JEM JOURNAL of EDUCATIONAL MEASUREMENT
NCME national council on measurement in education

DOI: 10.1111/jedm.12265
Journal of Educational Measurement

A Latent Class Signal Detection Model for Rater Scoring with Ordered Perceptual Distributions

Lawrence T. DeCarlo
Columbia University
Xiaoliang Zhou
Australian Council for Educational Research

In signal detection rater models for constructed response (CR) scoring, it is assumed that raters discriminate equally well between different latent classes defined by the scoring rubric. An extended model that relaxes this assumption is introduced; the model recognizes that a rater may not discriminate equally well between some of the scoring classes. The extension recognizes a different type of rater effect and is shown to offer useful tests and diagnostic plots of the equal discrimination assumption, along with ways to assess rater accuracy and various rater effects. The approach is illustrated with an application to a large-scale language test.

Constructed response (CR) items are often used in assessments and can consist of written essays, musical performances, videos of language speaking, and so on. CR items differ from supply-response items, such as multiple choice, in that raters are needed to score CR items. Thus, models of rater scoring need to consider effects introduced by the use of human raters. For example, the possible presence of various rater effects has long been recognized (e.g., Saal, Downey, & Lahey, 1980) and recently reviewed (e.g., Myford & Wolfe, 2004; Wind & Peterson, 2018; Wolfe, 2014).

Rater scoring has typically been approached by applying an item response theory (IRT) model, with raters taking the place of items. For example, a common approach is to use the Facets model, or extensions, as the rater model (e.g., Engelhard, 1994; Linacre, 1989; Myford & Wolfe, 2004). Another approach is to use a signal detection theory (SDT) model as the rater model (e.g., DeCarlo, 2005; DeCarlo, Kim, & Johnson, 2011; Patz, Junker, Johnson, & Mariano, 2002). This approach arises from a psychological conceptualization of how raters arrive at their scores. The approach brings rater scoring within the realm of SDT, which has successfully been used since the 1950s in many areas of research and applications in psychology (Green & Swets, 1988; Macmillan & Creelman, 1991; Wickens, 2002). A recent review of both IRT and SDT models for rater scoring, along with their implementation in R, is provided by Robitzsch and Steinfeld (2018).

A useful aspect of the SDT approach to rater scoring is that it provides simple measures of rater accuracy and rater effects and, more importantly, identifies the *source* of the effects. In particular, two basic components of an SDT model are a *decision* component, as reflected by the raters' use of response criteria, and a *perceptual* component, as reflected by how well the raters discriminate between latent classes defined by the scoring rubric. Thus, rater effects such as severity, leniency, centrality, extremity, end effects, and so on are all handled by the decision side of SDT, that is, the effects reflect the placement of *response criteria*. On the perceptual side, SDT provides a *discrimination parameter* that measures how well the raters discriminate between the latent classes defined by the scoring rubric; the parameter is a measure of the distance between underlying perceptual distributions and provides a measure of *rater accuracy*, with higher values indicating higher detection, and so higher classification accuracy. Another example of a rater effect from the perceptual side of the model is the *halo* effect (Myford & Wolfe, 2004; Saal et al., 1980; Wolfe, 2014), which can occur with analytic scoring. In the halo effect, responses within raters are correlated across dimensions of the analytic scoring rubric, such as "organization" and "clarity," which can be handled by a multivariate extension of the SDT rater (SDTr) model (DeCarlo, 2003). Furthermore, the SDT perspective provides a simple account of the effect, which is that the nonzero correlations arise because the raters do not *perceive* "organization" and "clarity," for example, as being independent dimensions.

The extended SDT model discussed here recognizes another possible rater effect, also from the perceptual side of the theory. In particular, the model allows for the possibility that raters may not be able to discriminate equally well between some of the latent classes, that is, they might be able to distinguish between some of the classes better or worse than others. For example, for a rubric with four classes, a rater may discriminate well between latent classes 1, 2, and 3, but may not be able to discriminate very well between the two top classes of 3 and 4. Note that the SDT models considered in Patz et al. (2002) and DeCarlo et al. (2011) restrict discrimination to be equal across the classes, as do IRT models such as the two-parameter logistic (2PL) model. The extended SDT model offered here relaxes this assumption and suggests some new and useful plots, as well as more formal ways to assess the equal distance assumption.

From a statistical perspective, the original SDTr model with equal distances is a model with *proportional odds* (for the logit link) or *proportional hazards* (for the complementary log-log link) because the discrimination parameters are equal across the response categories (see DeCarlo, 1998). As noted earlier (DeCarlo, 2002, 2005), the SDTr model is also related to latent class versions of *association models* (Agresti, 2002; Clogg & Shihadeh, 1994), *located latent class models* (Lazarsfeld & Henry, 1968; Uebersax, 1993), and *discrete latent trait IRT models* (Heinen, 1996). The model presented here relaxes the equal distance assumption and only requires that the discrimination parameters monotonically increase across the classes. Thus, the extended model only assumes that the locations of the perceptual distributions are ordered and will be referred to as an SDTr model with ordered perceptual distributions (SDTr-o). The extension is related to what has been referred to as *ordered cluster* or *ordered class models* in the latent class literature (e.g., Croon, 1990; Vermunt & Magidson, 2016), and *nonparametric IRT models* (e.g., Mokken, 1971; Sijtsma & Molenaar, 2002). Note, however, that the latent classes are considered to be ordered in both the original SDTr model and the extended SDTr-o model, in that the "equal spacing" in the original model is in the perceptual distributions Ψ and not in the latent classes η. For the SDTr-o model, the perceptual distributions are not assumed to be equally spaced but only ordered (with respect to location), and so it is a model with ordered perceptual distributions, to clarify that the new model differs with respect to an assumption about the underlying perceptual distributions, and not about the latent classes.

The next section reviews the models and shows how they can be used to examine various rater effects. The extended SDTr-o model is then introduced. Simulations are conducted to answer two basic questions. First, how good is parameter recovery for the extended model, given that it increases the number of discrimination parameters from J to $J(K{-}1)$, where J is the number of raters and K is the number of classes. Second, what happens if the true model is the (new) unequal distance model but the equal distance model is instead fit, and in particular how does this affect the detection of various rater effects and rater accuracy? The model also suggests new diagnostic plots, which provide detailed information about each rater, and ways to test the equal distance assumption.

Background

The Latent Class SDTr Model

The latent class signal detection model (DeCarlo, 2002, 2005) can be applied to the situation where several raters score a CR item and is,

$$p(Y_j \leq k|\eta) = F(c_{jk} - d_j\eta), \tag{1}$$

where Y_j is the response of jth rater, which is a discrete score k with K classes (to simplify notation, the number of response classes is assumed to be the same across different raters and items, as is often the case in practice, although this need not be the case), η is a latent categorical variable with M values, such as scores from 0 to $M - 1$ (or centered scores of $m - (M + 1)/2$), F is a cumulative distribution function (CDF) for a location-family of distributions, such as the logistic or normal, d_j is a discrimination parameter for the jth rater, c_{jk} are $K - 1$ strictly ordered response criteria, $c_{j1} < c_{j2} < \ldots < c_{j,K-1}$, for the jth rater and kth response category, with $c_{j0} = -\infty$ and $c_{jK} = \infty$. Note that d_j and c_{jk} are scaled differently depending on the link function (the inverse of F), for example, for the logit link, the square root of the variance of the logistic distribution, $\pi^2/3$ is used as a scale factor, whereas 1 is used for the probit link, and $\pi^2/6$ for the complementary log-log link.

