fully crossed designs, which provide a useful reference point for the incomplete designs examined below. Estimation and classification are examined using a range of values for the rater parameters that is consistent with that found for real-world data.

Methods

The simulated data were generated using SAS software macros written by the author (used in DeCarlo, 2005, and modified as needed for the current studies). Data for 10 raters discriminating between six latent classes by giving one to six responses were simulated. The latent class sizes were chosen to approximate a normal distribution (see Appendix A), which is consistent with the results found for the exams analyzed below. A range of values of d from 2–5

was used, which covers a range of detection from moderate to excellent (for the logistic model) and is consistent with that found for real-world data. For example, for the large-scale assessment examined below, the values of d for 44 raters ranged from 1.8–5.3. One also has to make decisions about the location of the criteria for the different values of d . A general approach, used here, is to locate the criteria 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; some conditions where the criteria are not at the intersection points are also examined below. Relative to d , this means that the first through last criteria are located at $\frac{1}{2}d$, $1\frac{1}{2}d$, $2\frac{1}{2}d$, and so on. That is, it should be obvious that the intersection points for symmetrical distributions are at the point midway between the 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 . A sample size of 1,080 was used for all conditions; the size of 1,080 was used instead of simply 1,000 because the incomplete design examined below is fully balanced for 10 raters with a sample size of 1,080. Each condition consisted of 100 replications.

Data generation consisted of three steps. First, values for the latent variable $X^\#$ (i.e., 0, 1, 2, ..., $K-1$) were generated using a multinomial distribution, where the latent class sizes were used as the probabilities for each latent-class category. Next, the generated values of $X^\#$ were used in Equation 1 along with the population parameters c_{jk} and d_j to get cumulative response probabilities for each rater and response category, using a logistic distribution for F . To generate an observed response, the probabilities were compared to values obtained from a uniform random variable generated on an interval from 0–1. If the value was less than or equal to the probability for the lowest response category, then a response of 1 was assigned; if it was greater than the probability for the lowest category, but less than or equal to the value for the second category, then a response of 2 was assigned, and so on.

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). Some small simulations indicated that Latent Gold tended to have good performance for the models considered here, and so it was used. In particular, a prerelease version of Latent Gold (demo Version 4.5), made available to the author, was used; Version 4.5 allows one to use syntax to specify a wide range of models, including the latent-class SDT model. Latent Gold uses the

expectation-maximization algorithm followed by the Newton-Raphson procedure to obtain maximum likelihood estimates of the parameters (unless Bayes constants are used; see the incomplete simulation below). 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.

One complication that must be recognized is known as *label switching* (McLachlan & Peel, 2000). In the current context, label switching has to do with the coding of the latent categorical variable $X^{\#}$, in that it is arbitrary as to which class is assigned a value of zero. For example, in some cases, the classes will be labeled as 0, 1, ..., $K-1$, and in other cases as $K-1$, $K-2$, ..., 0. The maximized log likelihood has the same value for the switched solution, the main consequence for the latent class SDT model is that the sign of d is reversed as is the order of the latent classes. In addition, when label switching occurs, one has to add $K-1$ times d to the obtained criteria estimates in order to obtain estimates of c . The SAS macro that stripped out and summarized the data checked for label switching and adjusted the computations appropriately.

Results

Rater parameters and latent-class sizes. Appendix A presents, for the rater parameters and latent-class sizes, the population parameters, the mean parameter estimates, the bias, the (absolute) percent bias (the parameter estimate minus the population value, divided by the population value, times 100; the absolute percent bias is shown here because the direction is obvious from the sign of the bias), and the mean squared error (MSE) for fits of the model to the 100 sets of simulated data.

Tables A1–A3 show that estimation is excellent for values of d from 2–4. The bias and MSE for the rater parameters are small, with a percent bias of generally less than 1% for d and less than 2% for c ; percent bias less than 5% is usually viewed as trivial, values from 5–10% as moderate, and over 10% as large (e.g., Flora & Curran, 2004); here, percent bias of 10% or less is viewed as acceptable. The MSE is small, being less than 0.10 for values of d from 2–4, and generally less than 0.50 for the c . Estimation of the latent class sizes is also quite good, with a percent bias of less than 2%.

For $d = 5$, there were problems with convergence, with convergence for only 48 out of the 100 replications; Table A4 is based on the 48 cases where the program converged. It is not surprising to encounter estimation difficulties of this sort with larger values of d because the tables being analyzed become more sparse; that is, there tend to be more zero and low-count cells as d increases. In terms of a two-by-two table, for example, it should be apparent that the entries will concentrate more along the diagonal as d increases (and will all be on the diagonal for perfect discrimination). The fact that estimation problems arise with large values of the slope parameter (i.e., d) is well known in, for example, logistic regression (e.g., Hosmer & Lemeshow, 1989; Rindskopf, 2002).

Table A4 shows that, for the converged cases, estimation of d is very good, with a percent bias of 3% or less and a MSE of less than 0.10. With respect to the response criteria, the percent bias is larger but is generally less than 10%, except for the first criterion, which tends to have larger percent bias, around 12%. However, the MSEs for the response criteria are considerably larger, in the range of 10–20, which indicates that the estimates of c have large variability across replications. Estimation of the latent-class sizes is also problematic, with a small percent bias for the middle classes but large percent bias ($> 20\%$) for the end classes.

Tables A5 and A6 show examples where shifted criteria were used for d s of 2 and 3. In this case, the criteria for 2 of the 10 raters were shifted down from the intersection points locations by 2, the criteria for 2 other raters were shifted down by 1, the criteria for another 2 raters were shifted up by 1, the criteria for another 2 raters were shifted up by 2, and the criteria for the remaining 2 raters were left at the intersection points. Tables A5 and A6 show that shifting the locations of the criteria had little effect on estimation, with the percent bias being below 5% for both the rater parameters and the latent class sizes.

Standard errors. Appendix B presents results for the evaluation of the standard errors of d and the latent-class sizes (the response criteria are not of central interest; there are also some complexities with respect to evaluating the standard errors of c in a simulation because of label switching). Tables B1–B4 show that estimation of the standard errors of d is good, with a percent bias of 10% or less for values of d from 2–4; the standard errors of the latent class sizes also appear to be well recovered. For d of 5, the percent bias is larger, up to about 20% for the standard error of d . The percent bias is also larger for the latent-class sizes, particularly for the first and last classes, where it is around 90%. Table B4 shows that, for a d of 5, the bias is

consistently negative for the standard errors of both d and the latent class sizes, and so the standard errors tend to be underestimated. Tables B5 and B6 show that, for the shifted criteria conditions, the bias for the standard errors is again small for values of d of 2 and 3.

