Journal of Educational Measurement
Fall 2011, Vol. 48, No. 3, pp. 333–356

A Hierarchical Rater Model for Constructed Responses, with a Signal Detection Rater Model

Lawrence T. DeCarlo
Teachers College, Columbia University
YoungKoung Kim
The College Board
Matthew S. Johnson
Teachers College, Columbia University

The hierarchical rater model (HRM) recognizes the hierarchical structure of data that arises when raters score constructed response items. In this approach, raters' scores are not viewed as being direct indicators of examinee proficiency, but rather as indicators of essay quality; the (latent categorical) quality of an examinee's essay in turn serves as an indicator of the examinee's proficiency, thus giving a hierarchical structure. Here it is shown that a latent class model motivated by signal detection theory (SDT) is a natural candidate for the first level of the HRM, the rater model. The latent class SDT model provides measures of rater precision and various rater effects, above and beyond simply severity or leniency. The HRM-SDT model is applied to data from a large-scale assessment and is shown to provide a useful summary of various aspects of the raters' performance.

Constructed response (CR) items, such as essay questions, are widely used in large-scale assessments, such as the SAT®, which includes one essay, and the GRE®, which includes two essays. CR items are also used in various national and international assessments, such as the National Assessment of Educational Progress (NAEP), Trends in International Mathematics and Science Study (TIMSS), and the Programme for International Student Assessment (PISA). CR items add an extra layer of complexity to the scoring process, because they require raters to score them, in contrast to multiple choice items, responses to which can simply be machine scored as right or wrong.

A typical way to deal with rater scores for CR items is to use an item response theory (IRT) model, such as the generalized partial credit model (Muraki, 1992) or the Rasch model (Linacre, 1989). For example, the generalized partial credit model is used for CR items in NAEP, whereas (a version of) the Rasch model is used for PISA. In these approaches, rater scores are used as direct indicators of examinee proficiency. A basic problem with this approach, however, is that it follows that more precise measurement of an examinee's proficiency can be obtained simply by using more raters, instead of by giving the examinee more items! This was shown formally by Mariano (2002) in terms of accumulated Fisher information, and was also noted by Patz (1996) and Patz, Junker, Johnson, and Mariano (2002), "as the number of raters per item increases, IRT facets models appear to give infinitely precise measurement of the examinee's latent proficiency θ_i" (p. 348).

The above problem forces one to recognize that raters do not really provide direct information about an examinee's proficiency, but rather provide information about the quality of an essay (or other CR) produced by the examinee. The (categorical) quality of the essay, in turn, provides information about the examinee's proficiency. Thus, there is a *hierarchical structure* to the data: in the first level, the raters' scores are ordinal indicators of the "true" category that an essay belongs to, whereas in the second level, the latent categories are ordinal indicators of examinees' proficiency. A hierarchical rater model (HRM), introduced by Patz (Patz, 1996; Patz et al., 2002), explicitly recognizes the hierarchical structure of the data; the HRM uses a signal detection model for the first level of the model (the rater model) and an IRT model for the second level (the item model).

The problem noted above does not arise with the HRM because obtaining more raters provides more information about which category a particular essay belongs to, and not directly about examinee proficiency, and so one cannot obtain infinitely precise estimates of proficiency simply by using more raters. In fact, Mariano (2002; also see Patz et al., 2002) showed that, for the HRM, the standard errors of the proficiency estimates could never be smaller than those obtained by using the true categories in the Level 2 IRT model (i.e., for an infinite number of raters or for perfect detection).

There are, however, some limitations to the particular model that was used in Level 1 of the HRM. For example, Patz et al. (2002) noted that, when rater discrimination was high, there were problems obtaining estimates of the rater severity parameter (the severity parameter indicates whether a rater is strict or lenient). Another limitation is that the model only allows for rater effects in terms of severity or leniency, whereas other rater effects often appear in real-world data. These problems do not arise, however, if a model based on an extension of traditional signal detection theory (SDT) to situations involving the detection of latent classes is used (e.g., DeCarlo, 2002, 2005, 2008a). Here it is shown that the approach offers advantages when used for the first level of the HRM, giving what will be referred to as an HRM-SDT model. For example, the latent class SDT model can deal with various rater effects that appear in real-world data, beyond simply severity or leniency, as shown below. The model is also straightforward to implement in standard software, given that the latent class SDT model is simply a generalized linear model with a latent categorical predictor. The approach also brings the well-established framework of SDT, as widely used in psychology, to the rater model of the HRM. The HRM-SDT is compared, in an analysis of a large-scale language assessment, to the original version of the HRM used by Patz et al. (2002) and others (e.g., Mariano & Junker, 2007). A partly Bayesian approach to estimation, posterior mode estimation (PME), is also discussed.

