Journal of Educational Measurement
Spring 2005, Vol. 42, No. 1, pp. 53–76

A Model of Rater Behavior in Essay Grading Based on Signal Detection Theory

Lawrence T. DeCarlo
Teachers College, Columbia University

An approach to essay grading based on signal detection theory (SDT) is presented. SDT offers a basis for understanding rater behavior with respect to the scoring of construct responses, in that it provides a theory of psychological processes underlying the raters' behavior. The approach also provides measures of the precision of the raters and the accuracy of classifications. An application of latent class SDT to essay grading is detailed, and similarities to and differences from item response theory (IRT) are noted. The validity and utility of classifications obtained from the SDT model and scores obtained from IRT models are compared. Validity coefficients were found to be about equal in magnitude across SDT and IRT models. Results from a simulation study of a 5-class SDT model with eight raters are also presented.

Essays are widely used in performance assessments, such as entrance examinations, placement examinations, and in college courses. The use of essays in assessment necessitates the use of raters to score the essays, which in turn raises issues of rater training, rater reliability, scoring, and so on. These issues have been discussed in the various research areas that have been concerned with essay grading, such as the measurement literature, the writing assessment literature, the language testing literature, and the medical literature. Unfortunately, there has often not been a consensus with regard to various issues. For example, there have been arguments both for and against emphasizing agreement in rater training (see Weigle, 1998, and the references therein). Reliability has also been approached in different ways, such as through classical test theory, generalizability theory, or via the Rasch model (e.g., see MacMillan, 2000). Issues with respect to scoring have also been recurrent (Clauser, 2000). In addition, different models have been used to analyze and score essays, such as the Rasch model (e.g., Congdon & McQueen, 2000; Engelhard, 1994, 1996; Linacre, 1989; Weigle, 1998), the generalized partial credit model, the graded response model, multilevel or Bayesian extensions of these models (e.g., Donoghue & Hombo, 2000; Johnson, 1996; Johnson & Albert, 1999; Patz, Junker, Johnson, & Mariano, 2002; Verhelst & Verstralen, 2001), or models based on generalizability theory (e.g., Longford, 1994, 1995).

With respect to addressing issues related to the use of raters in performance assessment, it would be of great help if we had an understanding of what raters actually do when rating essays. Currently used approaches, such as item response theory (IRT) or generalizability theory, are essentially silent on this point. For example, Hambleton, Swaminathan, and Rogers (1991) noted that "Much of the IRT research to date has emphasized the use of mathematical models that provide little in the way of psychological interpretations of examinee item and test performance" (p. 154). Similarly, Goldstein and Wood (1989) noted, with respect to IRT, "But what sort of theory is it? As the title of Lord and Novick's (1968) book made clear, the theory is statistical, not psychological" (p. 139). Nevertheless, the importance of understanding psychological processes involved in performance assessment has been widely recognized. For example, in the psychometrics literature, van der Linden and Hambleton (1997) noted that there is an increased interest among IRT researchers in models of cognitive processes and, with respect to measurement models, "If, in addition, such models also have a spin off to psychological theory, so much the better. They may help to better integrate measurement and substantive research—areas that to date have lived too apart from each other" (p. 22). In the language testing literature, Cumming (1990) noted that "*Direct* validation of the judgement processes used in these assessment methods has not been possible because there is insufficient knowledge about the decision making or criteria which raters or teachers actually use to perform such evaluations" (p. 32). In the writing assessment literature, Barritt, Stock, and Clark (1986) simply asked "What do we, as teachers who read to evaluate, do when we judge student essays holistically?" (p. 316).

The purpose of this article is to show that signal detection theory (SDT), which has been widely and successfully used in psychology and medicine (e.g., see Gescheider, 1997; Green & Swets, 1988; Macmillan & Creelman, 1991; Swets, 1996), provides a psychological theory about what raters do when they score essays. In particular, a latent class extension of the standard SDT model is applied to essay grading; the model and examples of applications to psychological studies and medical diagnosis have been given in DeCarlo (2002a), where the relevance of the approach to essay grading was also noted, as were implications for rater training. Here it is shown that, in addition to providing a model of psychological processes involved in essay grading, SDT also provides answers or at least guidance with respect to some of the issues noted above. The approach is illustrated with a set of real-world data: essays obtained as part of a final exam in a college course. The latent class SDT model is compared and contrasted to IRT models that are commonly used to score essays. Different approaches to classifying and scoring are compared, and evidence as to the validity of the resulting classifications and scores is obtained.

It should be noted that the present approach differs somewhat from those currently in use with respect to the conceptualization of the construct being assessed. In particular, the view here is that many of the constructs considered in psychology and education can usefully be thought of as consisting of ordered categories, in that they are not formulated precisely enough or richly enough to be considered as quantitative. Although constructs are usually treated in psychometrics as being continuous and quantitative, this is nevertheless an assumption. In applications of IRT models or structural equation models, for example, it is not shown that the construct is quantitative, it is assumed. Michell (1997) has criticized psychometrics (and psychology) for just this, that is, for assuming that constructs are continuous and quantitative without presenting evidence for this assumption. In addition, it has long been recognized that there can be difficulties in distinguishing between latent class and latent trait models. For example, Bartholomew and Knott (1999) noted that, for a widely analyzed set of data that consisted of items from the Law School Admissions Test, the fits (and parameter estimates) for a latent trait model with a normally distributed latent trait and a latent class model with two classes were similar; they also gave an example where the covariance structure for a structural equation model was identical to that for a model where all the latent variables were categorical. Similarly, Molenaar and von Eye (1994) showed that it is not possible to distinguish (in terms of the covariance structure) between a latent profile model (with t classes), which treats the latent variable as categorical, and a factor analysis model (with $t - 1$ factors), which treats the latent variable as continuous, and so whether one considers the latent variable to be categorical or continuous is arbitrary. Further comparisons of latent class and latent trait models can be found in Hagenaars and McCutcheon (2002), Heinen (1996), Langeheine and Rost (1988), Lindsay, Clogg, and Grego (1991), Marcoulides and Moustaki (2002), Rost and Langeheine (1997), and von Eye and Clogg (1994).

The view here is that whether a construct should be considered as being categorical or continuous is an open question, on theoretical, statistical, and empirical grounds. It is also more informative, in my view, to compare different models and conceptualizations, rather than to simply take one model or view as "correct." The present study compares a model that treats the construct as categorical to models that treat the construct as continuous.

Given the assumption of a latent categorical variable, the application of SDT to essay grading is straightforward: SDT views each rater as attempting to discriminate between latent classes of essays. Figure 1 illustrates the basic ideas of SDT for three latent classes and a 1–4 response. A rater's decision is viewed as being based in part on his or her perception of the overall quality of an essay (for holistic scoring). The perception of an essay's quality can be viewed as being a realization from a probability distribution on an underlying continuum, with a different probability distribution associated with each latent class. The distances d between the distributions, shown in Figure 1, are of primary interest, in that they reflect a rater's ability to discriminate between the latent classes. It is assumed that, when rating an essay, a rater compares their perception of the essay's quality to response criteria, shown as vertical lines in the figure, and gives a response of "1", for example, if the realization is below the lowest criterion, "2" if it is between the first and second criteria, and so on. Thus, an observed response reflects both perceptual and decisional aspects, namely the rater's

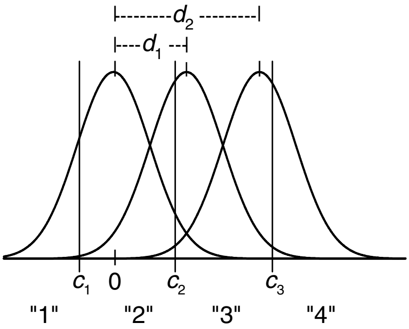

FIGURE 1. *An illustration of signal detection theory with three latent classes and a 1–4 rating response.*

perception of an essay's quality and the rater's use of response criteria; an important aspect of SDT is that it separates these two aspects. For further discussion of the basic SDT model, its interpretation, and applications, see Gescheider (1997), Macmillan and Creelman (1991), or Wickens (2002).