To summarize, Appendixes A and B show that estimation of d and its standard error is quite good for values of d in the range of 2–5, whereas the response criteria and latent-class sizes are accurately estimated for values of d in the range of 2–4 but are less well estimated for a value of d of 5. Overall, the results indicate that if one wishes to assess the performance of the raters, in which case d is of primary interest, then one can obtain a good idea of rater performance, in that d is accurately estimated for the range of values examined here, which is similar to the range found in practice.

Classification. Table 1 shows the classification accuracy (proportion correctly classified) for values of d ranging from 2–5. PC_{pred} is the predicted proportion correctly classified and is obtained from the posterior probabilities (it is basically the average of the maximum posterior probabilities across cases); this value is obtained when the model is fit and is therefore available for both simulated and real-world data. In contrast, PC_{obt} is only available in a simulation and is the obtained proportion of cases that were actually correctly classified in the simulation, where the cases are classified into the class with the maximum posterior probability. Table 1 also shows lambda (see Dayton, 1998; DeCarlo, 2002), both predicted and obtained, and two measures of association with the true latent classes, namely the Pearson correlation r and τ_b . Lambda adjusts the proportion correct using the largest latent class size,

$$\lambda = [PC - \max p(X^\#)]/[1 - \max p(X^\#)], \quad (4)$$

and reflects the improvement in classification accuracy over and above simply classifying all of the cases into the class with the largest size.

Table 1 shows that the predicted proportion correctly classified (PC_{pred}) is .92 for a d of 2 and is over .98 for d s from 3–5. A high value of PC is expected because the accuracy of classification increases with the number of raters, and 10 raters per essay is a relatively large number (as compared below to an incomplete design with only 2 raters per essay). For values of d from 2–4, the obtained proportion correctly classified (PC_{obt}) is close to the predicted value. Table 1 also shows that PC_{pred} overestimates PC_{obt} (the difference is very small in this case), as was also found by DeCarlo (2005). For a d of 5, the obtained PC is considerably smaller (.34);

this likely occurs in part because of poor estimation of the latent class sizes for a d of 5, particularly for the end classes, as was noted above. The problem appears to be that in situations with large d (and small d , though not shown), one or more of the estimated latent-class sizes tends to zero or near zero, and so the classifications tend to be off by one class (or more). This can be shown by computing the proportion correctly classified within one class, which was .99 or larger in every case, including $d = 5$. Note that, even with poor classification accuracy for a d of 5, the Pearson correlation (.93) and τ_{b} (.93) are high, and so the classifications still reflect the order of the latent classes.

Table 1

Proportion Correctly Classified and Correlations With True Latent Classes, Fully Crossed Design

d	PC _{pred}	PC _{obt}	λ_{pred}	λ_{obt}	τ_{b}	r
Intersection-point criteria						
2	.919	.916	.891	.887	.978	.960
3	.986	.985	.981	.979	.996	.993
4	.998	.988	.997	.984	.999	.999
5 ^a	.986	.335	.982	.104	.931	.932
Shifted criteria						
2	.911	.908	.880	.876	.975	.957
3	.981	.981	.975	.974	.995	.991

Note. Ten raters per essay. d is the SDT discrimination parameter; PC_{pred} is the predicted proportion correct; PC_{obt} is the obtained proportion correct.

^aThe $d = 5$ condition includes only 48/100 replications where the program converged.

Table 1 also includes conditions for d s of 2 and 3 where the response criteria were shifted from the intersection points for 8 out of 10 raters (up or down by 1 or 2, see Appendix A). It is interesting to compare classification accuracy in these conditions to that for the intersection point criteria conditions. Table 1 shows that, for a d of 2 and 3, shifting the criteria has little effect on PC, either predicted or obtained, with a reduction in PC for shifted criteria of less than 1%. This

shows that, for a fully crossed design, the criteria locations have little effect on classification accuracy.

Agreement. Table 2 shows the relation between d and agreement statistics. The agreement proportions and weighted kappas are the average, for each replication, across the 45 pairs of rater combinations. That is, they are the average pairwise agreement, which in turn is then averaged over the 100 replications. For weighted kappa, Cicchetti-Allison weights were used, as documented in the FREQ procedure of SAS. A value of Kendall’s coefficient of concordance W (Kendall & Smith, 1939), based on all 10 raters, was also computed for each replication and then averaged over the 100 replications. In contrast to the simple agreement statistic and Kappa, which only consider pairwise relations, W is a measure of agreement across all of the raters (W examines agreement in rankings across the raters, where the rankings are obtained within each rater by using their scores); the fact that it takes into account that there are 10 raters is likely why, as shown next, W tends to be larger than the agreement statistic or kappa.

The upper part of Table 2 is for the conditions with intersection-point criteria locations; the table shows that agreement increases from .27 to .75 as d varies from 1–5; weighted kappa ranges from .28 to .84 and Kendall’s W ranges from .47 to .90. In contrast, Table 1 shows that PC is greater than .90 for values of d from 2–5. For example, for $d = 2$, agreement is .37 (from Table 2), whereas the proportion correctly classified is .92 (from Table 1). Thus, agreement can be low while classification accuracy is high.

Table 2
Average Pairwise Agreement and Weighted Kappa

Criteria	d				
	1	2	3	4	5
Intersection-point criteria					
Agreement	.267	.366	.502	.636	.752
Weighted kappa	.283	.492	.645	.754	.836
Kendall’s W	.469	.717	.838	.899	.936
Shifted criteria					
Agreement	—	.290	.380	.482	—
Weighted kappa	—	.388	.533	.638	—
Kendall’s W	—	.711	.829	.881	—

The lower portion of Table 2, for three conditions with shifted criteria, shows that agreement again increases with d . Most important, Table 2 shows that the percent agreement and kappa are smaller for the shifted criteria as compared to the intersection-point criteria. For example, for a d of 3, agreement decreases by 12%, that is, from 50% for intersection-point criteria to 38% for shifted criteria, which shows that agreement is heavily affected by the criteria locations, as is weighted kappa (an interesting result is that the shifted criteria only appear to have a small effect, less than .02, on Kendall's W). In contrast, Table 1 shows that, for a d of 3 with shifted criteria, PC decreases by less than one half of a percent. This shows that shifting the response criteria has a large effect on agreement but only a small effect on classification accuracy, which suggests limitations of using agreement in practice.