The Hierarchical Rater Model

The HRM (Patz, 1996; Patz et al., 2002) recognizes that the use of raters in CR scoring leads to a hierarchical data structure. In particular, in the HRM, the scores provided by raters are not direct indicators of examinee ability, as in the Rasch model or other IRT approaches to rater scoring, but rather are indicators

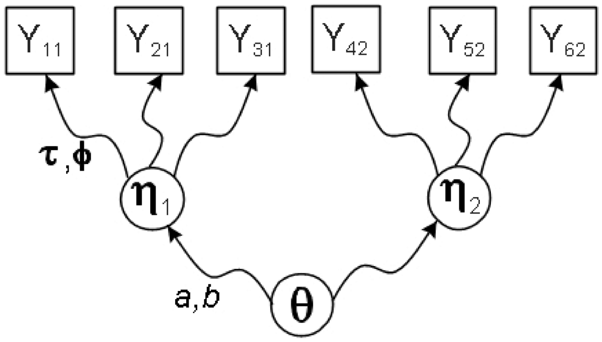

Figure 1. A representation of the HRM of Patz (1996) and Patz et al. (2002).

of the true or "ideal" category that each essay belongs to, where the true category is defined by the scoring rubric. For example, raters are trained to score the SAT using categories of 1 = *very little or no mastery*, 2 = *little mastery*, 3 = *developing mastery*, 4 = *adequate mastery*, 5 = *reasonably consistent mastery*, and 6 = *clear and consistent mastery* (a score of zero is used for essays not written on the essay assignment); detailed descriptions of each category are also provided (e.g., at http://professionals.collegeboard.com/testing/sat-reasoning/scores/essay/guide). Of course, one would not extensively train raters with a particular scoring rubric if one did not believe to some extent that there are in fact true categories of little mastery, adequate mastery, and so on.[1]

For a given essay, however, the true category is not directly observed or known, that is, the true category is *latent*; the latent categorical variable is denoted here as η. It follows that the raters' task is one of signal detection: given a particular essay, a rater's task is to determine (detect) which category the essay belongs to. Thus, a signal detection model is a natural candidate for the rater model in the first level of the HRM. A signal detection approach recognizes that raters' judgments are fallible indicators of the true category. The approach also provides a measure of a rater's ability to detect the true categories and recognizes that raters often have various response tendencies, as discussed below.

In the second level of the HRM, the latent categories serve as ordinal indicators of examinee proficiency, via an IRT model (note that the situation differs from the usual IRT model in that the indicators are latent rather than observed). Thus, an essay produced by an examinee is viewed as belonging to one of the categories defined by the scoring rubric, and the (true) category that the essay belongs to is in turn an indicator of the examinee's proficiency. In the notation used here, the observed rater scores Y are indicators of the true essay category η, and the true category η in turn is an indicator of examinee proficiency θ.

Figure 1 shows a representation of the HRM for the situation where each examinee responds to two items, say essays, with each essay scored by three raters, for a total of six raters. The rater scores are observed ordinal responses (e.g., a 1 to 6 score), which are shown in Figure 1 as Y_{jl} for the jth rater and lth item. The raters attempt to detect the true latent category for each essay, indicated by the latent categorical variable η_1 for the first item and η_2 for the second item. Note that the arrows from η_l to Y_{jl} are curved to indicate that the relation is nonlinear, and in particular the probability of Y_{jl} is connected to η_l via a nonlinear function (see below). The basic Level 1 (signal detection) parameters are τ_{jl} and ϕ_{jl}, as shown in Figure 1, which are rater precision and rater severity parameters, respectively.

In the second level of the HRM, the true essay category, η_l, serves as an ordinal indicator of an examinee's proficiency θ. The arrows are again curved to indicate that the latent categorical variables η_l (i.e., their probability) have a nonlinear relation to examinees' proficiency θ, via an IRT model (the generalized partial credit model is used here; other IRT models can also be used). The Level 2 parameters are a_l and b_{lm}, which are the usual discrimination and category step (transition) parameters, respectively.

As noted above, the HRM addresses the problem that arises when an IRT approach to rater scoring is used, which is that increasing the number of raters gives increasingly precise estimates of proficiency. In particular, as shown in Figure 1, obtaining more raters provides more information about η_l, and not directly about θ, whereas in an IRT approach obtaining more raters provides more information directly about θ. Note that this problem arises even if the test includes only one CR item.

When a test includes more than one CR item, another problem arises, as discussed by Patz et al. (2002). As shown in Figure 1, the first three raters are nested in the first item, whereas the second three raters are nested in the second item. If one simply uses the scores from the six raters in a model such as the Facets model (Linacre, 1989), which is commonly used for rater data, then the fact that raters are nested within items is ignored. The nesting means that the scores from Raters 1, 2, and 3, for example, are correlated in part because the three raters all score the same item (the first item), and similarly for Raters 4, 5, and 6, who all score the second item. Ignoring the correlation that arises due to nesting will give estimates of the standard errors of proficiency that are biased downward, as was noted by Patz et al. (2002) and others (e.g., Donoghue & Hombo, 2000; Wilson & Hoskens, 2001). The HRM, on the other hand, recognizes the nesting and corrects for the downward bias. The next sections introduce the components of the HRM in more detail.