The model follows directly from a signal detection conceptualization of rater scoring. To start, it is assumed that there are M latent classes of η that are defined by the scoring rubric. For example, the rubric for scoring SAT essays specifies four ordered classes with descriptions of qualities of essays in each class. From the reading scoring guide, part of the description of scores of 1, 2, 3, and 4 are that a 1 shows “little or no” comprehension, 2 shows “some” comprehension, 3 shows “effective” comprehension, and 4 shows “thorough” comprehension. These qualities, along with others, are assumed to define four ordered latent “true” classes, η, that CRs (e.g., essays) can be classified into. The raters’ task is to detect these M latent classes using responses Y from 1 to K where $K = M$ for the SDTr model (the more general latent class SDT model does not have this restriction). The only assumption about the M latent classes of η is that they are ordered.

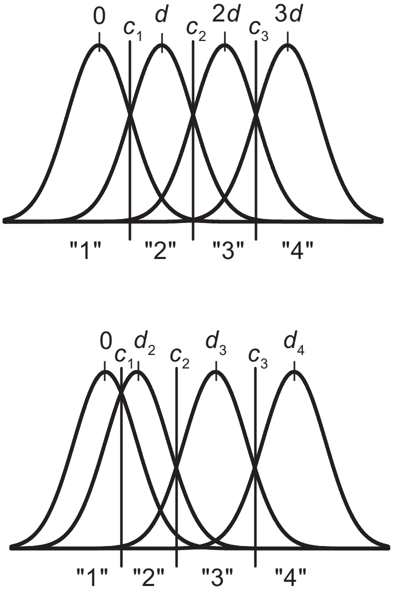

Figure 1. Perceptual distributions and parameters *c* and *d* for signal detection theory with equal spacing (top) and unequal spacing (bottom), with four latent classes.

The view in SDT is that the decision of an observer is based in part on his or her *perception* of a stimulus, which can be represented by a latent continuous variable, Ψ, with perceptual distributions being represented by probability distributions. In the application to rater scoring, for example, it is assumed that, when a rater reads an essay from one of the latent classes of η, he or she has a perception of the overall quality of the essay (for holistic scoring). As shown in the top panel of Figure 1 (the bottom panel is for the extended model discussed below), the perception is drawn from one of several perceptual distributions that are associated with each latent class; Figure 1 shows the situation where there are four latent classes defined by the scoring rubric, and so there are four associated perceptual distributions. Thus, a psychological part of SDT is the introduction of a latent continuous variable, Ψ, which represents the raters' perceptions, and which intervenes between the observed ordinal response Y and the latent ordinal categorical variable η.

Suppose, for example, that Class 1 of the scoring rubric defines a class of essays that show "little or no" comprehension and Class 2 defines a (better) class of essays that show "some" comprehension. According to SDT, when a rater reads an essay from Class 1, he or she has a perception of the overall quality of the essay, which is a realization from the first probability distribution shown in Figure 1. This is a basic idea in SDT and psychophysics, which goes back to Fechner (1860/1966) and Thurstone (1927), and accounts for randomness in responses that was observed early on in psychophysics (i.e., it is due to randomness in psychological perceptions Ψ).

The rater then compares the perceptual realization to response criteria; if it is below the first criterion, c_1, the essay is scored as "1," if it is above the first criterion but below the second criterion, c_2, the essay is scored as "2," and so on. If an essay is from Class 2 (e.g., some comprehension), then the perceptual realization comes from the second distribution shown in Figure 1, and so on for the other classes.

Rater accuracy. As shown in Figure 1, it is assumed that, as the latent class η moves from a lower class to a higher class, the location of the perceptual distribution also shifts upward, which simply reflects that raters tend to perceive essays from higher classes as being of higher quality. The amount of shift is given by the *discrimination* parameter d_j, which provides a measure of *rater accuracy* (for the jth rater) and so it is of primary interest. Raters with higher values of d_j discriminate better and their classifications are more accurate, whereas smaller values of d_j indicate that a rater cannot discriminate as well between the latent classes.

Although one can allow the distances to vary between the distributions (as done here), it was suggested earlier that a simplifying assumption is to assume that the spacing is equal (DeCarlo, 2002, 2005), which greatly reduces the number of parameters. That is, as shown in Figure 1, the distance between adjacent distributions is d in all cases, and so the distributions are equally spaced. A simple way to implement this restriction is to score η from 0 to $K - 1$ (where $K = M$), and so the first distribution is at 0, the second at d, the third at $2d$, and so on. Note that the scoring of η is done solely to implement the equal distance restriction on the parameter d, it does not mean that η is assumed to be quantitative—as noted above, the latent classes η are only assumed to be ordered, the equal spacing is in the perceptual distributions Ψ, not η. It should also be noted that there can be estimation advantages to using centered scoring, $k - (K + 1)/2$, given that this results in smaller values of the parameters.

Rater effects. An interesting aspect of the SDT representation is that it suggests natural reference points for the response criteria. For example, in the case of one rater and two classes, placing the criterion at the intersection point of the two underlying distributions maximizes the proportion correct. The situation gets more complex with more raters and distributions, but in any case, it was suggested earlier that the intersection points provide a useful reference point for the criteria and allow one to define various rater effects (DeCarlo, 2005; DeCarlo et al., 2011).

For the SDTr model, the equal distance restriction means that the intersection locations are also equally spaced, as shown in the top panel of Figure 1. The intersection locations, however, will vary across raters depending on d_j. For example, with the first distribution at zero, if d_1 for Rater 1 is 2, then the intersection points are at 1, 3, 5, and so on, whereas if d_2 for Rater 2 is 4, then the intersection points are at 2, 6, 10, etc. To account for this and make the locations comparable across raters, it was earlier suggested to rescale by dividing the estimates of c_{jk} by $(K - 1) \times \hat{d}_j$ so that the highest distribution is always at 1 (the lowest is at 0), which gives "relative" criteria locations (DeCarlo, 2005). This allows one to compare the raters' criteria locations both to the intersection points and to each other, all in one plot. Furthermore, the plot can reveal any and all rater effects due to response criteria usage, such as severity,

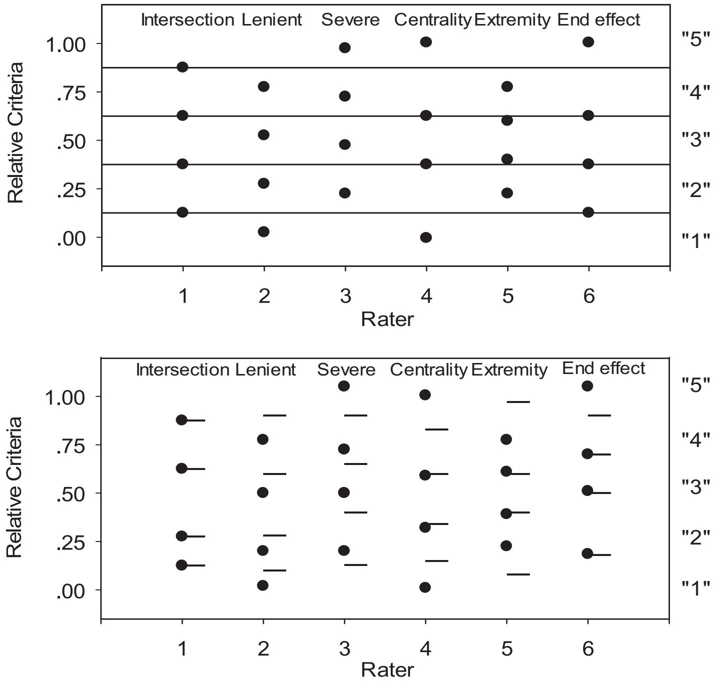

Figure 2. Plots of the relative criteria locations (filled circles), 1–5 response, showing various rater effects. The top panel is for the equal distance model, the bottom panel is for the unequal distance model. The lines show the underlying distribution intersection points.

leniency, centrality, extremity, and other patterns, such as "end effects," or patterns without names.