It should be noted that, although the latent class SDT model considered here has not been used for essay grading, to my knowledge, approaches to essay grading via models that in essence include a signal detection component have been considered in the measurement and statistics literature. For example, Patz et al. (2002) presented a hierarchical rater model (HRM) where one level of the model was a latent class signal detection model, although the model was parameterized differently than the usual SDT model. Patz et al. (2002) basically viewed raters as detecting "ideal ratings," where the ideal ratings were latent classes defined by the scoring rubric. In this way, both the latent class SDT model considered here and the hierarchical rater model view raters as attempting to discriminate between latent classes of essays. Note that rater effects are treated as random in the HRM of Patz et al. (2002), whereas rater effects are treated as fixed in the latent class SDT model; Donoghue and Hombo (2000) considered a version of the HRM with fixed rater effects. Johnson and Albert (1999) considered a multirater model for essay grading that in essence used a signal detection conceptualization; their approach differs somewhat from that presented here in that the latent variable was treated as continuous and normally distributed.

The next section introduces the models. The Rasch and IRT models are well known and will only be briefly summarized. The latent class SDT model is presented in more detail.

The Models

Signal Detection Theory with Latent Classes

The latent class signal detection model generalizes the traditional signal detection model (Green & Swets, 1988) to latent classes of events. Specifically, consider the situation where J raters examine N cases (essays, slides, etc.) and assign a discrete score k_j to each case, where $1 \leq k_j \leq K_j$ and K_j is the number of response categories for rater j. For the simple case with two latent classes, the latent class signal detection model can be written as

$$p(Y_j \leq k_j \mid X^* = x_t^*) = F\left(\frac{c_{jk} - d_j x_t^*}{\tau_j}\right), \tag{1}$$

for $1 \leq k_j \leq K_j - 1$, where Y_j is the response variable for rater j, X^* is a latent categorical variable that takes on values of $x_t^* = 0$ or 1 for $t = 1$ or 2, $p(Y_j \leq k_j \mid X^* = x_t^*)$ is the cumulative probability of a response of k_j or less from rater j given x_t^*, c_{jk} are $K_j - 1$ response criteria for the jth rater with $c_{j0} = -\infty$, $c_{jK} = \infty$, and $c_{j1} < c_{j2} < \cdots < c_{j,K-1}$, d_j is a discrimination parameter for the jth rater (i.e., the distance between the underlying perceptual distributions), F is a cumulative distribution function (CDF), and τ_j is a scale parameter, which can be set to unity without loss of generality. The logistic distribution is used here for F (via a logit link function, see below); normal and other distributions have also been considered in SDT (see DeCarlo, 1998).

As noted above, the task is conceptualized in SDT as consisting of two basic components: a perceptual component, which is the rater's perception of the quality of an essay, and a decision component, which involves the placement of response criteria. The focus in research in psychology has primarily been on the discrimination parameter d_j, which can be viewed as a measure of the rater's ability to discriminate between the latent classes. The response criteria c_{jk} are often viewed as depending on arbitrary and uncontrolled factors. As discussed below, the finding of rater drift across sessions or failures of rater training to improve agreement (e.g., see Congdon & McQueen, 2000) are not surprising from the perspective of SDT, in that these are analogous to arbitrary differences in raters' use of response criteria.

The SDT model can be extended to more than two latent classes in several ways. One approach simply allows for more than two classes, without any further restrictions. In that case, the latent classes are unordered categories of events (the latent variable is treated as nominal) and the SDT model provides estimates of the distances between the perceptual distributions for each latent class; note that the same approach has been used in SDT with (multiple) observed signals (see DeCarlo, 1998). The model is closely related to latent class cluster models discussed by Vermunt and Magidson (2000), with the difference that the SDT model uses a cumulative link function for the response probabilities, whereas this is not the case for latent class cluster models. The use of a cumulative link in SDT follows from the conceptualization of the situation in terms of a latent underlying variable, in that the cumulative link function is the inverse of a cumulative distribution function for the latent variable (see DeCarlo, 1998), whereas the cluster models (and factor models) are motivated directly in terms of response probabilities.

Another way to extend the model to more than two latent classes places restrictions on the parameters. For example, in an extension referred to as an equal-distance SDT model (DeCarlo, 2002a), it is assumed that the raters perceive the latent classes as being equally spaced. In this approach, for a model with T latent classes, values of $0, 1, \ldots, T - 1$ are used for x_t^*. In terms of SDT, this places a constraint on the discrimination parameters so that the underlying perceptual distributions are equally spaced; for example, $d_2 = 2d_1$ in Figure 1. Note that the latent classes are only assumed to be ordered, they are not assumed to be equally spaced; the equal spacing is in the raters' perceptions, not the latent classes. This extension represents a new type of SDT model. From a statistical perspective, it can be viewed as a latent class extension of a uniform association model for cumulative odds ratios noted by Agresti (1990); it is also related to latent class factor models discussed by Vermunt and Magidson (2000; with the difference that the SDT model uses a cumulative link function). For the data considered here (also see DeCarlo, 2002a), the equal-distance SDT model was consistently favored (in terms of information criteria) over the unrestricted SDT extension noted above, and so only the equal-distance SDT model is considered.

Equation 1 specifies the model for each rater. The model can be incorporated into a restricted latent class model (e.g., Clogg, 1995; Dayton, 1998) as follows. For J raters, the observed data are response patterns that consist of J elements. A latent class model assumes that there are $t = 1$ to T mutually exclusive and exhaustive latent classes X^* so that the probabilities of the response patterns can be obtained by summing over the latent classes,

$$\begin{aligned} p(Y_1 = k_1, Y_2 = k_2, \ldots, Y_J = k_J) \\ = \sum_{t=1}^{T} p(Y_1 = k_1, \ldots, Y_J = k_J, X^* = x_t^*) \\ = \sum_{t=1}^{T} p(X^* = x_t^*) p(Y_1 = k_1, \ldots, Y_J = k_J \mid X^* = x_t^*), \end{aligned} \tag{2}$$

where $p(Y_1 = k_1,\ Y_2 = k_2, \ldots, Y_J = k_J)$ is the probability of response pattern $(k_1, k_2, \ldots, k_J)$, $p(Y_1 = k_1, Y_2 = k_2, \ldots, Y_J = k_J \mid X^* = x_t^*)$ is the conditional probability of the response pattern given $X^* = x_t^*$, and $p(X^* = x_t^*)$ is the size (mixing proportion) of latent class t with $p(X^*) > 0$ for all t and $\sum_t p(X^* = x_t^*) = 1$. Furthermore, conditional on the latent class, responses are assumed to be independent, so that

$$p(Y_1 = k_1, \ldots, Y_J = k_J \mid X^* = x_t^*) = \prod_{j=1}^{J} p(Y_j = k_j \mid X^* = x_t^*), \tag{3}$$

where $p(Y_j = k_j \mid X^* = x_t^*)$ is the conditional probability of response k_j for rater j given $X^* = x_t^*$ and $\sum_k p(Y_j = k_j \mid X^*) = 1$. Equation 3 reflects a basic assumption of latent class analysis, which is that the J response variables are independent given the latent class.

Finally, to incorporate the SDT model into the latent class model, differences of the cumulative probabilities of Equation 1 are used for the conditional probabilities of Equation 3,

$$\begin{aligned} p(Y_j = k_j \mid X^* = x_t^*) &= F(c_{jk} - d_j x_t^*) \qquad k_j = 1 \\ p(Y_j = k_j \mid X^* = x_t^*) &= F(c_{jk} - d_j x_t^*) - F(c_{jk-1} - d_j x_t^*) \qquad 1 < k_j < K_j\,. \\ p(Y_j = k_j \mid X^* = x_t^*) &= 1 - F(c_{jk-1} - d_j x_t^*) \qquad k_j = K_j \end{aligned} \tag{4}$$

Thus, the full model consists of an SDT model for each rater incorporated into a restricted latent class model.

Rater precision. From the perspective of SDT, it is of basic interest to assess how well a rater can discriminate between the latent classes, and estimates of d_j provide information on exactly that. In particular, one can view d_j (or its inverse) as providing a measure of the precision of a rater, with larger values indicating greater precision. Using terminology suggested by Clogg and Manning (1996), the discrimination parameter (or a correlation-type transform of it, such as Yule's Q; see DeCarlo, 2002a) provides a measure of rater-level reliability.