Discussion

The simulations show that, for a fully crossed design with 10 raters, the rater parameters are accurately recovered for a range of discrimination from 2–4, with d being accurately estimated for a range of 2–5. The response criteria and latent-class sizes are well recovered for d s from 2–4 but are less well recovered for values outside of this range. Of course, these results depend in part on the software being used and the particular set of parameters. For example, Latent Gold 4.5 accurately recovered the rater parameters for a range of d from 2–4, but there were problems even within this range with LEM (Vermunt, 1997), mostly a problem of obtaining latent-class size estimates of zero. Latent Gold has options, such as Bayes constants, that can help to ameliorate problems of this sort; this will be examined in future studies (also see the section on incomplete designs below).

An important result is that classification accuracy increases with the raters' level of discrimination, as does agreement. Thus, classification accuracy can be increased by improving raters' discrimination. The levels of agreement, however, are not very informative about classification accuracy, as they are not meant to be. For example, for a d of 3 in a fully crossed design, agreement is 50% and weighted kappa is 64%, which are low to moderate, whereas classification accuracy is 98%, which is excellent. Thus, classification accuracy is high, yet average pairwise agreement is poor to moderate.

Another important result shown in Table 2 is that agreement is heavily affected by the response criteria locations; for example, agreement dropped by over 10% when the criteria were shifted for some of the raters, whereas the effect on classification accuracy was negligible (less

than 1%). These results support the view that the discrimination parameter d is more informative in terms of evaluating the raters' performance and the resulting classification accuracy. For example, the simulation shows that, with 10 raters in a fully crossed design, classification accuracy is high for values of d of 2 or more. The results also show the advantage of using model-based classifications over simply averaging the scores, in that rater differences in response criteria have little effect on the model-based classifications, as shown in Table 1 (more on this below).

Simulated Data: Balanced Incomplete Block (BIB) Design

The above simulations offer basic information about parameter recovery and classification in fully crossed designs. However, in practice, the designs used are incomplete, in that not all the raters rate all of the essays. This section examines the performance of the latent-class SDT model in situations with incomplete data. A BIB design, which is very efficient, is used. This provides information about the effects of incompleteness in a best case scenario and provides an important reference point for future studies using other types of incomplete designs, such as unbalanced designs.

Methods

The data generated for the fully crossed design were used. A SAS macro was used on the data to create missing values according to a BIB design for 10 raters and 1,080 cases, with each rater scoring 2 essays. The incomplete aspect of the design is that each essay is scored by only 2 out of 10 raters, whereas the balanced aspect is that (a) each essay is scored by 2 raters, (b) each rater scores 216 essays, and (c) each rater is paired with every other rater an equal number of times. Each condition consisted of 100 replications.

Data that are missing by design, as in the BIB, are missing completely at random (Rubin, 1976). Latent Gold 4.5 was again used to fit the latent-class SDT model to the incomplete data, and SAS macros were used to strip out and summarize the data. Some pilot simulations showed that estimation problems occurred, and in particular latent-class sizes of zero or large values of d (with large or indeterminate standard errors) were found. Using Bayes constants of one (for the latent and categorical options; see Vermunt & Magison, 2007) appeared to eliminate these problems (use of Bayes constants smooths the parameter estimates and helps to prevent boundary problems). Thus, they were used for the incomplete simulations presented next; note

that, with the use of Bayes constants, one is using posterior-mode estimation, which includes log priors in the log likelihood function. The priors act as a penalty for solutions that are too close to the boundary of the parameter space, with the result that the parameter estimates are smoothed away from the boundary, as noted by Vermunt and Magidson (2005).

Results

Rater parameters and latent-class sizes. Appendix C presents, for the BIB conditions, the mean estimated parameters, the bias, the percent bias, and the MSE. Tables C1–C3 show that, for values of d from 2–4, the percent bias is generally 10% or less for d , but ranges to over 60% for c ; similarly, the MSE is generally less than 0.3 for d but is larger for c . Some patterns also appear across the tables—the percent bias tends to be largest for the first criterion, whereas the MSE tends to be largest for the last criterion. Tables C1–C3 also show that the bias for the latent-class sizes is generally less than 10% for latent classes of 2–5 (at least for d s of 2 and 3) but is large for the first and last classes (which have the smallest sizes), with a percent bias of up to 80%. Table C4 shows a condition with shifted criteria for a d of 3; the percent bias for d is again generally 10% or less. However, the percent bias for the response criteria and latent-class sizes tends to be larger, and so shifting the criteria led to somewhat larger bias in the criteria and size estimates.

Compared to the fully crossed design, the bias and MSE are larger for the BIB design, as expected, because of the large number of missing values. Overall, however, Appendix C shows that estimation of d is good for values of d from 2–4. Estimation of the latent-class sizes is also adequate; however, the smallest latent-class sizes tend to be overestimated.

Standard errors. Appendix D presents, for the BIB conditions, tables that examine the performance of the standard errors for d and the latent class sizes. The bias is generally small to moderate, 10–15% or less, which indicates that the standard errors are reasonably well estimated. A comparison of Appendix D to Appendix B shows that the standard errors are larger for the BIB conditions than for the fully crossed conditions; for example, the standard errors are about 0.10 for a d of 3 in a fully crossed design but are about 0.45 for a d of 3 in a BIB design. This reflects the fact that less information is available in the BIB design with two raters per essay, as compared to a fully crossed design.

Classification. Table 3 presents the proportion correctly classified for the BIB design. As for the fully crossed design, the proportion correctly classified increases as d varies from 2–4,

with the obtained PC ranging from about 52–80%; the Pearson correlation and τ_b also increase, from about .80 to over .90. The condition with shifted criteria shows that the criteria only have a small effect on the PC and association measures. For example, for a d of 3, PC_{obt} is 69.6% for intersection-point criteria and 68.2% for shifted criteria, which is a difference of less than 2%. Thus, shifting the criteria has little effect on classification accuracy (for model-based classifications, see below). Similarly, τ_b and r are both around .90 and differ across intersection-point and shifted criteria by only about .01. Table 3 also shows that PC_{pred} overestimates PC_{obt} , as found above for the fully crossed design and by DeCarlo (2005); for example, for a d of 3, PC_{pred} is .74, whereas PC_{obt} is .70.