For example, the top panel of Figure 2 shows a plot of relative criteria for several raters exhibiting different types of rater effects (the bottom panel is for the extended model and is discussed below). In this case, there are four criteria that delineate responses of 1–5 (as in the example presented below). The solid lines represent the intersection point locations and the filled circles show the relative criteria locations for each rater. Figure 2 shows that Rater 1's criteria locations, for example, are all at the intersection points. In contrast, Rater 2's criteria are all shifted downward from the horizontal lines (and Rater 1). This corresponds to a shift of the criteria shown in Figure 1 to the left, and so the second rater is more "lenient" and tends to give higher scores. The third rater shows an upward shift, which corresponds to a rightward shift of the criteria shown in Figure 1, and so the third rater is more "severe" and tends to give lower scores. For Rater 4, the first criterion is lower and the last criterion is higher. This corresponds to the first criterion being shifted to the left and the last criterion being shifted to the right in Figure 1, and so the rater tends to give fewer responses of 1 and 5, and more responses of 2, 3, and 4, which is "centrality." Rater 5's first and last criterion are shifted up and down, respectively, which means that Rater 5 tends to overuse the first and last classes of "1" and "5," which is "extremity." The sixth rater in Figure 2 shows what can be referred to as an "end effect" or "restriction of the range," in that only the last criteria is higher, and so this rater tends to not use "5" on the 1–5 scale.

In addition to the examples shown in Figure 2, the criteria locations can show other types of patterns that may not currently have labels. Thus, by estimating the relative criteria locations for each rater and by plotting them, commonly found rater effects such as leniency/severity and centrality/extremity are revealed. Note that, given a definition of a rater effect as a particular type of pattern, as in Figure 2, a summary measure that detects the effect can also be derived. For example, one can look at the *average deviation* of each estimated criteria from the intersection points as a measure of severity/leniency; this would be zero for Rater 1, indicating no severity/leniency, negative for Rater 2, indicating that the rater was (relatively more) lenient, and positive for Rater 3, indicating that the rater was severe. Robitzsch and Steinfeld (2018) recently did something similar using the hierarchical rater signal detection model (HRM-SDT) (they computed the average of the criteria) and found a correlation of .99 with the severity parameter of the Patz et al. model, and so this derived measure from the SDT parameters clearly detects leniency/severity. One can also formulate measures of centrality, extremity, and other rater effects. Thus, given a definition of a rater effect, a measure can be computed from the SDT parameter estimates (and the model does not need to be modified to obtain the measure, which avoids a proliferation of models for rater effects). Note that a disadvantage of the "measure" approach is that it limits what one can detect only to what the measure is made to detect, whereas the plot noted above can reveal any and all types of rater effects (due to criteria placement), as shown in Figure 2.

Relation to the Patz et al. Model

The rater model offered by Patz et al. (2002), for the first level of the hierarchical rater model (HRM), can be written (for a single item) as

$$p(Y_j = k|\eta) \propto \exp\left\{-\frac{1}{2\psi_j^2}[k - (\eta + \phi_j)]^2\right\}, \tag{2}$$

where Y_j is the response of jth rater to the item, with the response being a discrete score k with K classes, η is a latent a categorical variable (ideal ratings), ψ_j^2 is a variance parameter for rater j, and ϕ_j is a rater severity parameter. The inverse of the variance parameter provides a measure of rater precision, $\tau_j = 1/(2\psi_j^2)$, whereas positive values of ϕ_j indicate severity and negative indicate leniency (zero indicates no bias), in that the rater tends to give low or high scores, respectively. As noted by Patz et al., the rater model is a discrete signal detection model; note that the model applies to the probabilities for each response category, whereas Equation 1 is for cumulative probabilities.

It is useful to rearrange terms slightly and rewrite the Patz et al. model as

$$p(Y_j = k|\eta) \propto \exp\{-\tau_j[(k + \phi_j) - \eta]^2\}. \tag{3}$$

Equation 3 shows that τ_j serves the same role as the discrimination parameter a_j in IRT or d_j in SDT, whereas ϕ_j is analogous to the difficulty parameter b_j in IRT.

The term $(k + \phi_j)$ plays a role similar to the criteria c_{jk} in SDT—if the criteria are restricted to be equally spaced across the K classes, and a mean shift in the criteria locations is allowed for, then c_{jk} can be replaced by c_j, the mean shift for each rater, which reflects severity or leniency, in the same manner as ϕ_j in Equation 3. This reflects that the rater model of Patz et al. can only detect severity or leniency, and not centrality/extremity and other effects that appear in real-world data (see DeCarlo et al., 2011); for example, centrality and extremity can give a severity of zero, in which case they will not be detected by a severity (bias) parameter. On the other hand, the criteria parameters c_{jk} of the SDTr model of Equation 1 can capture many different types of rater effects, as shown above.

SDT with Ordered Perceptual Distributions

A simplifying assumption made in both the Patz et al. (2002) and DeCarlo et al. (2011) versions of the SDTr model is that the raters are assumed to discriminate equally between all of the latent classes. That is, both models have only one discrimination parameter per rater, τ_j, in the first model and d_j in the latter model. The current extension allows discrimination to vary across the latent classes and replaces d_j with d_{jm} where m indicates the latent class. This possibility was recognized earlier (e.g., DeCarlo, 2002, 2005; Patz et al., 2002) but has not been developed or examined in any detail.

The lower panel of Figure 1 shows the theory with distributions that are unequally spaced. In this case, the distributions associated with Classes 1 and 2 are closer together than the other distributions, and so the rater cannot discriminate as well between Classes 1 and 2 as compared to the other classes. Compared to the equal spacing model shown in the top panel of Figure 1, the common (across the classes) distance d_j is now replaced with class-specific d_{jm}, which is the distance of each distribution from the first distribution, which is at zero. It is useful to look at the differences between the d_{jm}s, that is $d_{jm+1} - d_{jm}$ for $m = 1$ to $M - 1$, which reflect how well the rater discriminates between adjacent classes; any inequality of the distances between the distributions will then be readily apparent (e.g., see Tables 1 and 2 below).

Allowing for unequal spacing allows the model to account for several realistic possibilities. For example, a rater may discriminate well between essays in the top three classes, say essays that show “some,” “effective,” or “thorough” comprehension, but may discriminate poorly between essays in the first two classes, say essays that only show “little or no” versus “some” comprehension. This means that the first two distributions are closer together, as shown in the lower panel of Figure 1. Other possibilities are that the rater cannot discriminate well on the high end, say between “effective” and “thorough” essays, and so the top distributions are closer together, or that the rater can discriminate the best and worst essays, but not those in the middle, and so the middle distributions are closer together, and so on.

Relaxing the equal distance assumption gives a version of the SDTr model where the underlying distributions are not assumed to be equally spaced across the latent classes. The model extends Equation 1 by replacing d_j with d_{jm} (for category-specific discrimination parameters) and treating η as having unordered categories (nominal).

The nominal aspect of the latent variable is implemented by using *M* latent dummy variables η_m, and so the model can be written as

$$p(Y_j \leq k|\eta_m) = F\left(c_{jk} - \sum_{m=1}^{M} d_{jm}\eta_m\right), \tag{4}$$

where *M* is the number of latent classes, which for the rater model is the same as the number of response categories, $M = K$. For example, for four latent classes, the four rows of η_1, η_2, η_3, and η_4 (corresponding to the four distributions) are (0,0,0,0), (0,1,0,0), (0,0,1,0), and (0,0,0,1). Note that one can also get the discrimination parameters in difference form directly by coding the four latent eta's as (0,0,0,0), (0,1,0,0), (0,1,1,0), and (0,1,1,1).