Classification accuracy. Another aspect of reliability concerns the accuracy of the classifications. The statistic lambda is useful in this regard, in that Clogg and Manning (1996) noted that lambda can be viewed as a nonparametric measure of the reliability of the classifications obtained from the set of raters (items). Lambda is computed using an estimate of the proportion correctly predicted, P_C, which in turn is computed using posterior probabilities (see Dayton, 1998; DeCarlo, 2002a). The proportion correctly predicted provides a measure of classification accuracy, whereas lambda indicates the relative increase in the proportion correctly predicted as compared to the largest latent class size. Note that there is an upward bias in the estimate of P_C (and therefore lambda); the simulation presented below provides some information about the magnitude of the bias for the data considered here. The size of lambda depends on several factors, with the magnitude of the discrimination parameters and the number of raters per essay having a large influence.

Latent Trait Models

The latent trait models considered here are the partial credit (PC) model (Masters, 1982), the generalized partial credit (GPC) model (Muraki, 1992), and the graded response (GR) model (Samejima, 1969). The GR model is closely related to Equation 1. A general version of the model can be written as

$$p(Y_j \leq k_j \mid \theta) = F[a_j(b'_{jk} - \theta)] = F(b_{jk} - a_j\theta), \tag{5}$$

where F is the logistic CDF, θ is a continuous latent variable, a_j is a discrimination parameter, and b_{jk} differs from the usual difficulty parameter b'_{jk} in that it equals $a_j b'_{jk}$; the parameterization used in Equation 5 (cf. McDonald, 1999) is useful for showing how the model is related to the SDT model (i.e., b_{jk} is analogous to the response criteria c_{jk} and a_j is analogous to d_j). A comparison of Equations 1 and 5 shows that a basic difference is that the latent variable θ is continuous in latent trait models whereas the latent variable X^* is discrete in the latent class SDT model. From the perspective of IRT, the latent class SDT model can be viewed as a semiparametric version of the graded response model (see Heinen, 1996). Note that Equation 5 can also be written as a generalized linear model (McCullagh & Nelder, 1989) by applying the inverse of F, say $g = F^{-1}$, to both sides, which gives

$$g[p(Y_j \leq k_j \mid \theta)] = b_{jk} - a_j\theta, \tag{6}$$

where g is a link function. Using the logit link, $g = \log[p/(1-p)]$ gives the GR model; other link functions can also be used (Mellenbergh, 1994). The logit link is also used for the latent class logistic SDT model considered here.

The GPC model, with a parameterization similar to that of Equation 6 (cf. Heinen, 1996), can be written as

$$\log\left(\frac{p(Y_j = k_j + 1 \mid \theta)}{p(Y_j = k_j \mid \theta)}\right) = b_{jk} - a_j\theta. \tag{7}$$

A comparison of Equations 6 and 7 shows that the basic difference is that the GPC model uses adjacent-category logits (see Agresti, 1990; Heinen, 1996) in lieu of a logit link; note that the GPC model can also be rewritten as a log-linear model with scores (see Clogg & Shihadeh, 1994). The PC model restricts the discrimination parameter a_j in the GPC model of Equation 7 to be equal across the raters.

Data Analysis

The models discussed above are examined here with a set of real-world data. The data consisted of 125 essays obtained in a college class, where the goal was to assign grades. An example of an application of latent class SDT to essay grading with a larger data set (and more raters), along with evidence as to criterion validity, can be found in DeCarlo (2002b).

Method

The essays were from 125 students in a graduate introductory measurement course who, as part of a final exam, wrote a one-page essay on how they would evaluate a new questionnaire. The students were given 1 hour to write the essay in class. Eight raters (professor and seven graduate students) graded each essay on a 1–4 scale, with 1 = definitely below average, 2 = average to slightly below average, 3 = average to slightly above average, and 4 = definitely above average. The raters were instructed to grade on content, and to try to ignore other aspects of the essay, such as handwriting quality, spelling, or length of the essay (which was restricted to be no more than one page). The average score on three in-class multiple-choice exams was used as a criterion to assess the validity of the essay scores.

The Design

The design is fully crossed, in that all of the raters graded all of the essays (for an example with an incomplete design, see DeCarlo, 2002b). This design was used because the focus here is on the application of SDT as a model of rater behavior; eight raters were used because some small simulations (with a 2-class model) suggested that more than five raters would be needed in order to obtain acceptable estimation precision of the raters' parameters for a sample size of around 100; this is reinforced by the simulation presented below (note that increasing the number of raters compensates to some extent for a small sample size, see DeCarlo 2002a). Designs with one essay and many raters are often used in research studies, such as studies of rater training (e.g., Weigle, 1998), rater reliability (e.g., Blok, 1995; Shohamy, Gordon, & Kraemer, 1992), computer-based scoring (e.g., see Johnson, 1996), and in situations where new models of rater behavior are introduced, such as the multirater model of Johnson and Albert (1999). The SDT approach can also be extended to more than one essay and/or more than one response per rater, but the focus here is limited to the simple multiple rater, one-essay situation; an extension to more than one response per rater is given in DeCarlo (2003).

Fitting the models. The latent class and latent trait models discussed above were fit using LEM (Vermunt, 1997), which is a general package for categorical data analysis that provides maximum likelihood estimates of the parameters of latent class models by means of the EM algorithm. The programs were run several times with different starting values, because of a well-known problem in latent class analysis with local maxima (e.g., Aitkin, Anderson, & Hinde, 1981; McLachlan & Peel, 2000). The latent trait models were fit using marginal maximum likelihood (MML); 41 nodes and quadrature weights from −4.5 to 4.5 were used.

Classification and scoring. For the latent class SDT model, the posterior probabilities of each latent class, given the response pattern, were used to assign each essay to the latent classes, that is, each essay was assigned to the latent class with maximum posterior probability. For the latent trait models, Multilog (Thissen, 1991) was used to assign maximum *a posteriori* (MAP) scores. The SDT classifications and item response scores are compared below.

Validation. Correlations of the classifications and scores with a criterion variable were examined; the average score on three in-class multiple-choice exams was used as the criterion. In addition to the usual Pearson correlation, a nonparametric measure of association, Kendall's τ (Kendall, 1945), was examined. τ_b is useful for assessing the degree to which the relationship between the classifications (or scores) and the criterion is monotonic; the square of τ_b also has an interpretation, like the square of Pearson's correlation, in terms of a proportional reduction in error (Wilson, 1969).[1]

Results

Model Selection

Because of the sparseness of the data (i.e., there are 4^8 possible patterns of ratings and only 125 observations), likelihood ratio and chi-square goodness-of-fit statistics are not useful for assessing the absolute fit of the models (one cannot assume that the statistics follow their asymptotic distributions for sparse data). Information criteria, on the other hand, can be used to compare the different models (Lin & Dayton, 1997; Sclove, 1987). Although this does not assess the fit of the models to the data, it allows a comparison of the fit of the latent class SDT model to models that are currently used for essay grading, such as the Rasch model and IRT models (i.e., it assesses the relative fit of the models). It should also be noted that model selection is a complex topic and other approaches have been proposed or are under development.

Table 1 shows, for latent class logistic SDT models with from one to six latent classes and for the IRT models, information criteria, which can be used to compare nonnested (and nested) models, with smaller values indicating a preferred model (see Agresti, 1990; Burnham & Anderson, 2002; Dayton, 1998). The table shows values of a version of Akaike's information criterion with a small sample bias correction (AIC_c) and the Bayesian information criterion (BIC). Burnham and Anderson (2002) recommended the use of AIC_c over AIC when the ratio of the sample size to the number of parameters is small (say <40), as is the case here. Specifically, $\text{AIC}_c = -2\log L + 2p[N/(N - p - 1)]$, where L is the likelihood, p is the number of parameters, and N is the sample size, whereas $\text{BIC} = -2\log L + p\log N$. Guidelines with respect to interpreting the magnitude of differences of AIC_c from the smallest value in a set of models are given by Burnham and Anderson (2002);

TABLE 1
Information Criteria for the Various Models ($N = 125$)

Model	No. of Parameters	AIC_c	BIC
Logistic SDT, 2 classes	32	2160.24	2227.78
Logistic SDT, 3 equal distance	33	2080.94	2149.61
Logistic SDT, 4 equal distance	34	2072.47	2142.19
Logistic SDT, 5 equal distance	35	2057.74	2128.42
Logistic SDT, 6 equal distance	36	2059.93	2131.48
Graded response (41 nodes)	31	2042.88	2109.22
Generalized partial credit (41 nodes)	31	2051.26	2117.61
Partial credit (41 nodes)	24	2102.63	2158.51

AIC_c = small sample bias corrected Akaike's information criteria; BIC = Bayesian information criterion.