Table 3
Proportion Correctly Classified and Correlations With True Latent Classes, Balanced Incomplete Block Design

d	PC_{pred}	PC_{obt}	PC_{av}	λ	λ_{obt}	τ_b	r
Intersection-point criteria							
2	.623	.525	.575	.478	.360	.792	.866
3	.744	.699	.708	.656	.594	.871	.926
4	.843	.799	.803	.788	.729	.911	.951
Shifted criteria							
3	.721	.682	.617	.627	.572	.871	.925
4	.812	.801	.707	.747	.731	.912	.950

Note. Two raters per essay.

A comparison of Table 3 and Table 1 shows the effects on classification of using a BIB design over a fully crossed design. For example, for a d of 3, PC_{obt} is about 70% for the BIB with 2 raters per essay and is 98% for the fully crossed design with 10 raters per essay. Thus, there is a large effect of the number of raters per essay on classification accuracy, as expected. Table 3 also shows the proportion correctly classified using the average of the two scores, PC_{av} . This is of interest because the simple average is commonly used in practice. To assess classification accuracy for cases where the average had in-between values, such as 2.5, the scores were rounded both up and down; the results differed by less than .003, and the larger values of PC_{av}

are reported here. Table 3 shows that, as before, classification accuracy increases with d . However, in contrast to the model-based classifications, PC_{av} is considerably lower for the shifted criteria. For example, for a d of 3, PC_{av} was 70.8% for intersection-point criteria locations but only 61.7% for shifted criteria locations, a decrease of almost 10%. Thus, the results for PC_{av} show that the proportion correctly classified drops considerably for average scores if there are differences in the response criteria locations across raters, as in the shifted criteria condition. In contrast, criteria shifts have little effect on model-based classifications (PC_{obt}). Thus, an advantage of the model-based approach is that classification accuracy is not affected by idiosyncrasies in raters' response usage, whereas it is affected if average scores are used.

Agreement. Table 4 shows agreement statistics, the proportion of exact agreement, and weighted kappa for the BIB simulation (note that Kendall's W cannot be computed for the BIB design because of the missing values). Table 4 shows agreement between pairs of raters averaged over the 100 replications. A comparison of Table 4 to Table 2 shows only trivial differences, which is as expected. The only difference is that the fully crossed design first averages over the 45 rater pairs for each essay, and then over the 100 replications, whereas for the BIB design, there is only 1 rater pair per essay, and so the averaging is only over the 100 replications. Table 4 shows that agreement increases with the discrimination parameter d and that agreement is heavily affected by the response criteria, decreasing, for example, by 12% (50% to 38%) for a d of 3 and by 16% (64% to 48%) for a d of 4; the weighted kappas are also smaller.

Table 4
Agreement Proportions and Weighted Kappa

Criteria	d				
	1	2	3	4	5
Intersection-point criteria					
Agreement	.267	.368	.505	.636	.751
Weighted kappa	.268	.488	.643	.748	.832
Shifted criteria					
Agreement	—	.292	.381	.483	—
Weighted kappa	—	.422	.563	.655	—

Table 4 suggests that agreement statistics are informative about pairwise agreement, as they should be, but not about classification accuracy. For example, for a d of 3, Table 4 shows that agreement is only about 50% for intersection-point criteria and 38% for shifted criteria. However, Table 3 shows that classification accuracy is about 70% in both cases, and the measures of association are about .90. Thus, low agreement does not necessarily mean that the raters are performing poorly; the discrimination parameter is more informative in this regard, in that it provides information about how well the raters discriminate the latent classes and about how accurately the items are classified.

Discussion

The simulations provide basic information about various aspects of an approach to essay grading via SDT. First, the simulations show that estimation of the rater parameters, particularly the discrimination parameter, is good for both complete and incomplete designs (at least with the use of Bayes constants for incomplete designs), for the range of values examined here (d from 2–5), which are comparable to those found in practice (see below). It should be noted that there tend to be estimation problems in incomplete designs, but the use of Bayes constants (of one) and posterior-mode estimation gave good results. Estimation of the latent-class sizes also appears to be adequate across values of d from 2–4 (again with the use of Bayes constants for incomplete designs), at least for the normal-like distribution of latent-class sizes examined here (as found for real-world data below). Larger values of d , such as 5, can lead to convergence problems and poor estimation of the response criteria and latent-class sizes. Classification is also adversely affected; d , however, still appears to be adequately estimated, which means rater performance can still be evaluated. An argument can be made for the use of average ratings in the situation where estimation problems arise because of large values of d , in that classification accuracy should be high (for average ratings) even with differences in the criteria locations; the use of larger values of Bayes constants in that situation also can be explored.

Tables 1–4 provide useful information about expected performance in a signal-detection task as a function of rater discrimination. For a fully crossed design with 10 raters, Table 1 shows that classification accuracy is excellent (>90%) even for the lowest value of discrimination examined; this occurs because there are many raters per essay. Measures of agreement, such as the percent agreement and weighted kappa, tend to be considerably smaller. Of greater practical interest is that, for a BIB design with 2 raters per essay, Table 3 shows that

70% or more of the essays are correctly classified for values of d of 3 or larger, whereas agreement is 50% or more and weighted kappa is 64%. Table 3 provides guidelines as to the levels of classification accuracy and agreement associated with a particular level of rater performance.

In light of the above, requiring a specified level of agreement is shown to be a conservative approach. For example, suppose a minimum of 70% agreement is required. Table 4 shows that this is associated with a level of discrimination of greater than 4, which is quite good, whereas Table 3 suggests that a d of over 4 is associated with an obtained classification accuracy of over 80%. Thus, a specified level of agreement is a strict criterion, which is fine as long as this is understood. In some situations, it might be more useful simply to consider expected classification rates rather than agreement levels. For example, if classification accuracy is desired to be 70% or greater, then requiring (an average) rater discrimination of 3 or larger (for the logistic model) seems quite reasonable. The above also shows that agreement is heavily affected by the response criteria locations, as expected, and so agreement can be misleading with respect to how good classification is. Estimates of d and c , on the other hand, provide important information about classification accuracy and whether raters are performing adequately or not.