The first distribution serves as the reference and so $d_{j1} = 0$. A positive monotonicity constraint is placed on the d_{jm}s so that they are ordered, that is, $d_{j1} \leq d_{j2} \ldots \leq d_{jM}$, which in turn means that the underlying perceptual distributions have ordered locations. Note that without the monotonicity constraint, the perceptual distributions are not ordered (they are nominal); a nominal latent class SDT model and corresponding association models were fit and compared to other models in DeCarlo (2002). Given that Equation 4 only assumes that the perceptual distributions are ordered, rather than equally spaced as in Equation 1, it is referred to as an SDTr-o.

Rater Effects

The SDTr-o model of Equation 4 raises questions about how to present the relative criteria plots shown earlier for the SDTr model. Note that the lines representing the intersection points in the top panel of Figure 2 will no longer be equally spaced, because of the (possibly) unequal distances. One can still scale the top distribution to be at one, as for the relative criteria, and the bottom to be at zero, however all the intersection points will have different spacings across the raters, if d_{jm} varies across the classes. To address this, the approach here is to show the different intersection point locations separately for each rater, using short line segments as shown in the lower panel of Figure 2, with filled circles showing the relative criteria estimates as before.

Note that the rater effects shown in the lower panel of Figure 2 appear essentially the same as those shown in the top panel, except that one has to use the line segments for each rater for comparison. For example, it is apparent in the lower panel that Rater 1's relative criteria estimates are all at the intersection points (i.e., on the line segments), whereas Rater 2's relative criteria are all below the line segments, and so Rater 2 is more lenient than Rater 1. Rater 3's relative criteria are all above the line segments, and so this rater is relatively more severe; Rater 4's first and last relative criteria estimates are below and above the line segments, respectively, and so this rater tends to underuse the end categories of "1" and "5," which is centrality, and so on for other effects. Thus, the modified relative criteria plot for the SDTr-o model can be used in exactly the same manner as the relative criteria plot for the SDTr model.

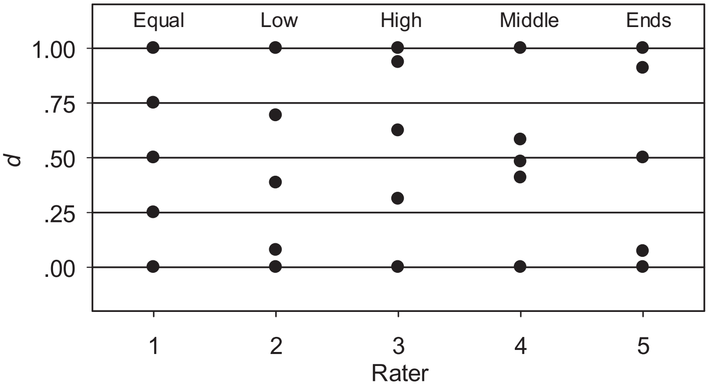

Figure 3. Discrimination plots, 1–5 response, with various effects; lines show the locations of five equally spaced distributions, filled circles show the distribution locations.

Rater accuracy

In addition to the extended relative criteria plot discussed above, it is also useful to provide a plot of the estimates of d_{jm} for each rater, which provides a convenient way to visually assess the equal distance assumption. Figure 3 shows this type of *discrimination plot* for the situation with five latent classes. Given that the spacing between the distributions is of interest, the first distribution is set at zero and the last distribution at one, and horizontal lines show the equal spacing locations.

Figure 3 shows that the perceptual distributions for Rater 1 are equally spaced. Rater 2, however, shows poor discrimination between Classes 1 and 2 (i.e., the bottom two filled circles are close together), but good discrimination between the rest of the classes; Rater 3 shows poor discrimination between Classes 4 and 5, but good discrimination elsewhere; Rater 4 shows good discrimination between Classes 1 and 2, and 4 and 5, but poor discrimination between the middle classes of 2, 3, and 4; Rater 5 shows good discrimination between the middle classes of 2, 3, and 4, but poor discrimination between the end classes, that is, between 1 and 2, and 4 and 5. Of course there are many other possibilities, Figure 3 simply illustrates a few that can be easily classified. A discrimination plot such as Figure 3 is useful to see if any raters show discrimination problems between some of the classes and to get an idea of whether or not the equal spacing assumption is reasonable.

A test of equal *d*

Given that the SDTr model with equal d_j is nested in the SDTr-o model with unequal d_{jm} (Equation 4), the assumption of equal d_j across the M classes can be tested by using a likelihood ratio (LR) test, in the same manner that the LR test is used to test the proportional odds assumption in logistic regression models (Agresti, 2002). Thus, one fits both models, the restricted SDTr model and the unrestricted SDTr-o model, and the LR statistic is computed by subtracting the minus two log likelihoods (−2LL), that is, $\text{LR} = -2\text{LL}_{\text{restricted}} - (-2\text{LL}_{\text{unrestricted}})$. This can be tested against a chi-squared distribution with degrees of freedom given by the difference in the number of parameters between the models, which in this case is $J(K - 1) - J = J(K - 2)$. Note that the log likelihoods used here for are from posterior mode estimation. We examine the performance of the LR statistic in simulations presented below.

Estimation

A well-known problem in latent class analysis (with maximum likelihood estimation, MLE) is that one frequently encounters *boundary problems* (e.g., Clogg & Eliason, 1987; DeCarlo, 2011; Vermunt & Magidson, 2016), in that parameter estimates go to the "boundaries," such as positive/negative infinity, and the standard errors tend to be large or indeterminate. Several authors have suggested using *posterior mode estimation* (PME) to deal with this problem (e.g., DeCarlo, 2011; DeCarlo et al., 2011; Galindo-Garre & Vermunt, 2006; Maris, 1999; Vermunt & Magidson, 2016). PME is a partly Bayesian approach in that, rather than deriving the full posterior, one simply finds the mode of the posterior. This is less computationally intense than a full Bayesian analysis and smooths the offending parameters away from the boundary. PME was used here with a Bayes constant of one, which is analogous to adding one observation to the data; earlier simulations showed good parameter recovery with PME and Bayes constants of one for the SDTr model (e.g., DeCarlo, 2008, 2010; also see DeCarlo et al., 2011, for a review). The SDTr and SDTr-o models were also implemented in a full Bayesian analysis and the results were compared to those obtained with PME.

Simulations

A basic question is how well the parameters are recovered, given that the SDTr-o model has $J(K-1)$ discrimination parameters, whereas the original SDTr model only has J discrimination parameters. This is examined in two simulations—one for a fully crossed design, where each essay is scored by every rater, and one for an incomplete design, where each essay is scored by only a subset of the raters. For the fully crossed design, there were 10 raters, a 1–4 scoring rubric, and 1,000 examinees (e.g., essays). For the incomplete design, there were 10 raters, a 1–4 response, and 2,000 examinees, but each essay was scored by only two of the 10 raters, with the rest of the values set to missing, and so 80% of the data is missing by design, as compared to the fully crossed design. A balanced incomplete block design was used to assign rater pairs, except that the design is slightly unbalanced for 2,000 examinees (more realistic) in that 2,025 examinees would be needed for the design to be perfectly balanced (see DeCarlo, 2010). For the current simulations, raters generally scored around 400 essays (which reflects the slight lack of balance; for 2,025 essays, each rater would score exactly 405 essays each). We examine parameter recovery using PME in Latent Gold for both fully crossed and incomplete designs.