Kass and Raftery (1995) suggested similar guidelines for differences in twice the log Bayes factors (BIC can be viewed as an approximation to the Bayes factor).

Among the latent class SDT models, Table 1 shows that both AIC_c and BIC are smallest for the model with five classes. Among the total set of models, both information criteria suggest the GR model as best. The parameter estimates given below show that the estimated latent class sizes for the 5-class SDT model are close in value to the weights used in Gaussian quadrature (as was noted by a reviewer), which is why the GR model, which assumes a normal distribution for the latent trait (and so has fewer parameters than the SDT model), is likely favored over the latent class SDT model in this case (as noted above, the two models only differ with respect to the treatment of the latent variable). Of course, this need not generally be the case (the latent class SDT model in essence allows for arbitrary distributional shapes). It is also interesting to note that the partial credit model, which is probably the most widely used model for essay grading in practice, fares poorly as compared to the other models; this occurs in part because there are clearly differences in discrimination across the raters, as shown next.

As noted above, a decision as to whether a latent variable is continuous or discrete involves more than just statistical considerations, in that theory and practical utility also play a role. The next section examines parameter estimates and the utility of scores and classifications obtained from the models.

Parameter Estimates

The left panel of Figure 2 shows, for the 5-class SDT model, a plot of the estimates of the discrimination parameters and their standard errors. The figure shows that the point estimate of d_j is largest for rater B and smallest for raters E and H. The standard error of d_j is rather large for rater B; the simulation presented below shows that the standard errors tend to be large for large values of d_j (which reflects a well-known problem). A likelihood ratio test of a restricted model with equal discrimination parameters across the raters, which is nested within a model with unrestricted parameters, rejects the restricted model (LR = 77.04, $df = 7$, $p < 0.01$). Thus, the raters differed with respect to their ability to discriminate between the latent classes.

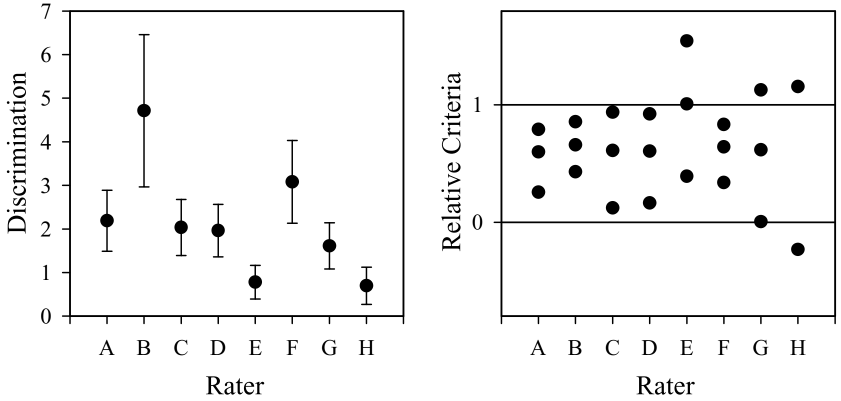

FIGURE 2. *The top left and right panels show, respectively, a plot of the estimated discrimination parameters and response criteria for the raters in the present study, for a 5-class SDT model.*

The response criteria are defined in Equation 1 with respect to their distance from the lowest latent class (see DeCarlo, 1998, for comments on an alternative parameterization). Note that when discrimination varies across the raters, as here, it is not very informative to simply compare the (absolute) criteria locations. A more useful approach is to set the distance between the highest and lowest underlying distributions to be equal across raters, so that the relative locations of the response criteria (i.e., relative to the highest and lowest underlying distributions) can be compared. This was done by dividing the estimates of the response criteria by the estimated distance between the highest and lowest distributions for each rater (which is simply the estimate of d_j times $T-1$). I refer to these as the *relative* response criteria locations.

The right panel of Figure 2 shows a plot of the relative criteria locations for the eight raters, again for the 5-class SDT model. The horizontal reference lines show the location of the lowest distribution (at zero) and the highest distribution (at 1). The figure shows that the relative locations of the response criteria were similar across the raters, with the exception that rater E's criteria for responses of 3 and 4 were higher than those for the other raters. Thus, rater E was stricter than the other raters with respect to giving higher scores. Also note that there are only two criteria for rater H, because the rater used only three response categories.

In summary, the raters' precision ranged from low (<1.0) to high (≥ 3.0) and the relative locations of the response criteria were similar across most of the raters, but also differed in some cases. Note that, in terms of agreement, computing weighted kappa (Cohen, 1968) for each pair of raters gave values that ranged from .03 to .46, and so agreement ranged from very poor to moderate, using guidelines suggested by Landis and Koch (1977). More importantly, the results for weighted kappa were predictable from the SDT parameter estimates shown in Figure 2. For example, rater E, who had low detection and response criteria that differed from the other raters, gave the lowest values of weighted kappa (when paired with other raters), whereas weighted kappa was largest for the two raters with the largest estimates of d_j (raters B and F, who also had similar response criteria, as can be seen in Figure 2). Thus, the

SDT parameters account for why agreement ranged from poor to moderate, in that the level of agreement depends on the rater parameters d_j and c_{jk}.

With respect to the latent class sizes, the estimates for the 5-class SDT model (with standard errors in parentheses) are .071 (.035), .218 (.047), .362 (.053), .293 (.047), and .056 (.025) for the lowest to highest latent classes, respectively. With respect to classification accuracy, the estimate of the proportion correctly predicted for the 5-class model is .90. The value of lambda is .85, which means that there is an 85% increase in correct classification by using the posterior probabilities as compared to classifying all of the essays into the largest latent class, which would give only 36% correct classification in this case. The simulation presented below shows, however, that there is likely a considerable upwards bias in lambda for a sample size of 125.

Classification and scoring. Further insight into the different models can be obtained by comparing the scores or classifications obtained from the models to the scores obtained by simply averaging the raters' ratings. Figure 3 presents plots of (a) the five classifications obtained from the SDT model, (b) the scores obtained from the PC model, (c) the scores obtained from the GPC model, and (d) the scores obtained from the GR model, all plotted against the average of the eight rater's scores.

The upper right panel of Figure 3 shows that the scores obtained from the PC model are an order-preserving nonlinear transform of the average score. This is a

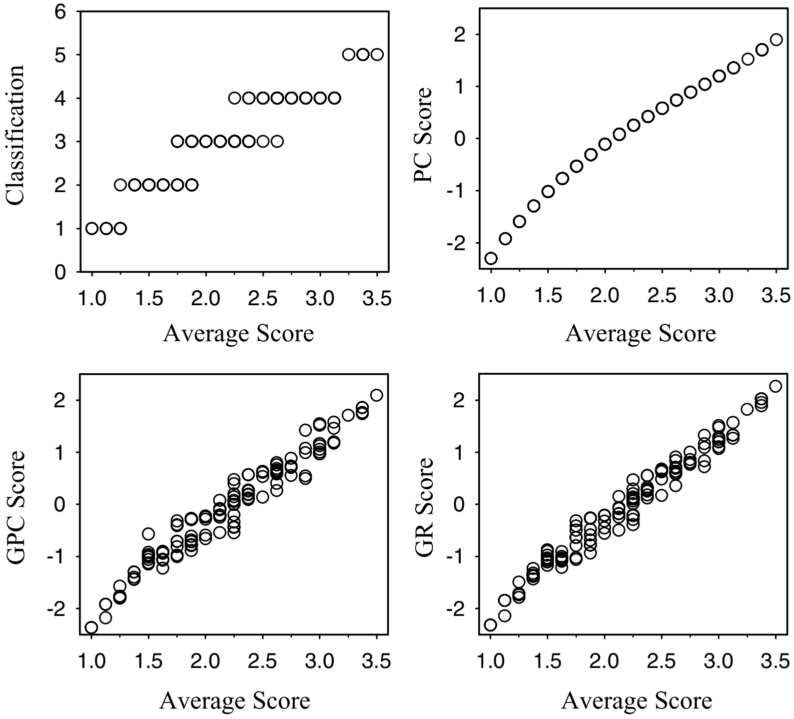

FIGURE 3. *Plots of the classifications from a 5-class SDT model, PC scores, GPC scores, and GR scores against the average scores.*

basic characteristic of the PC model that follows from the use of adjacent category logits and the restriction of equal discrimination parameters across the raters. In contrast, the two lower panels of Figure 3 show that scores from the GPC model and the GR model do not preserve the order of the average scores. For example, for an average score of 2.25, the GPC scores varied from -0.55 to 0.47. Thus, two essays with the same average score can receive higher or lower scores, depending on the response pattern (of course, the standard errors should also be taken into account). In addition, an essay with a lower average score can receive a higher GPC or GR score than an essay with a higher average score.