ETS Data

This section applies the latent-class signal-detection model to the writing section of several ETS datasets. This application provides information about parameter values found in practice.

Example 1: Writing Assessment

The data examined here are scores given to essays written by 10,647 examinees as part of a large-scale writing assessment (note that 17 essays were dropped because of one or more missing scores and 4 more were dropped because 2 raters scored only 2 essays each). The essays were scored by 44 raters, who used a 1–6 response scale (a response of zero is also possible but was not used for the subset of essays examined here). Each essay was scored by 2 raters, with the 44 raters each scoring anywhere from 33–1404 essays.

Differences in response category usage were noted across the 44 raters. For example, 9 raters used all of the response categories, Categories 1–6, whereas 27 raters used Categories 2–6; 5 raters used Categories 2–5; and 1 rater each used Categories 1–5, 3–6, or 3–5. In some cases,

the restricted response range likely occurred because of a small sample size, for example, the rater who only used Categories 3–5 scored just 33 essays, and the rater who used Categories 3–6 scored 137 essays. Yet, this was not always the case—the rater who used Categories 1–5 scored 897 essays. From the signal-detection perspective, the differences in response category usage reflect individual differences in the response criteria locations. Lack of response category usage has been discussed in the measurement literature as an issue of *null categories* (see Wilson & Masters, 1993); for the analysis presented here, the response categories were downcoded (i.e., 2–6 becomes 1–5), which has no effect on the estimates of d (and c_2 becomes c_1 , etc.). The effect, if any, of downcoding on classification accuracy in the context of the latent-class SDT model is being examined in current research.

The typical approach to arrive at a score for each essay is simply to add or average the two scores. This approach essentially treats the pool of raters for the first and second scores as being equivalent (for each score); that is, the data are pooled across raters and so are treated as being from a fully crossed design (i.e., there are two scores, collapsed across raters, for all essays). On the other hand, fitting the latent-class SDT model to the data in incomplete form, where d_j and c_{jk} are treated as rater-specific fixed effects, allows examination of any differences across the 44 raters who actually provided the scores.

Figure 3 shows a histogram of the estimates of d_j for the 44 raters, obtained by fitting the model to the data in incomplete form (again using Bayes constants of one). The estimates of d have a mean of 3.5 with a range of 1.9–5.4 and a standard deviation of 0.9. The estimates are approximately normally distributed, with a (Fisher’s g) skew of 0.05 (SE of 0.36) and kurtosis of -0.43 (SE of 0.70). Thus, there appear to be differences in discrimination across the 44 raters.

Figure 4 presents a plot of the relative criteria (DeCarlo, 2005) for the 44 raters who scored the test. The relative criteria are

$$\text{rel } c_{jk} = c_{jk} / (K-1) d_j, \tag{5}$$

where K is the number of latent classes and the estimates obtained for c_{jk} and d_j are used in the above. Equation 5 equalizes the location of the highest and lowest distributions across raters; for example, the lowest distribution is set at 0 and the highest at 1 in Figure 4. The horizontal lines show the intersection-point locations of the five response criteria, that is, the crossover points of the symmetric underlying distributions. Thus, Figure 4 compactly shows the locations of each

rater's response criteria, relative to the intersection points of the six underlying distributions, which is informative about the raters' use of the response categories. For example, for Rater 1, Figure 4 shows that the first two criteria are below the intersection point of the first and second distribution (and so Rater 1 is somewhat conservative with respect to giving responses of Categories 1 and 2), whereas the third, fourth, and fifth criteria are at the second, fourth, and fifth intersection points.

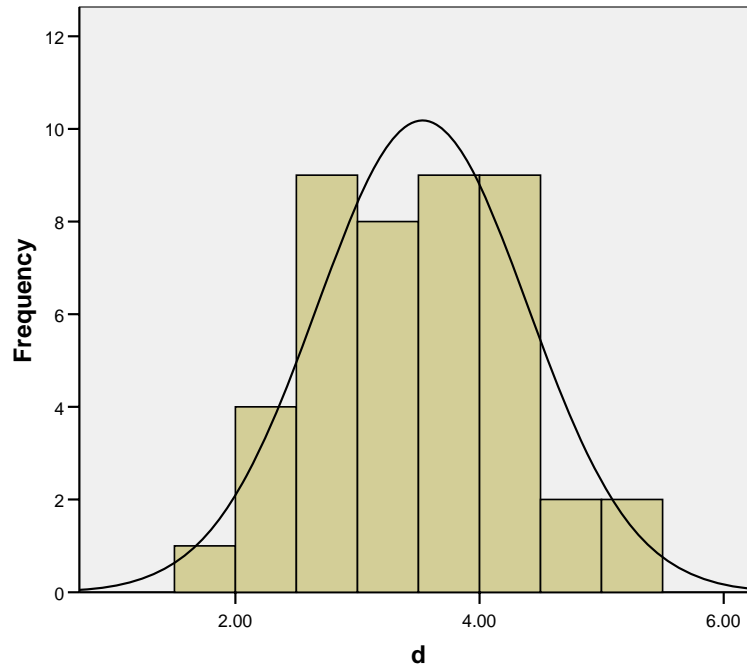


Figure 3. Distribution of d for the writing test.

Overall, Figure 4 shows differences between raters in response category usage, both in terms of criteria locations and the number of categories used (as noted above). An interesting result that is apparent in the figure is that the raters tend to be conservative with respect to their use of the lower response categories, in that the first and sometimes second criteria tend to be well below the intersection-point locations (keeping in mind that in situations with four instead of five criteria, the first response was usually 2, and so the first criterion shown is actually c_2 , not c_1). Figure 4 also shows that, in general, there are criteria that tend to lie on the second and fourth intersection points, whereas the first and last criteria tend to be below and above the first and last intersection points, respectively. Thus, Figure 4 shows that, according to the SDT model, raters appear to be conservative with respect to using categories such as 1 and 6, in that the

corresponding criteria tend to be well above or below the intersection point (which means that those responses are used less frequently).²