The typical incomplete design used in large-scale testing has many raters, say 30 or more, but because of the large number of CRs (e.g., essays) to be scored, typically thousands, each essay is usually scored by only two raters; the SAT is an example. If there are only two raters per essay, then the SDTr model is not identified without additional parameter constraints (e.g., equal d_j across the raters). Here it is noted that, in a Bayesian approach, the use of priors can help with identification, though one must exercise caution (e.g., Neath & Samaniego, 1997). In the Bayesian context, the “identification” issue has to do with how much information is available in the data, and whether the data (i.e., the likelihood) add something to the priors (see DeCarlo, 2011, for an example of this problem in a related context). The partly Bayesian PME approach used here deals with this problem; it has previously been shown in prior simulations that PME with Bayes constants of one gave reasonable parameter recovery for incomplete designs with two raters (DeCarlo, 2010), though more work on this needs to be done. The present study adds to this information.

The data were generated according to the SDTr-o model of Equation 4 using SAS macros written by the first author. For both the fully crossed and incomplete designs, eight of the 10 raters showed one of the four rater effects depicted in Figure 2, that is, leniency, severity, centrality, or extremity. In addition, eight of the 10 raters had unequal discrimination across some of the classes, and in particular, examined were basic situations where the rater had poor discrimination either on the low end (i.e., between 1 and 2), the high end (between 3 and 4), the middle classes (between 2 and 3), or the end classes (between 1 and 2, and between 3 and 4).

For each condition, 100 data sets were generated and Latent Gold 5.1 (LG5.1, Vermunt & Magidson, 2016) was used to fit Equation 4; PME was used with Bayes constants of 1. The means, bias, percent bias, and mean squared error (MSE) across the 100 replications were examined. Also examined was the performance of the LR test. Note that Latent Gold provides a check of local identification by examining a necessary condition (in addition to the necessary condition of having equal or fewer parameters than observed response patterns), which is that the rank of the Jacobian matrix must be greater than or equal to the number of parameters; if it is less than the number of parameters, then the model is not identified. The check is performed for both PME and MLE. The test indicated identification problems for the balanced incomplete block (BIB) design for MLE (but not for the fully crossed design), as expected, but not for PME. Note that a failure to converge occurred for a few of the 100 BIB data sets; this was addressed by increasing both the expectation-maximization and Newton Raphson iterations, in which case convergence occurred.

Results

Parameter recovery. Table 1 shows the population parameters, parameter estimates, bias, percent bias, and MSE for the fully crossed condition; note that the *d*s are presented as differences (between adjacent distributions), in which case unequal values are more readily apparent. The table shows that recovery of the SDTr-o model parameters is excellent, in that the estimated parameters are all close to the population values, generally with small bias, and the MSE is small. The latent class sizes are also well recovered. Thus, for a fully crossed design, estimation appears to be excellent, and the parameter estimates pick up both the various rater effects and the unequal discrimination across classes.

Table 2 shows results for the incomplete design. In this case, estimation is not as good as for the fully crossed design, in that the bias and MSE tend to be larger, reflecting the lower amount of information available in the data. However, the

Table 1

Parameter Recovery for the SDTr-o Model, 1–4 Response, Fully Crossed, N = 1,000

Parameter	Value	Estimate	Bias	%Bias	MSE
d_{11}	1.00	1.02	.02	2.47	.05
$d_{12}-d_{11}$	1.00	.99	−.01	1.19	.02
$d_{13}-d_{12}$	1.00	.98	−.02	1.81	.05
d_{21}	1.00	1.01	.01	1.09	.06
$d_{22}-d_{21}$	2.00	2.05	.05	2.34	.03
$d_{23}-d_{22}$	2.00	1.95	−.05	2.73	.10
d_{31}	3.00	3.04	.04	1.49	.14
$d_{32}-d_{31}$	3.00	3.03	.03	.94	.05
$d_{33}-d_{32}$	1.00	.96	−.04	3.85	.03
d_{41}	1.00	1.00	<.01	.39	.06
$d_{42}-d_{41}$	4.00	4.04	.04	.90	.05
$d_{43}-d_{42}$	1.00	1.01	.01	.53	.04
d_{51}	5.00	5.31	.31	6.27	1.24
$d_{52}-d_{51}$	1.00	.99	−.01	1.33	.03
$d_{53}-d_{52}$	5.00	5.40	.40	7.94	1.19
d_{61}	5.00	5.30	.30	5.99	.97
$d_{62}-d_{61}$	5.00	4.98	−.02	.42	.08
$d_{63}-d_{62}$	5.00	5.73	.73	14.50	1.25
d_{71}	1.00	1.07	.07	6.82	.18
$d_{72}-d_{71}$	4.50	4.49	−.01	.29	.07
$d_{73}-d_{72}$	4.50	4.59	.09	1.97	.42
d_{81}	3.50	3.51	.01	.25	.16
$d_{82}-d_{81}$	3.50	3.54	.04	1.24	.05
$d_{83}-d_{82}$	1.00	1.00	<.01	.02	.29
d_{91}	1.00	.98	−.02	2.12	.03
$d_{92}-d_{91}$	2.50	2.54	.04	1.64	.03
$d_{93}-d_{92}$	1.00	.99	−.01	.60	.05
d_{101}	1.50	1.50	<.01	.06	.06
$d_{102}-d_{101}$	1.00	1.03	.03	3.01	.03
$d_{103}-d_{102}$	1.50	1.49	−.01	.35	.06
c_{11}	.50	.51	.01	2.99	.04
c_{12}	1.50	1.52	.02	1.54	.04
c_{13}	2.50	2.52	.02	.76	.04
c_{21}	−.50	−.50	<.01	.89	.03
c_{22}	1.00	1.01	.01	.80	.04
c_{23}	3.00	3.03	.03	1.06	.05
c_{31}	2.50	2.53	.03	1.18	.11
c_{32}	5.50	5.54	.04	.80	.13
c_{33}	7.50	7.55	.05	.71	.15
c_{41}	−.50	−.51	−.01	2.40	.03
c_{42}	3.00	3.03	.03	.91	.05
c_{43}	6.50	6.54	.04	.66	.08
c_{51}	4.50	4.81	.31	6.83	1.20
c_{52}	5.50	5.80	.30	5.42	1.21

(Continued)

Table 1
Continued

Parameter	Value	Estimate	Bias	%Bias	MSE
c_{53}	6.50	6.80	.30	4.66	1.20
c_{61}	.50	.50	<.01	.20	.04
c_{62}	7.50	7.79	.29	3.81	.92
c_{63}	14.50	14.98	.48	3.29	1.99
c_{71}	2.50	2.57	.07	2.62	.15
c_{72}	5.25	5.31	.06	1.18	.19
c_{73}	9.75	9.88	.13	1.35	.50
c_{81}	−.25	−.24	.01	3.25	.03
c_{82}	3.25	3.25	<.01	.12	.16
c_{83}	5.50	5.52	.02	.32	.17
c_{91}	.50	.49	−.01	1.04	.03
c_{92}	2.25	2.26	.01	.64	.04
c_{93}	4.00	4.01	.01	.37	.05
c_{101}	1.50	1.51	.01	.84	.05
c_{102}	2.00	2.01	.01	.70	.05
c_{103}	2.50	2.51	.01	.39	.05

Latent Class Sizes

Parameter	Value	Estimate	Bias	%Bias	MSE
Class 1	.15	.15	<.01	.52	<.01
Class 2	.35	.35	<.01	.01	<.01
Class 3	.35	.35	<.01	.29	<.01
Class 4	.15	.15	<.01	.18	<.01

estimates are still useful for detecting effects in some cases. For example, Rater 3 has poor discrimination between the top two distributions (the spacing between the four distributions is 3, 3, and 1, respectively), and the parameter estimates of 2.76, 2.91, and 1.13 clearly pick this up. However, in other cases the estimates of d_j miss the effect or suggest a different effect. The same holds for the criteria estimates, estimation is again poorer, but some of the rater effects and discrimination differences still appear. For example, a plot of the criteria estimates clearly shows that Rater 2 is “lenient,” that Rater 3 is “strict,” and so on.