The top left panel of Figure 3 shows that the classifications obtained from the 5-class SDT model also do not preserve the order of the average scores. Points that can be connected by a vertical line represent cases where two essays with the same average score are in different latent classes. The figure shows that one cannot find a vertical line that cuts across more than two latent classes, and so in this case two essays with the same average score never differ by more than one latent class. Also note that an essay with a higher average score can end up in a lower latent class than an essay with a lower average score. The figure shows that this happened primarily with scores in the middle of the range; for example, there is no overlap for the fourth and fifth latent classes, and very little for the first and second latent classes, whereas the greatest region of overlap is for the third and fourth classes, for cases with average scores of from 2.25 to 2.6.

From the perspective of SDT, the reason why the classifications do not necessarily preserve the order of the average scores is straightforward. If one essay received a higher average score than a second essay, for example, but the first essay received high ratings from raters with poor discrimination whereas the second essay received (say fewer) high ratings from raters with good discrimination, then the first essay might end up in a lower latent class, because the classifications take into account (via the posterior probabilities) differences in the raters' discrimination and response criteria. In contrast, scores from the PC model preserve the order of the average scores, and so with respect to the example just given, the first essay would always receive a higher score. Thus, one can argue that the PC model is limited in the way that it adjusts scores to account for rater differences, in that it can only adjust the spacing between scores and not the order of the scores (for a fully crossed design).

It is also informative to compare the SDT classifications and the IRT scores. Figure 4 shows the classifications obtained from the 5-class SDT model plotted against the scores obtained from the GPC and GR models. The left panel shows that the latent class SDT model partitions the GR scores into clusters that do not overlap, whereas the right panel shows that the latent class model does not partition the GPC scores, which reflects the effect of the use of a different link function. Note, however, that there is vertical overlap for only one case for the GPC scores (at $x = -0.5$ and $y = 2$) and so the results are similar to those obtained with the GR model in this case. If it is assumed that the latent variable is quantitative, then the challenge is to show that the variation of IRT scores within the latent classes offers additional useful information.

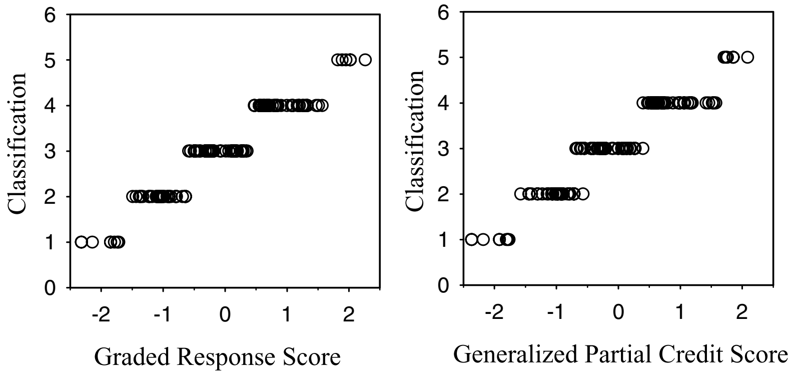

FIGURE 4. *Plots of the classifications from a 5-class SDT model against the GPC and GR scores.*

Criterion validity. Evidence of criterion validity has often not been a part of studies that have compared latent class and latent trait models. Table 2 shows measures of association (Pearson correlations and Kendall's τ_b) between the classifications (from the SDT model) and scores (from the IRT models) and a criterion, which was the average score on three exams. The Pearson correlation is 0.55 for the 5-class SDT model and is similar in magnitude to the values obtained for the average score and IRT models. τ_b is largest for the 3- and 5-class SDT models and is slightly smaller for the average score, PC model, GPC model, and GR model. Overall, with respect to criterion validity, Table 2 suggests that there is little or no difference between using continuous IRT scores or average scores versus simply classifying the essays into three to five ordered latent classes.

It should be noted that a limitation of simply using the essays' class membership in measures of association, as in Table 2, is that it ignores uncertainty in the class membership, as noted by Aitkin et al. (1981) and Clogg (1995), for example. Possible ways to address this are by using the posterior probabilities rather than just class membership (as suggested by Aitkin et al., 1981), by using multiple imputation

TABLE 2
Pearson Correlation and Kendall's τ_b for the Classifications or Scores with a Criterion (Exam Average)

Model	r	τ_b
Logistic SDT, 2 classes	0.39	0.30
Logistic SDT, 3 equal distance	0.53	0.41
Logistic SDT, 4 equal distance	0.51	0.37
Logistic SDT, 5 equal distance	0.55	0.41
Logistic SDT, 6 equal distance	0.54	0.41
Graded response	0.55	0.36
Generalized partial credit	0.54	0.36
Partial credit	0.58	0.38
Average Score	0.56	0.38

(using the posterior probabilities from the latent class model), or by incorporating the criterion variable directly into the model (e.g., DeCarlo, 2002b).[2] This requires further study.

Discussion

The scoring rubric for the present study consisted of four categories, and so one could argue that the rubric defined four latent classes. The information criteria in Table 1, however, suggest that a 5-class SDT model was best among the latent class models. Note that the 4-category scoring rubric used middle categories with labels of 2 = "average to slightly below average" and 3 = "average to slightly above average." The results suggest that these two categories might be divided into three categories, such as slightly below average (2), average (3), and slightly above average (4), giving a total of five latent classes. Thus, one can interpret the latent classes as consisting of an "average" class surrounded by higher and lower classes (which are smaller in size). Of course, it should be kept in mind that the interpretation of the latent classes goes beyond the model and involves other information and considerations.

The estimates of the detection parameters showed that rater precision varied considerably across the raters. The response criteria were generally similar across raters, but at least two of the raters differed from the others. Because of these differences, pairwise agreement ranged from poor to moderate. Nevertheless, classification, as assessed by the estimate of P_C and lambda, appeared to be adequate (even when bias is considered; see the simulation below), thanks in part to the large number of raters. Correlations of the latent classes with average exam scores provided evidence as to the validity of the classifications, with the results suggesting little difference between treating the latent variable as categorical or continuous. Thus, the current study joins earlier studies, noted above, which have shown more similarities than differences between latent class and latent trait models. The results suggest that classifications from the latent class SDT model might have as much utility as scores from latent trait models or average scores. Some issues and limitations of the current approach were also noted.

SDT models with ordinal latent classes have not previously been used, to my knowledge. The next section presents a simulation that was conducted in order to obtain information about the performance of the estimators and the accuracy of classification for a 5-class SDT model, as used above. The simulation provides additional information about the results just presented.

A Simulation

Method

Simulation design. The simulation design followed from the results found above. The number of raters was eight, with the raters using a 1–4 ordinal response scale. The population model was a logistic SDT model with five ordered latent classes with the population values for d_j and c_{jk} being the same (with rounding) as those obtained above (the population values are given in the tables below). The latent class sizes used were, in order from lowest to highest class, 0.07, 0.22, 0.36, 0.29, and 0.06, which are the same as those found above (rounded to two decimal places).

The number of replications (i.e., samples generated) was 100. Two sample sizes (i.e., number of essays) were examined, one of 125, as in the present study, and one of 300; the sample size of 300 was included to obtain an idea about how much estimation and classification might improve with a modest increase in sample size.