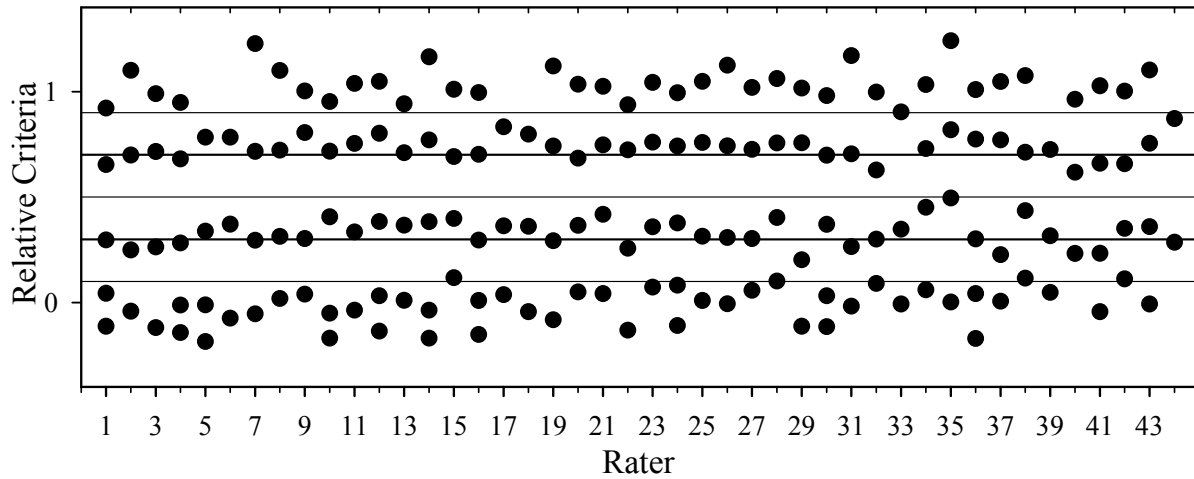


Figure 4. Relative criteria locations for 44 raters.

The estimates of the latent-class sizes (with standard errors in parentheses) are .02 (.002), .14 (.01), .19 (.01), .46 (.02), .17 (.01), and .02 (.006). Thus, latent-class sizes of 1 and 6 are small, with the largest latent-class size being 4. Note that although latent classes of 1 and 6 have small sizes, the standard errors are small (because of the large sample size). Also note that the latent classes are approximately normally distributed, with some small, negative skew. The predicted proportion correctly classified, PC_{pred} , is .74, which is consistent with values found in the simulation for d s of 3–4. For example, PC_{obt} in Table 3 suggests that 70–80% of the cases might be classified correctly. A simulation using the obtained parameter estimates can be conducted to gain more detailed information about likely classification accuracy.

In sum, an application of latent-class SDT to a writing assessment offers new and interesting results. First, it suggests that rater performance for the test is very good (average d of 3.5), with some differences across raters in discrimination and response criteria. This, along with PC_{pred} , suggests that classification accuracy is likely 70% or more. Second, the estimates of the response criteria suggest that the raters are conservative with respect to use of the lowest response categories and the highest category; this has never, to my knowledge, been noted before and merits further attention. It could occur, for example, because of the scoring rubric or other instructions given to the raters (e.g., that might lead them to believe that the lowest and highest

categories occur less frequently than they actually do). It is also interesting to note that this was not found for the analysis of data from the next test examined. Third, the estimated latent-class sizes suggest that most of the essays are classified into Category 4, with only about 2% classified into Categories 1 and 6; the distribution of the latent classes is also close to that for a normal distribution (with a small negative skew).

Example 2: Writing Assessment

For the second large-scale writing test examined, the scoring rubric consisted of categories from 1–5 (there is also a 0 category for essays that have, for example, little or no text or are not on the assigned topic; in this example, only 124 essays out of over 42,000 received scores of zero and so were not included in the analysis). This section presents an analysis of data from 42,608 examinees (after dropping 69 cases because of missing values and 124 more cases with zeroes), with 2 essays per examinee. Each essay had two scores (from various raters, the pooled data are analyzed here), with a third score given by an adjudicator when the two scores differed by 2 or more. For the first writing task, 3.9% of the essays had third (adjudicated) scores, whereas 2.6% of the essays for the second task had third scores. 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 long as a sufficient number are available, as is typically the case for large-scale assessments. Latent-class SDT models with five latent classes are fit to the data. In this analysis, 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 results for pooled data, as are commonly analyzed. Also, the data are being used in that form in current research on a hierarchical rater model.

Table 5 presents results for the writing task. A latent-class, logistic, SDT model was fit to the data using all three scores, where the third score was only available for the adjudicated cases (3.9%). Table 5 shows that the estimates of d are similar across the three scores, being 3.8, 3.9, and 3.4, respectively. Note that the standard errors are larger for Score 3 because 96.1% of the scores were missing (there were still 1,661 third scores available). It is interesting to note that, despite the high degree of missing scores, the estimates of c_{jk} and d_j for the third score are close

to those obtained for the other two scores, which indicates that the behavior of the adjudicators was similar to that of the other raters.

Table 5

Results for the Second Writing Test Treated as a Fully Crossed Design

Parameter	Score 1		Score 2		Score 3	
	Estimate	SE	Estimate	SE	Estimate	SE
d	3.77	0.05	3.88	0.06	3.39	0.20
c_1	1.73	0.07	1.84	0.07	1.26	0.29
c_2	5.47	0.11	5.59	0.12	4.43	0.33
c_3	8.98	0.12	9.22	0.14	7.86	0.45
c_4	12.35	0.16	12.67	0.17	11.03	0.59

Table 5 shows that discrimination is again in the range of 3–4 (keeping in mind that five latent classes were used for this example, whereas six were used for the first example). It is also apparent that discrimination for the adjudicated score is about the same in magnitude as for the other two scores. The criteria estimates are also similar in magnitude across the three scores, with the criteria for Score 3 being slightly to the left of the other two scores, which indicates that the criteria for the third score were slightly more liberal (i.e., higher responses were used) than those for the other two scores. It is also interesting to note that, in all cases, the response criteria estimates in Table 5 are close to their intersection-point locations. For example, using the parameter estimate for d for the first score (3.8), intersection-point criteria locations will be at 1.9, 5.7, 9.5, and 13.3, which are close in value to the estimates shown in Table 5 (1.7, 5.5, 9.0, and 12.3). Thus, it appears that, for pooled data, the response criteria tend to be located close to the intersection points of the underlying logistic distributions, which to my knowledge has not been noted before.