An estimation problem appeared for the SDTr-o model in the BIB condition (but not in the fully crossed condition), which is that some of the distances between adjacent distributions were fixed to zero. This occurs because of the monotonicity constraint—when an estimate goes below zero during the iterations it is fixed to zero. This is reported as an “activated constraint” in Latent Gold and the number of parameters for the information criteria and the degrees of freedom for the LR test are adjusted.

LR test: Type I error and power. Table 3 shows comparisons of Type I error rates and power for the LR test of equal discrimination between the classes; SDTr

Table 2

Parameter Recovery for the SDTr-o Model, 1–4 Response, Incomplete (BIB), N = 2,000

Parameter	Value	Estimate	Bias	%Bias	MSE
d_{11}	1.00	1.02	.02	2.11	.79
$d_{12}-d_{11}$	1.00	.84	−.16	15.99	.32
$d_{13}-d_{12}$	1.00	1.09	.09	8.68	.59
d_{21}	1.00	1.35	.35	34.99	.87
$d_{22}-d_{21}$	2.00	1.74	−.27	13.25	.77
$d_{23}-d_{22}$	2.00	2.51	.51	25.50	2.01
d_{31}	3.00	2.76	−.24	7.88	1.38
$d_{32}-d_{31}$	3.00	2.91	−.09	2.86	.84
$d_{33}-d_{32}$	1.00	1.13	.13	12.75	.83
d_{41}	1.00	1.26	.26	26.30	1.31
$d_{42}-d_{41}$	4.00	4.19	.19	4.63	1.13
$d_{43}-d_{42}$	1.00	1.14	.14	13.81	.95
d_{51}	5.00	3.47	−1.53	30.64	3.66
$d_{52}-d_{51}$	1.00	.62	−.38	38.35	.45
$d_{53}-d_{52}$	5.00	3.93	−1.07	21.43	2.78
d_{61}	5.50	4.35	−1.15	20.93	2.81
$d_{62}-d_{61}$	5.50	4.54	−.96	17.54	2.52
$d_{63}-d_{62}$	5.50	4.99	−.51	9.29	1.98
d_{71}	1.00	1.62	.62	61.80	2.86
$d_{72}-d_{71}$	4.50	4.34	−.16	3.49	1.34
$d_{73}-d_{72}$	1.00	4.13	3.13	313.02	11.32
d_{81}	2.50	3.38	.88	35.34	2.56
$d_{82}-d_{81}$	3.50	3.42	−.08	2.36	.93
$d_{83}-d_{82}$	3.50	1.24	−2.26	64.51	6.55
d_{91}	2.50	1.16	−1.34	53.70	2.60
$d_{92}-d_{91}$	1.00	2.31	1.31	130.97	2.18
$d_{93}-d_{92}$	2.50	1.12	−1.38	55.14	2.65
d_{101}	1.50	1.74	.24	15.97	1.42
$d_{102}-d_{101}$	1.50	.75	−.75	49.89	.96
$d_{103}-d_{102}$	1.00	1.71	.71	70.61	2.20
c_{11}	.50	.40	−.10	19.98	.37
c_{12}	1.50	1.43	−.07	4.51	.45
c_{13}	2.50	2.48	−.02	.87	.47
c_{21}	−.50	−.50	<.01	.18	.15
c_{22}	1.00	1.10	.10	10.49	.27
c_{23}	3.00	3.24	.24	7.89	.44
c_{31}	2.50	2.04	−.46	18.60	1.24
c_{32}	5.50	5.24	−.26	4.67	1.65
c_{33}	7.50	7.36	−.14	1.92	1.92
c_{41}	1.50	−.59	−2.09	139.51	4.60
c_{42}	3.00	3.31	.31	10.38	1.22
c_{43}	4.50	7.12	2.62	58.26	8.49
c_{51}	4.50	2.72	−1.78	39.62	4.44
c_{52}	7.50	3.75	−3.75	49.98	15.43

(Continued)

Table 2
Continued

Parameter	Value	Estimate	Bias	%Bias	MSE
c_{53}	10.50	4.78	−5.72	54.49	34.24
c_{61}	.75	−.08	−.83	110.27	1.13
c_{62}	6.25	6.55	.30	4.85	2.43
c_{63}	11.75	13.23	1.48	12.62	6.11
c_{71}	.50	2.99	2.49	497.39	7.39
c_{72}	3.25	5.85	2.60	80.14	8.56
c_{73}	6.00	10.16	4.16	69.35	19.99
c_{81}	.25	−.67	−.92	366.01	1.18
c_{82}	4.25	2.86	−1.39	32.72	3.63
c_{83}	8.75	5.29	−3.46	39.57	13.90
c_{91}	−.75	.48	1.23	164.40	1.84
c_{92}	3.00	2.30	−.70	23.45	.97
c_{93}	6.75	4.05	−2.70	39.99	7.92
c_{101}	.75	1.56	.81	107.90	1.58
c_{102}	2.25	2.08	−.17	7.36	.98
c_{103}	3.50	2.60	−.90	25.60	1.77

Latent Class Sizes

Parameter	Value	Estimate	Bias	%Bias	MSE
Class 1	.15	.20	.05	33.59	.01
Class 2	.35	.30	−.05	13.86	.01
Class 3	.35	.32	−.03	9.71	<.01
Class 4	.15	.18	.03	21.42	<.01

models with from three to six latent classes were examined. The Type I error rate is assessed by fitting the SDTr and SDTr-o models to data generated with equal distances, that is the SDTr model, and seeing how often (for 100 replications) the LR test incorrectly rejects (at the .05 level) the null hypothesis of equal distances. Power is assessed by fitting the models to data generated with unequal distances, that is, the SDTr-o model, and seeing how often the LR statistic correctly rejects the null hypothesis of equal distances. The rater effects were kept the same across the equal and unequal distance conditions.

The Type I error rate is of particular interest, given that one does not want to fit the more complex SDTr-o model if the equal distance assumption of the simpler SDTr model is reasonably satisfied, and so one wants Type I error rates that are at or below the nominal level. The top part of Table 3 shows results for fully crossed designs with from three to six latent classes. The Type 1 error rate of the LR test in every case is at or below the nominal .05 level, and the power is 1. Thus, the LR test appears to be useful for testing the equal distance assumption for fully crossed designs. For the incomplete (BIB) design, the Type I error rates of the LR test are again close to the nominal .05 level in every case, which is important, however the power is lower and

Table 3
Type I Error Rates and Power for LR Test of Equal Distances, Three to Six Class Models

Number of Classes:	Three	Four	Five	Six
Fully Crossed Design				
Type I Error	.02	.05	.00	.00
Power	1.00	1.00	1.00	1.00
BIB Design				
Type I Error	.04	.02	.04	.08
Power	.63	.79	.76	.63

Note. N = 1,000 for fully crossed and 2,000 for balanced incomplete block (BIB), results are based on 100 replications per condition.