The data were simulated with SAS (Release 8.2) using a macro written by the author. The simulation involved three steps. First, for the 5-class SDT model, values for the latent variable X^* (i.e., $x_t^* = 0-4$) were generated according to a multinomial distribution, with the estimates of the latent class sizes given above used for $p(X^* = x_t^*)$. Second, the values of x_t^* together with the population values of the SDT parameters c_{jk} and d_j were used in Equation 1 with a logistic distribution for F to generate probabilities for each response category and rater. Third, these values were compared to values obtained from a uniform random variable generated on an interval of 0 to 1. If the uniform 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.

Each of the 100 replications was analyzed using LEM (Vermunt, 1997). A SAS macro written by the author was used to generate (100) input files for LEM and also a DOS batch file that was used to call LEM repeatedly. Other SAS macros stripped out information from the LEM output for each replication, and the results were combined in a file that was then used for the remaining analyses.

Some Considerations

Local maxima and convergence. The default value for convergence in LEM is a minimum increase in the log-likelihood of 0.000001. The convergence value was included in the summary output so that it could be determined if there were any cases where nonconvergence occurred. For the present simulation, the default number of iterations in LEM was increased from 1,000 to 20,000, and convergence occurred in all cases except one (for $N = 125$), which converged after repeated runs. A potential problem is that the converged solution could represent a local maxima, as noted above. This can be checked by re-running the program with different starting values, and checking for a change in the maximized log-likelihood (obtaining a larger value indicates a local maxima). For the present simulation, the program was re-run with different starting values several times for each of the 100 replications (for each condition) and the log-likelihoods were checked for any change. Local maxima were encountered in several of the replications for the sample size of 125, and in only a few cases for the sample size of 300.

Empirical identifiability. The models examined here are identified; however, problems can still arise with what can be termed empirical identifiability. This can be detected by examining the eigenvalues of the estimated information matrix; Goodman (1974) noted that nonzero values are sufficient for identification. For each replication of the simulation, the smallest eigenvalues were printed, so that the number of times the smallest eigenvalue had values at or below zero could be kept track of. This occurred in 36 of the 100 replications for the sample size of 125 (which led to estimation problems mostly for the largest discrimination parameter, see below) and in 1 of the 100 replications for the sample size of 300.

Redundant solutions. It should be recognized when fitting the model that there are two redundant solutions that will appear over repeated runs; the maximized log-likelihoods for the two solutions are identical (because one solution is simply a re-parameterization of the other). The two solutions arise because it is arbitrary whether the lowest or highest latent class is labeled as zero. This is not a problem, but must simply be kept track of. The order of the latent classes can be determined from the sign of the discrimination parameter. When the lowest class is coded as zero, then the model is the same as in Equation 1 above, and the estimates of d will be negative. When the coding is reversed (the lowest class is coded as $T-1$), the estimates of d will be the same, but the signs will be positive instead of negative; the criteria estimates are also not those given in Equation 1, but can be obtained by adding the estimate of d_j to each criteria estimate. For the present simulation, a SAS macro that stripped out and combined the data for final analysis also checked the signs of d and, if needed, adjusted the sign of d_j, the criteria, and the direction of the latent classes.

Outcome Measures

The estimators of the SDT parameters were assessed by examining the bias (estimated value minus true value) and mean squared error (MSE). Bias was also examined for the latent class sizes. The standard errors given by LEM were evaluated by comparing the standard deviation of the estimated parameters across replications to the mean of the estimated standard errors, with bias defined as the latter value minus the former value. The estimates of the standard errors of the discrimination parameters and the latent class sizes were evaluated in this way.

Of particular interest is the performance of P_C and lambda, since in real-world research the basic goal is often to classify cases (e.g., to make a selection, placement, or grading decision). Evaluation of P_C proceeded as follows. For each replication, the estimate of P_C was obtained by summing the estimated posterior probabilities over all the cases (note that this approach uses the observed frequencies in lieu of expected frequencies; see Clogg, 1995; DeCarlo, 2002a). Note that the classification for each case can be compared to the realized value of X^*, which is known in a simulation, and so the proportion of cases that are classified correctly in each replication can be determined; I refer to this simply as the *obtained* P_C. The mean of the estimates of P_C (which are obtained from the posterior probabilities, as noted above) over the 100 replications was compared to the mean of the obtained values of P_C over replications. This was also done for lambda: the mean of the estimates of lambda over the 100 replications was obtained (using the estimates of P_C and the estimates of the largest class size) and was compared to the mean of the obtained lambdas; the obtained lambdas were computed for each replication by using the obtained P_C and the obtained maximum latent class size.

Note that, with respect to evaluating the usefulness of the classification, lambda is somewhat restrictive in that it only considers exact agreement between the classifications and the latent classes. Pearson correlations between the classifications and the obtained latent classes, and Kendall's τ_b for the classification and the obtained latent classes were also examined.

Results and Discussion

Table 3 presents, for simulations with sample sizes of 125 and 300, the population values of the SDT parameters, the mean estimates across the replications, the bias, and the MSE. With respect to the discrimination parameters, the left side of the table shows that, for $N = 125$, the bias is small (i.e., less than 10%) for the estimates of

TABLE 3
SDT Parameter Estimates, Bias, and MSE for 5-class SDT Simulation

		$N = 125$			$N = 300$		
Parameter	Value	Estimate	Bias	MSE	Estimate	Bias	MSE
d_1	0.7	0.670	−0.030	0.040	0.670	−0.031	0.018
d_2	0.8	0.750	−0.050	0.047	0.772	−0.028	0.018
d_3	1.6	1.521	−0.079	0.150	1.578	−0.022	0.046
d_4	2.0	1.886	−0.114	0.166	2.030	−0.030	0.068
d_5	2.0	1.882	−0.118	0.156	1.990	−0.010	0.072
d_6	2.2	2.055	−0.145	0.269	2.208	−0.008	0.085
d_7	3.0	2.838	−0.162	0.547	2.970	−0.030	0.126
d_8	4.7	5.912	1.212	9.207	4.799	0.099	0.475
c_{11}	−0.6	−0.594	0.006	0.295	−0.632	−0.032	0.111
c_{12}	1.5	1.551	0.051	0.293	1.482	−0.018	0.118
c_{13}	3.2	3.267	0.067	0.359	3.184	−0.016	0.128
c_{21}	1.2	1.231	0.031	0.367	1.156	−0.044	0.117
c_{22}	3.2	3.260	0.060	0.439	3.161	−0.039	0.130
c_{23}	4.8	4.897	0.097	0.573	4.802	0.002	0.177
c_{31}	0.0	−0.105	−0.105	0.718	−0.102	−0.102	0.250
c_{32}	4.0	4.091	0.091	1.193	4.010	0.010	0.338
c_{33}	7.3	7.543	0.243	1.965	7.346	0.046	0.452
c_{41}	1.3	1.236	−0.064	1.359	1.302	0.002	0.333
c_{42}	4.7	4.814	0.114	1.757	4.855	0.155	0.518
c_{43}	7.2	7.338	0.138	2.033	7.402	0.202	0.715
c_{51}	1.0	0.941	−0.059	1.103	0.962	−0.039	0.434
c_{52}	5.0	5.105	0.105	1.392	5.036	0.036	0.575
c_{53}	7.6	7.741	0.141	1.747	7.627	0.027	0.738
c_{61}	2.3	2.272	−0.028	1.488	2.298	−0.001	0.491
c_{62}	5.2	5.169	−0.031	1.585	5.280	0.080	0.664
c_{63}	7.0	7.023	0.023	1.993	7.132	0.132	0.766
c_{71}	4.2	4.246	0.046	4.916	4.182	−0.018	0.917
c_{72}	8.0	8.194	0.194	6.275	8.051	0.051	1.247
c_{73}	10.0	10.280	0.280	7.082	10.077	0.077	1.352
c_{81}	8.0	11.225	3.225	56.239	8.268	0.268	2.610
c_{82}	12.0	16.356	4.356	87.773	12.450	0.450	3.579
c_{83}	16.0	21.535	5.535	126.488	16.631	0.631	4.667

MSE = mean squared error.