With respect to the latent-class sizes (Table 6), the end categories are smallest, being .12 and .08 for Categories 1 and 5, respectively, and are larger than that found for the first test examined above. The largest latent-class size is for Category 3, followed by Category 4; note that the latent-class sizes are also approximately normally distributed. The estimate of PC_{pred} is .84. Just considering Scores 1 and 2, the agreement proportion is .58 and weighted kappa is .67.

Table 6***Latent Class Sizes for the Second Writing Test***

Category	p_1	p_2	p_3	p_4	p_5
Estimate	.12	.17	.34	.29	.08
SE	< .01	< .01	< .01	< .01	< .01

In sum, the results for the writing sections of several ETS tests showed consistent results with respect to rater parameters—discrimination appears to be in the range of 3–4 for a logistic SDT model (with five or six latent classes). Note that the SDT approach focuses attention on effect sizes for the raters, namely the magnitude of discrimination as measured by d , which is informative about rater performance and classification accuracy. The response criteria for raters in the second test appeared to be close to their intersection-point locations. The results also showed that the distribution of latent classes is slightly asymmetric (negatively skewed), but close to (discrete) normal.

Summary and Conclusions

The present report lays out the scope and potential of an approach to essay grading via a latent-class extension of SDT. The simulations provide basic information about parameter estimation and about the relation between discrimination, classification, and agreement; the real-world analyses provide information about values of the rater parameters and latent-class sizes that are found in practice.

The approach via SDT also informs several issues. For example, why is agreement of interest in scoring tasks such as essay grading? The answer is that high agreement suggests that the raters are detecting a construct, such as the latent classes defined in the scoring rubric. However, as shown here, agreement is at most only an indirect indicator of rater performance, in that it depends not only on the raters' ability to discriminate between the latent classes, but also on their use of response criteria. For example, as shown here, the raters' discrimination can be quite high and classification accuracy can be high, yet agreement can be quite low (e.g., if the criteria differ across raters). Thus, agreement only provides an indirect assessment of what is really of interest, which is how well the raters classify the essays. Here it is noted that estimates of the raters' parameters, particularly d , are informative about rater performance and

classification accuracy. The recommendation is to supplement agreement statistics with estimates of the rater parameters d and c ; at the least, this might provide information as to *why* raters disagree (see DeCarlo, 2002, for another example).

An implication of the above for rater training, which was noted earlier (DeCarlo, 2002), is that it is probably more effective to monitor raters in terms of their discrimination parameter than by their level of agreement. Use of agreement might be unnecessarily strict. For example, it might suggest rater retraining or elimination in situations where it is not necessary, in that the rater discriminates adequately but has different response criteria. The estimate of d will provide valuable information about rater performance in that case, and the use of model-based classifications will likely have benefits as well. Thus, the approach via SDT might have cost benefits with respect to reducing unnecessary elimination or retraining of raters. Given that the above simulations showed that the discrimination parameter was accurately estimated for the range of values that appear to be found in practice, the latent-class SDT model should be a useful tool for monitoring rater performance.

It also was shown that adjudicated cases can be included in the analysis. For an analysis of essays pooled across raters, discrimination was about the same for adjudicated essays as it was for the other essays. This finding makes sense for ETS tests, because the adjudicators are sometimes chosen from the general pool of raters, and so they should have similar discrimination to other raters. However, in some cases in the psychometrics literature, adjudicators are assumed to be experts; note that the latent-class SDT model allows assessment of expertise by using the data of all of the raters (or scores), as done above, and comparing the parameters across raters (experts should show large values of d and appropriate criteria locations). In the same way, the latent-class SDT model allows one to evaluate presumed gold standards used in medical and other research, as noted earlier (DeCarlo, 2002).

Similarities across tests used by ETS were also found. For example, rater similarities were found across the essays used in the writing sections of the first and second tests, in that discrimination tended to be in the range of 3–4 for the logistic model. Note that this was also found for an analysis of a large sample of SAT essays (where d s of 3.5 and 3.1 were found; DeCarlo & Kim, 2008). It is also interesting to note that, for an analysis of a small sample (125) of college data scored by nonexperts (graduate students), the average value of d was 2.1 (for the logistic model; see DeCarlo, 2005), which is smaller than that found for the professional raters

used in the large-scale assessments examined here (where values of d from 3–4 were found). This difference could reflect differences in the raters' experience or differences in the quality of the essay item or the scoring rubric; further research on this is needed. In any case, there are clearly interesting patterns of results with respect to d , found both here and in previous studies. The latent-class SDT model offers a new perspective with which to examine CR data and suggests new directions for future research.

References

- Agresti, A. (2002). *Categorical data analysis* (2nd ed.). Hoboken, NJ: Wiley.
- Dayton, C. M. (1998). *Latent class scaling analysis*. Thousand Oaks, CA: Sage.
- 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., & 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.
- Flora, D. B., & Curran, P. J. (2004). An empirical evaluation of alternative methods of estimation for confirmatory factor analysis with ordinal data. *Psychological Methods*, 9, 466-491.
- Heinen, T. (1996). *Latent class and discrete latent trait models: Similarities and differences*. Thousand Oaks, CA: Sage.
- Hosmer, D. W., & Lemeshow, S. (1989). *Applied logistic regression*. New York: Wiley.
- Kendall, M. G., & Smith, B. B. (1939). The problem of m rankings. *The Annals of Mathematical Statistics*, 10, 275-287.
- Kline, R. B. (2005). *Principles and practice of structural equation modeling* (2nd ed.). New York: Guilford Press.
- McLachlan, G., & Peel, D. (2000). *Finite mixture models*. New York: John Wiley & Sons.
- Muthén, L. K., & Muthén, B. O. (2007). *Mplus user's guide* (5th ed.). Los Angeles: Muthén & Muthén.
- Rindskopf, D. (2002). Infinite parameter estimates in logistic regression: Opportunities, not problems. *Journal of Educational and Behavioral Statistics*, 27, 147-161.
- Rubin, D. B. (1976). Inference and missing data. *Biometrika*, 63, 581-592.
- Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores *Psychometrika Monograph No. 17*.