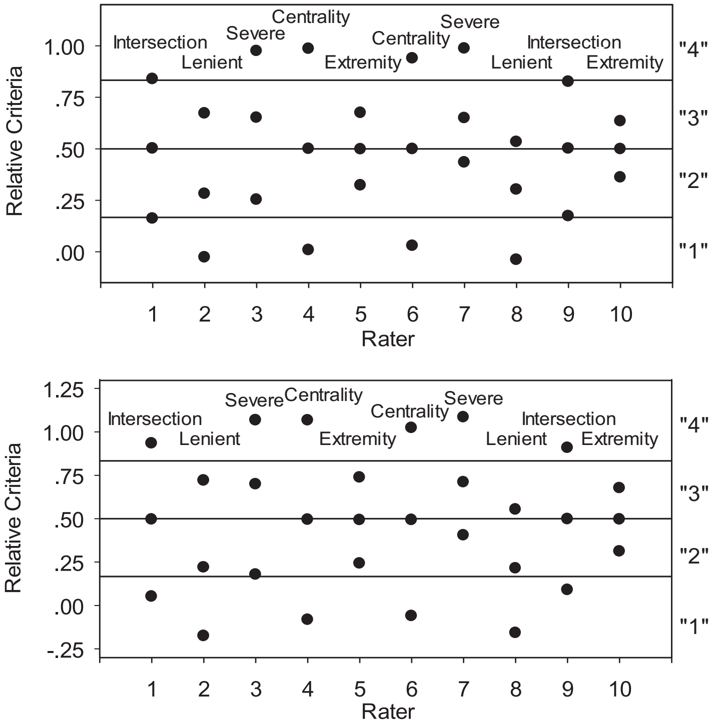

Figure 4. Plots of the relative criteria average estimates for a fit of the equal distance model to unequal distance data, 1–4 response, for fully crossed (top) and BIB (bottom).

ranges from .63 to .79. Overall, Table 3 shows that the LR test of equal distances appears to be quite useful for both complete and incomplete designs.

Robustness. It is also of interest to see what happens when the equal distance SDTr model is fit to data generated according to the unequal distance SDTr-o model; can one still detect rater effects? Figure 4 shows this type of plot for (average) results

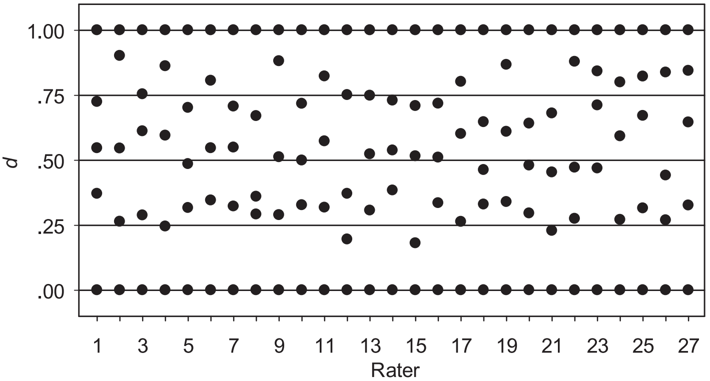

Figure 5. A discrimination plot, for five distributions, for the language data.

from both the fully crossed and incomplete simulations. The top panel for the fully crossed design shows that the distortion introduced by incorrectly assuming equal distances is relatively minor and one can still clearly see the various rater effects for each rater (noted at the top). The lower panel for the BIB design shows that the effects are again still mostly apparent. For example, Rater 2's criteria are all below the lines, and so this rater is lenient, Rater 3's criteria tend to be above the lines, showing severity, the first and last criteria for Rater 4 are below and above the lines, respectively, showing centrality, Rater 5's criteria are clustered in the middle, showing extremity, and so on. It is interesting to note that for the two raters with intersection point criteria locations (Raters 1 and 9), the first and last criteria are shifted down and up slightly for the BIB design, suggesting small centrality; this tendency was also noted in more extensive simulations (Zhou, 2019). Overall, the plots suggest that the relative criteria plots are fairly robust with respect to moderate violations of the equal distance assumption, and so a fit of the SDTr model is still useful for detecting rater effects.

Application to a Large-Scale Assessment

The example is data from a large-scale language test that was previously analyzed in DeCarlo et al. (2011). Analyzed here are data for one CR item (essay question) taken by 2,288 examinees and scored by 27 raters, with each essay scored by two raters.

Discrimination. Figure 5 shows a discrimination plot of the estimated discrimination parameters, scaled to range from zero to one, with horizontal lines showing the equal spacing locations. The plot gives detailed information about the distribution spacing for each rater. Note that some of the patterns shown in Figure 3 appear. For example, Rater 2 shows poor discrimination between Classes 4 and 5 (the two circles are close together at the top), whereas Rater 14 discriminates the end classes of 1 and 5 better than the middle classes of 2, 3, and 4, as shown by the larger spacing of the first and fifth circles from the middle three circles. Rater 8 shows poor discrim-

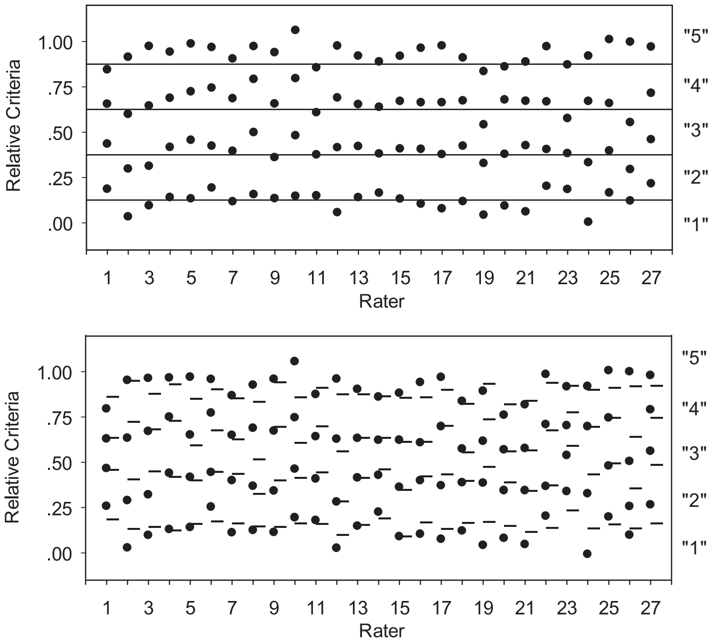

Figure 6. Relative criteria for a fit of the equal and unequal distance models to the language data, 1–5 response.

ination between classes 2 and 3. This is useful diagnostic information that could be helpful for rater training and monitoring. After identifying raters of potential interest from the plot, one should look at the parameter estimates and standard errors to get an idea if the deviations are of concern. Overall, Figure 5 suggests that the equal distance assumption is reasonable for these data; this conclusion is reinforced by the LR test given below.

Relative criteria. The top panel of Figure 6 shows a plot of the relative criteria estimates obtained for a fit of the equal distance SDTr model to the language data. The criteria estimates are generally close to the intersection points, with small effects appearing here and there. For example, Rater 10 and Rater 27's criteria estimates are all above the lines, and so these raters are relatively more severe. Rater 19's criteria estimates are all below the lines, and so this rater is relatively more lenient. Rater 12's estimates suggest centrality (though small), in that the first and last points are below and above the lines, respectively, and so this rater gives scores of "1" and "5" less frequently. Overall, the effects are few and relatively small and the estimated relative criteria locations are remarkably close to the (estimated) intersection points. The figure gives detailed information about each rater which could again be helpful with respect to rater training and monitoring.