TABLE 4
Latent Class Size Parameter Estimates and Bias for 5-class SDT Simulation

		$N = 125$		$N = 300$	
Parameter	Value	Estimate	Bias	Estimate	Bias
Class Size 1	0.07	0.105	0.035	0.086	0.016
Class Size 2	0.22	0.157	−0.067	0.202	−0.018
Class Size 3	0.36	0.330	−0.030	0.349	−0.011
Class Size 4	0.29	0.248	−0.042	0.283	−0.007
Class Size 5	0.06	0.160	0.100	0.080	0.020

d_1 to d_7 (which have values of 3 or less) and the MSE is also small, whereas the bias and MSE for the estimate of d_8 (which is 4.7) for rater 8 are large. The problem with estimation for d_8 reflects a well-known problem in conventional SDT that arises with large values of the discrimination parameter (because of small cell frequencies; see Macmillan & Creelman, 1991). Table 3 also shows that the bias and MSE for the response criteria are generally small for the first seven raters but are large for rater 8. The right side of the table shows that increasing the sample size to 300 results in a considerable improvement in estimation, in that the bias and MSE are clearly smaller in all cases.

Table 4 shows that the estimates of the latent class sizes are somewhat off for a sample size of 125, whereas estimation is considerably improved for a sample size of 300. With respect to the standard errors, Table 5 shows that the estimates of the SEs of the estimates of d_j are negatively biased for both sample sizes. For $N = 125$, the magnitude of the bias is small for the smallest values of d_j, gets larger as d_j increases,

TABLE 5
Evaluation of Standard Error Estimates for SDT Parameters and Latent Class Sizes, 5-class SDT Simulation

	$N = 125$			$N = 300$		
Parameter	SD	Mean SE	Bias	SD	Mean SE	Bias
d_1	0.200	0.176	−0.024	0.133	0.119	−0.014
d_2	0.213	0.188	−0.024	0.132	0.128	−0.004
d_3	0.381	0.273	−0.108	0.215	0.183	−0.032
d_4	0.393	0.308	−0.085	0.261	0.209	−0.052
d_5	0.378	0.307	−0.071	0.269	0.207	−0.062
d_6	0.501	0.326	−0.175	0.293	0.224	−0.069
d_7	0.725	0.463	−0.262	0.356	0.300	−0.056
d_8	2.796	1.147	−1.649	0.686	0.667	−0.019
Class Size 1	0.081	0.045	−0.036	0.046	0.029	−0.017
Class Size 2	0.089	0.054	−0.035	0.055	0.034	−0.021
Class Size 3	0.076	0.054	−0.021	0.043	0.034	−0.009
Class Size 4	0.101	0.048	−0.053	0.043	0.031	−0.009
Class Size 5	0.119	0.037	−0.082	0.064	0.019	−0.045

TABLE 6

Classification and Correlation Statistics for Classifications, 5-class 8-rater SDT Simulation

Sample Size	P_C	Obtained P_C	λ	Obtained λ	tau-b	r
100	0.902	0.685	0.847	0.502	0.934	0.916
300	0.895	0.841	0.838	0.751	0.944	0.926

and is large for the largest value, d_8 (note that the mean of the standard errors for the estimate of d_8 in Table 5 for $N = 125$ is based on only 76 of the 100 replications, since the standard errors were indeterminate in 24 replications). The right side of the table shows that, for $N = 300$, the bias is small in all cases. Table 5 shows that the estimates of the standard errors of the latent class size estimates are also negatively biased. Negative bias means that the estimated standard errors tend to underestimate the population standard errors, and so significance tests tend to be too liberal and confidence intervals too narrow.

Table 6 shows results for the proportion correct and lambda. The tables shows that, for both sample sizes, the estimates of P_C are close to 0.9. Note that, given the estimates of the SDT parameters and latent class sizes, one can determine a population value (i.e., for an infinite sample size) for P_C, which in this case was 0.894, and so the estimates of P_C shown in Table 6 are quite close to the large sample value. This, however, ignores finite sample bias. The obtained P_C and obtained lambda shown in Table 6 show that, for a sample size of 125, there is overestimation of the proportion correctly predicted, in that the mean (over replications) obtained P_C was 0.685 whereas the mean estimate of P_C was 0.902. As a result, the estimate of lambda (0.847) is also larger than the obtained lambda (0.502). In contrast, for the sample size of 300, the amount of overestimation is much smaller, with a mean value of P_C of 0.895 and an obtained value of 0.841.

As noted above, lambda is rather strict in that it only reflects exact agreement. Table 6 shows that the estimates of τ_b and r are large (>0.90) for both sample sizes, and so the order of the latent classes was well preserved by the classifications.

In summary, the simulation offers useful information and some practical guidelines. For one, it shows that estimation and classification with a 5-class SDT model with eight raters (and the given population values) were quite good for a sample size of 300. For a sample size of 125, estimation of the SDT parameters was good for values of d_j less than about 3, whereas the bias and MSE were large for a value of d_j of 4.7, and estimation of the latent class sizes was marginal. One is also more likely to encounter problems with empirical identification for smaller sample sizes. Thus, if the goal is to estimate the raters' discrimination parameters, then larger sample sizes (or more raters) are needed when large values of d are expected, such as in a study that involves experienced raters, or in other situations where good discrimination is expected. A possible remedy for the positive bias found for a large value of d might be to constrain d to be less than a certain value, but this is not possible in LEM (to my knowledge; the software Latent Gold allows one to use Bayes constants, but cumulative links are not offered; see Vermunt & Magidson, 2000). On the other hand, whether or not positive bias in estimates of large values of d is a problem depends on the purpose of the study. If the goal is to evaluate the raters, for example, then overestimating a large value of d (and obtaining a large standard error) is not really a problem, in that it will still be clear that the rater discriminated well (cf. Rindskopf, 2002). The simulation also shows that estimation of the latent class sizes, particularly for small class sizes, was marginal for $N = 125$ but was good for $N = 300$. Classification also appeared to be marginal for $N = 125$ but was good for $N = 300$. The values of τ_b and Pearson's correlation suggest that, even for the sample size of $N = 125$, the classifications were useful with respect to ranking the essays.

Conclusions

The approach via latent class SDT offers a useful supplement to the latent trait models that are currently in use for essay grading and in other situations that involve construct responses. The latent class SDT model provides a simple interpretation of the rater parameters. It also has practical implications. For example, the often-cited failure of extended training to improve agreement and the ongoing concern about rater drift are only issues when one expects rater parameters to be like item parameters, in that they are viewed as unchanging characteristics of the rater. In contrast, from the perspective of SDT, problems with agreement and drift simply reflect effects of the arbitrary use of response criteria by the raters; it has long been recognized in psychology that it is difficult to control the use of response criteria, and so the difficulties noted above are expected. Further, with respect to correct classification, the locations of the response criteria are not very important (see DeCarlo, 2002a). In short, from the perspective of SDT, the focus in rater training should be on increasing discrimination, rather than on maximizing agreement or eliminating rater drift, which have not been successful.

As shown here, the approach via latent class SDT provides measures of the precision of the raters and the reliability of the classifications. The present study also provides evidence as to the validity of the classifications. An interesting finding is that the results raise questions as to whether or not one obtains more information or practical utility by assuming that the latent variable is quantitative. Of course, this conclusion is limited to the situation examined here. Some limitations were also noted, such as the fact that simply using the classifications in other analyses ignores uncertainty in class membership, and that evidence with respect to criterion validity was based only on measures of association. The utility and validity of classifications obtained from the latent class SDT model and scores obtained from latent trait models should be further compared in future studies.

Notes

[1] A reviewer noted that the way τ_b handles ties might bias it in favor of the latent class model.

[2] For the present data, using the posterior probabilities (in several ways) appeared to make little difference with respect to Pearson's correlation or τ_b (the latter was slightly smaller).

References

Agresti, A. (1990). *Categorical data analysis*. New York: Wiley.

Aitkin, M., Anderson, D., & Hinde, J. (1981). Statistical modeling of data on teaching styles. *Journal of the Royal Statistical Society A, 144,* 419–461.

Barritt, L., Stock, P. L., & Clark, F. (1986). Researching practice: Evaluating assessment essays. *College Composition and Communication, 37,* 315–327.

Bartholomew, D. J., & Knott, M. (1999). *Latent variable models and factor analysis* (3rd ed.). New York: Oxford University Press.