- Vermunt, J. K. (1997). LEM: A general program for the analysis of categorical data [Computer software and manual]. Retrieved from the Tilburg University Web site: <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.
- Vermunt, J. K., & Magidson, J. (2007). *LG-Syntax user's guide: Manual for Latent Gold 4.5 syntax module*. Belmont, MA: Statistical Innovations.
- Wickens, T. D. (2002). *Elementary signal detection theory*. New York: Oxford University Press.
- Wilson, M., & Masters, G. N. (1993). The partial credit model and null categories. *Psychometrika*, 58, 87-99.

Notes

¹ Lawrence DeCarlo wrote this paper while under contract to ETS.

² It is interesting to note that the deviations from the intersection-point criteria locations appear to be in the direction of where optimal criteria would be located (see Wickens, 2002). Because Classes 1 and 6 have small sizes, the optimal criteria should be further to the left for c_1 and further to the right for c_5 , which is exactly what was found. The estimates are, however, further to the left or right than the optimal locations.

List of Appendixes

	Page
A – Parameter Estimates, Bias, Percent Bias, and Mean Squared Error for the Fully Crossed Conditions With 10 Raters.....	31
B – Evaluation of the Estimated Standard Errors for d and the Latent-Class Sizes.....	43
C – Parameter Estimates, Bias, Percentage Bias, and Mean Squared Error for the Balanced Incomplete Block (BIB) Design	46
D – Evaluation of the Estimated Standard Errors for d and the Latent Class Sizes, Balanced Incomplete Block (BIB) Design	54

Appendix A

Parameter Estimates, Bias, Percent Bias, and Mean Squared Error for the Fully Crossed Conditions With 10 Raters

Table A1

Intersection Point Criteria, Fully Crossed, $d = 2$, $N = 1,080$

Parameter	Value	Estimate	Bias	% Bias	MSE
Rater parameters					
d_1	2	2.022	0.022	1.100	0.006
d_2	2	2.003	0.003	0.150	0.008
d_3	2	1.992	-0.008	0.400	0.006
d_4	2	2.010	0.010	0.500	0.005
d_5	2	2.000	0.000	0.000	0.005
d_6	2	2.000	0.000	0.000	0.006
d_7	2	2.000	0.000	0.000	0.006
d_8	2	2.005	0.005	0.250	0.007
d_9	2	2.003	0.003	0.150	0.005
d_{10}	2	1.999	-0.001	0.050	0.006
c_{11}	1	1.006	0.006	0.600	0.031
c_{12}	3	3.036	0.036	1.200	0.030
c_{13}	5	5.049	0.049	0.980	0.057
c_{14}	7	7.065	0.065	0.928	0.075
c_{15}	9	9.088	0.088	0.977	0.103
c_{21}	1	1.009	0.009	0.900	0.031
c_{22}	3	3.014	0.014	0.466	0.033
c_{23}	5	5.022	0.022	0.440	0.046
c_{24}	7	7.020	0.020	0.285	0.088
c_{25}	9	9.041	0.041	0.455	0.131
c_{31}	1	0.981	-0.019	1.900	0.023
c_{32}	3	2.983	-0.016	0.533	0.031
c_{33}	5	4.980	-0.020	0.400	0.041
c_{34}	7	6.977	-0.023	0.328	0.063
c_{35}	9	8.981	-0.019	0.211	0.088
c_{41}	1	1.001	0.001	0.100	0.025
c_{42}	3	3.014	0.014	0.466	0.027
c_{43}	5	5.034	0.034	0.680	0.038
c_{44}	7	7.025	0.025	0.357	0.059
c_{45}	9	9.034	0.034	0.377	0.086
c_{51}	1	0.992	-0.008	0.800	0.022
c_{52}	3	3.013	0.013	0.433	0.028
c_{53}	5	5.004	0.004	0.080	0.045

(Table continues)

Table A1 (continued)

Parameter	Value	Estimate	Bias	% Bias	MSE
Rater parameters					
<i>c</i> ₅₄	7	7.014	0.014	0.200	0.071
<i>c</i> ₅₅	9	9.038	0.038	0.422	0.092
<i>c</i> ₆₁	1	0.998	-0.012	1.200	0.023
<i>c</i> ₆₂	3	3.007	0.007	0.233	0.033
<i>c</i> ₆₃	5	4.994	-0.006	0.120	0.048
<i>c</i> ₆₄	7	7.013	0.013	0.185	0.072
<i>c</i> ₆₅	9	9.009	0.009	0.100	0.093
<i>c</i> ₇₁	1	0.987	-0.013	1.300	0.023
<i>c</i> ₇₂	3	3.013	0.013	0.433	0.029
<i>c</i> ₇₃	5	5.011	0.011	0.220	0.052
<i>c</i> ₇₄	7	7.016	0.016	0.280	0.083
<i>c</i> ₇₅	9	9.023	0.023	0.255	0.113
<i>c</i> ₈₁	1	0.999	-0.001	0.100	0.030
<i>c</i> ₈₂	3	2.990	-0.010	0.333	0.033
<i>c</i> ₈₃	5	4.981	-0.019	0.380	0.045
<i>c</i> ₈₄	7	6.979	-0.021	0.300	0.072
<i>c</i> ₈₅	9	8.983	-0.017	0.188	0.099
<i>c</i> ₉₁	1	0.998	-0.002	0.200	0.025
<i>c</i> ₉₂	3	3.010	0.010	0.333	0.032
<i>c</i> ₉₃	5	5.013	0.013	0.260	0.042
<i>c</i> ₉₄	7	7.002	0.002	0.028	0.071
<i>c</i> ₉₅	9	8.995	-0.005	0.055	0.081
<i>c</i> ₁₀₁	1	1.007	0.007	0.700	0.030
<i>c</i> ₁₀₂	3	2.992	-0.008	0.266	0.029
<i>c</i> ₁₀₃	5	4.998	-0.002	0.040	0.050
<i>c</i> ₁₀₄	7	7.010	0.010	0.142	0.075
<i>c</i> ₁₀₅	9	9.024	0.024	0.266	0.091
Latent-class sizes					
Class 1	0.080	0.080	0.000	0.000	
Class 2	0.170	0.169	0.001	0.589	
Class 3	0.250	0.251	-0.001	0.400	
Class 4	0.250	0.250	0.000	0.000	
Class 5	0.170	0.170	0.000	0.000	
Class 6	0.080	0.081	-0.001	1.250	