The lower panel of Figure 6 shows, for purposes of comparison, a plot of the estimated relative criteria for a fit of the unequal distance SDTr-o model. The figure shows that, for the most part, the rater effects identified in the top panel also appear in the lower panel. For example, it is apparent in the lower panel that Raters 10 and 27 still show severity, in that the filled circles are all above the line segments; Rater 19 again shows leniency (circles all below the line segments), and Rater 12 still shows small centrality, in that the first and last points are below and above the line segments, respectively. The consistency of the results across the top and bottom panels of Figure 6 suggests that the relative criteria plot for the SDTr model (top panel) may be adequate in many cases.

LR test. Subtracting the minus two log likelihoods (using PME with Bayes constants of 1) obtained for fits of the SDTr and SDTr-o models gives an LR statistic of 82.50 with 81 degrees of freedom, which for the chi-squared distribution gives a *p*-value of .43, and so the null hypothesis of equal distances is not rejected. Thus, the LR test suggests that the equal distance SDTr model is adequate, as does visual inspection of the discrimination plot, as noted above. The estimated latent class sizes and standard errors (in parentheses) for the SDTr model are, for classes 1–5 respectively, .15 (.01), .14 (.01), .25 (.01), .24 (.02), and .23 (.01), and so there is some negative skew.

Discussion

Rater models apply to the situation where raters attempt to classify CRs such as essays into classes defined by a scoring rubric. It is assumed that the latent classes exist (the “true” classes), that they are ordered, and that the rater’s task is to detect instances from the classes. In the SDT approach, the rater is assumed to do this by comparing their perception of the quality of a particular essay to response criteria that delineate the classes. The equal distance SDTr model assumes that the perceptual distributions associated with essays from different classes have different locations, and that the difference between the locations of adjacent distributions is a constant, as shown in the top panel of Figure 1. The extended SDTr-o model relaxes this assumption and allows the distances to vary across the classes, with a monotonicity constraint, and so the locations of the perceptual distributions are only assumed to be ordered, as shown in the lower panel of Figure 1.

A useful aspect of the extended SDTr-o model is that it allows one to examine the equal spacing assumption of the SDTr model. This is a different type of rater effect, in that a rater may be able to discriminate between some of the latent classes better or worse than others. For example, a rater may be able to discriminate between different levels of good essays, but not between different levels of poor essays. The discrimination plots introduced here offer a simple visual way to assess this. For example, for the real-world data, it was apparent that the equal distance assumption was reasonable, which was also supported by the results of the LR test. The plots and test also offer useful diagnostic information about each of the raters.

For fully crossed data, estimation and testing was excellent, and so use of the SDTr-o model can be highly informative. For highly incomplete data, such as with only two raters per item, estimation problems appeared in some simulations (though not in the real-world data). Given that one has considerably less information with highly incomplete data, such as when only two raters score an item, it seems that a good strategy in that case is to use the SDTr model with the equal distance assumption, and then use the diagnostics offered by the SDTr-o model (the plots and test) for supplementary information about any possible discrimination problems.

An interesting extension of the model would be to relax the equal distance assumption for only *some* of the raters. One could possibly do a search over raters (e.g., by using the LR test on a case by case basis) to detect which raters violate the equal distance assumption, and then their discrimination parameters can be freed to see how the distances differ across the classes. Freeing the parameters for only a subset of the raters would result in the addition of considerably fewer parameters.

The model can also be used in situations with more than one CR item, such as the two essays required in the SAT. In that case, the SDTr-o model can be used in place of the SDTr model in Level 1 of the HRM-SDT (DeCarlo et al., 2011). This could provide information about the robustness of the HRM-SDT results, in both levels of the model, to the equal distance assumption, which could again be tested using the LR test (simultaneously or separately for two or more items).

Overall, the present research shows that the equal distance SDTr model appears to be useful for assessing rater accuracy and for detecting rater effects that commonly appear in real-world data, such as leniency/severity and centrality/extremity. The SDTr model appears to be useful even with moderate violations of the equal distance assumption. The SDTr-o model offers a useful extension that allows for the possibility that raters cannot discriminate between all of the latent classes equally well. It also offers useful diagnostics, such as plots and tests that supplement results obtained with the SDTr model.

References

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

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

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

Croon, M. (1990). Latent class analysis with ordered latent classes. *British Journal of Mathematical and Statistical Psychology*, *43*, 171–192.

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. (2003). 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.

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. (2008). *Studies of a latent-class signal detection model for constructed-response scoring*. ETS Research Report No. RR-08-63. Princeton NJ: Educational Testing Service.

DeCarlo, L. T. (2010). *Studies of a latent class signal detection model for constructed response scoring II: Incomplete and hierarchical designs*. ETS Research Report No. RR-10-08. Princeton NJ: Educational Testing Service.

DeCarlo, L. T. (2011). On the analysis of fraction subtraction data: The DINA model, classification, latent class sizes, and the Q-matrix. *Applied Psychological Measurement*, *35*, 8–26.

DeCarlo, L. T., Kim, Y., & Johnson, M. S. (2011). A hierarchical rater model for constructed responses, with a signal detection rater model. *Journal of Educational Measurement*, *48*, 333–356.

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

Fechner, G. (1860/1966). *Elements of psychophysics*. NY: Holt, Rinehart and Winston, Inc.

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

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

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

Lazarsfeld, P. F., & Henry, N. W. (1968). *Latent structure analysis*. New York: Houghton Mifflin Co.

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

Macmillan, N. A., & Creelman, C. D. (1991). *Detection theory: A user's guide* (2nd ed.). New York: Cambridge University Press.

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

Mokken, R. J. (1971). *A theory and procedure of scale analysis*. Berlin, Germany: De Gruyter.

Myford, C. M., & Wolfe, E. W. (2004). Detecting and measuring rater effects using many-Facet Rasch measurement: Part II. *Journal of Applied Measurement*, *5*, 189–227.

Neath, A. A., & Samaniego, F. J. (1997). On the efficacy of Bayesian inference for nonidentifiable models. *American Statistician*, *51*, 225–232.

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.

Robitzsch, A., & Steinfeld, J. (2018). Item response models for human ratings: Overview, estimation methods, and implementation in R. *Psychological Test and Assessment Modeling*, *60*, 101–139.

Saal, F. E., Downey, R. R., & Lahey, M. A. (1980). Rating the ratings: Assessing the psychometric quality of rating data. *Psychological Bulletin*, *2*, 413–428.

Sijtsma, K., & Molenaar, I. W. (2002). *Introduction to nonparametric item response theory*. Thousand Oaks, CA: Sage.

Thurstone, L. L. (1927). A law of comparative judgment. *Psychological Review*, *34*, 273–286.

Uebersax, J. S. (1993). Statistical modeling of expert ratings on medical treatment appropriateness. *Journal of the American Statistical Association*, *88*, 421–427.

Vermunt, J. K., & Magidson, J. (2016). *Technical guide for LatentGOLD 5.1: Basic, advanced, and syntax*. Belmont, MA: Statistical Innovations Inc.

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

Wind, S. A., & Peterson, M. E. (2018). A systematic review of methods for evaluating rating quality in language assessment. *Language Testing*, *35*, 161–192.

Wolfe, E. W. (2014). *Methods for monitoring rating quality: Current practices and suggested changes*. White Paper. Pearson Research Reports, Pearson.

Zhou, X. (2019). *Studies of extensions of HRM-SDT for constructed responses*. Unpublished doctoral dissertation.

Authors

LAWRENCE T. DECARLO is a Professor of Psychology and Education, Teachers College, Columbia University, 525 West 120th Street, New York, NY 10027; decarlo@tc.edu. His primary research interests include psychometrics and mathematical models in psychology.

XIAOLIANG ZHOU is a Research Fellow at Australian Council for Educational Research, 19 Prospect Hill Rd., Camberwell VIC 3124, Australia; xz2256@tc.columbia.edu. His primary research interests include psychometric methods and applications.