Blok, H. (1995). Estimating the reliability, validity, and invalidity of essay ratings. *Journal of Educational Measurement, 22,* 41–52.

Burnham, K. P., & Anderson, D. R. (2002). *Model selection and multimodel inference: A practical information-theoretic approach* (2nd ed.). New York: Springer.

Clauser, B. E. (2000). Recurrent issues and recent advances in scoring performance assessments. *Applied Psychological Measurement, 24,* 310–324.

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.

Clogg, C. C., & Manning, W. D. (1996). Assessing reliability of categorical measurements using latent class models. In A. von Eye & C. C. Clogg (Eds.), *Categorical variables in developmental research: Methods of analysis* (pp. 169–182). New York: Academic Press.

Clogg, C. C., & Shihadeh, E. S. (1994). *Statistical models for ordinal variables*. Thousand Oaks, CA: Sage.

Cohen, J. (1968). Weighted kappa: Nominal scale agreement with provision for scaled disagreement or partial credit. *Psychological Bulletin, 70,* 213–220.

Congdon, P. J., & McQueen, J. (2000). The stability of rater severity in large-scale assessment programs. *Journal of Educational Measurement, 37,* 163–178.

Cumming, A. (1990). Expertise in evaluating second language compositions. *Language Testing, 7,* 31–51.

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. (2002a). A latent class extension of signal detection theory, with applications. *Multivariate Behavioral Research, 37,* 423–451.

DeCarlo, L. T. (2002b, April). *A study of score validity for some latent class and latent trait models applied to essay grading*. Paper presented at the 2002 annual meeting of the American Educational Research Association, New Orleans, LA.

DeCarlo, L. T. (2003, July). *A multivariate extension of a latent class signal detection model*. Paper presented at the 2003 annual meeting of the Society for Mathematical Psychology, Ogden, UT.

Donoghue, J. R., & Hombo, C. M. (2000, April). *A comparison of different model assumptions about rater effects*. Paper presented at the 2000 annual meeting of the National Council on Measurement in Education, New Orleans, LA.

Engelhard, G., Jr. (1994). Examining rater errors in the assessment of written composition with many-faceted Rasch models. *Journal of Educational Measurement, 31,* 93–112.

Engelhard, G., Jr. (1996). Evaluating rater accuracy in performance assessments. *Journal of Educational Measurement, 33,* 56–70.

Gescheider, G. A. (1997). *Psychophysics: The fundamentals*. Hillsdale, NJ: Erlbaum.

Goldstein, H., & Wood, R. (1989). Five decades of item response modelling. *British Journal of Mathematical and Statistical Psychology, 42,* 139–167.

Goodman, L. A. (1974). The analysis of systems of qualitative variables when some of the variables are unobservable. Part I—a modified latent structure approach. *American Journal of Sociology, 79,* 1179–1259.

Green, D. M., & Swets, J. A. (1988). *Signal detection theory and psychophysics* (Rev. Ed.). Los Altos, CA: Peninsula Publishing.

Hagenaars, J. A., & McCutcheon, A. L. (Eds.). (2002). *Applied latent class analysis*. Cambridge, UK: Cambridge University Press.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory*. Newbury Park, CA: Sage.

Heinen, T. (1996). *Latent class and discrete latent trait models: Similarities and differences*. Thousand Oaks, CA: Sage.

Johnson, V. E. (1996). On Bayesian analysis of multirater ordinal data: An application to automated essay grading. *Journal of the American Statistical Association, 91,* 42–51.

Johnson, V. E., & Albert, J. H. (1999). *Ordinal data modeling*. New York: Springer.

Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association, 90,* 773–795.

Kendall, M. G. (1945). The treatment of ties in ranking problems. *Biometrika, 33,* 239–251.

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics, 33,* 159–174.

Langeheine, R., & Rost, J. (Eds.). (1988). *Latent trait and latent class models*. New York: Plenum.

Lin, T. H., & Dayton, C. M. (1997). Model selection information criteria for non-nested latent class models. *Journal of Educational and Behavioral Statistics, 22,* 249–264.

Linacre, J. M. (1989). *Many-faceted Rasch measurement*. Chicago, IL: MESA Press.

Lindsay, B., Clogg, C. C., & Grego, J. (1991). Semiparametric estimation in the Rasch model and related exponential response models, including a simple latent class model for item analysis. *Journal of the American Statistical Association, 86,* 96–107.

Longford, N. T. (1994). Reliability of essay grading and score adjustment. *Journal of Educational and Behavioral Statistics, 19,* 171–200.

Longford, N. T. (1995). *Models for uncertainty in educational testing*. New York: Springer-Verlag.

Macmillan, N. A., & Creelman, C. D. (1991). *Detection theory: A user's guide*. New York: Cambridge University Press.

MacMillan, P. D. (2000). Classical, generalizability, and multifaceted Rasch detection of interrater variability in large, sparse data sets. *Journal of Experimental Education, 68,* 167–190.

Marcoulides, G. A., & Moustaki, I. (Eds.). (2002). *Latent variable and latent structure models*. Mahwah, NJ: Erlbaum.

Masters, G. A. (1982). Rasch model for partial credit scoring. *Psychometrika, 47,* 149–174.

McCullagh, P., & Nelder, J. A. (1989). *Generalized linear models* (2nd ed.). New York: Chapman and Hall.

McDonald, R. P. (1999). *Test theory: A unified treatment*. Mahwah, NJ: Erlbaum.

McLachlan, G., & Peel, D. (2000). *Finite mixture models*. New York: Wiley.

Mellenbergh, G. J. (1994). Generalized linear item response theory. *Psychological Bulletin, 115,* 300–307.

Michell, J. (1997). Quantitative science and the definition of measurement in psychology. *British Journal of Psychology, 88,* 355–383.

Molenaar, P. C. M., & von Eye, A. (1994). On the arbitrary nature of latent variables. In A. von Eye, & C. C. Clogg (Eds.), *Latent variables analysis* (pp. 226–242). Thousand Oaks, CA: Sage.

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

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.

Rindskopf, D. (2002). Infinite parameter estimates in logistic regression. *Journal of Educational and Behavioral Statistics, 27,* 147–161.

Rost, J., & Langeheine, R. (Eds.). (1997). *Applications of latent trait and latent class models in the social sciences*. New York: Waxmann Münster.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monograph,* No. 17.

Sclove, S. L. (1987). Application of model-selection criteria to some problems in multivariate analysis. *Psychometrika, 52,* 333–343.

Shohamy, E., Gordon, C. M., & Kraemer, R. (1992). The effect of raters' background and training on the reliability of direct writing tests. *The Modern Language Journal, 76,* 27–33.

Swets, J. A. (1996). *Signal detection theory and ROC analysis in psychology and diagnostics: Collected papers*. Mahwah, NJ: Erlbaum.

Thissen, D. (1991). *Multilog user's guide* (Version 6.1). Chicago, IL: Scientific Software.

van der Linden, W. J., & Hambleton, R. K. (1997). Item response theory: Brief history, common models, and extensions. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory*. New York: Springer.

Verhelst, N. D., & Verstralen, H. H. F. M. (2001). An IRT model for multiple raters. In A. Boomsma, M. A. J. Van Duijn , & T. A. B. Snijders (Eds.), *Essays on item response theory* (pp. 89–108). New York: Springer.

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

Vermunt, J. K., & Magidson, J. (2000). *Latent Gold user's guide*. Belmont, MA: Statistical Innovations.

von Eye, A., & Clogg, C. C. (Eds.). (1994). *Latent variables analysis: Applications for developmental research*. Thousand Oaks, CA: Sage.

Weigle, S. C. (1998). Using FACETS to model rater training effects. *Language Testing, 15,* 263–287.

Wickens, T. D. (2002). *Elementary signal detection theory*. New York: Oxford University Press.

Wilson, T. P. (1969). A proportional-reduction-in-error interpretation for Kendall's tau-b. *Social Forces, 47,* 340–342.

Author

LAWRENCE T. DeCARLO is an Associate Professor of Psychology and Education, Department of Human Development, Box 118, Teachers College, Columbia University, New York, NY 10027-6696; decarlo@tc.edu. His primary research interests include measurement; statistics; psychometrics; and statistical models in psychology, education, and medicine.