Level 1: The Rater Model of Patz et al. (2002)

The signal detection-like model used by Patz et al. (2002; also see Mariano & Junker, 2007) for the first level of the HRM can be written as

$$p(Y_{jl} = k|\eta_l = \eta) \propto \exp\left\{-\frac{1}{2\psi_{jl}^2}[k - (\eta - \phi_{jl})]^2\right\}, \tag{1}$$

where Y_{jl} is the response of jth rater to the lth item, with the response being a discrete score k with K categories (the number of response categories 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_l is a latent categorical variable for the lth item, $\psi_{jl}^2$is a variance parameter for rater j (and item l) and ϕ_{jl} is a rater severity parameter (i.e., higher values indicate a more severe rater). Equation 1 is a signal detection-like model where the probabilities for each response category are approximately normally distributed. As noted by Patz et al. (2002), ψ_{jl}^2 is a measure of a rater's *lack* of reliability; its inverse, $\tau_{jl} = 1/(2\psi_{jl}^2)$, provides a measure of rater precision. The other rater parameter, ϕ_{jl}, is a severity parameter that indicates whether a rater is severe (positive values) or lenient (negative values), in that he or she tends to give low or high scores, respectively.

A problem with (1) that was recognized by Patz et al. (2002) is that the "most reliable raters" (i.e., those with small values of ψ_{jl}) tend to have the "least-well-estimated rater bias parameters" (p. 366). This occurs because, when ψ_{jl} is relatively small, the likelihood as a function of ϕ_{jl} is nearly constant over the range (–.5, .5) and is close to zero outside of that range (because the probability of a score in a response category other than the true category is near zero). Because the likelihood for ϕ_{jl} is nearly uniform from –.5 to .5, it is difficult to determine a unique value for ϕ_{jl} in that range. This was shown by Patz et al. (2002), in that there were problems determining a unique value for ϕ_{jl} for two highly reliable raters (with estimates of ψ_{jl} of .05 and .06)[2] because the posterior distribution of ϕ_{jl} was almost uniform.

Another limitation of (1) is that it cannot capture rater effects other than severity or leniency. There are, however, various types of rater effects that appear in large-scale assessments, such as *central tendency* or *restriction of the range* (e.g., Myford & Wolfe, 2004), as discussed below. A latent class model based on the traditional SDT model can easily deal with these types of effects.

Level 1: A Latent Class SDT Model of Rater Behavior

It has previously been suggested (DeCarlo, 2002, 2005) that psychological processes involved in CR scoring can be usefully understood within the framework of SDT (Green & Swets, 1988; Macmillan & Creelman, 2005; Wickens, 2002), which has been widely and successfully used in psychology and medicine. The application of SDT to CR scoring involves a latent class extension of SDT (DeCarlo, 2002); here it is noted that the model provides a useful alternative to the Level 1 model used by Patz et al. (2002).

CR scoring is conceptualized in SDT in terms of two basic aspects, namely a rater's perception of an essay's quality and his or her use of decision criteria. Figure 2 illustrates the basic ideas. Suppose that raters use responses of 1 to 4 to detect four latent categories. A basic idea in SDT is that a rater's decision is based upon his or her *perception*, Ψ, of the quality of an essay, where Ψ is a latent continuous random variable. It is assumed that the perceptions are realizations from a (location-family) probability distribution, such as the normal or logistic (for examples with other distributions, see DeCarlo, 1998). As shown in Figure 2, the distribution of Ψ has a different location for each of the latent categories, resulting in four locations for four categories. Thus, when presented with an essay from the first category, for example, the rater's perception (Ψ) of the quality of the essay is a realization from the first probability distribution; for an essay from the second category, the rater's perception is a realization from the second probability distribution, and so on.

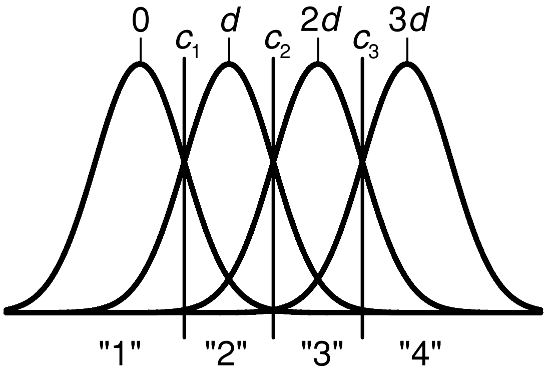

Figure 2. A representation of SDT, where raters make a 1 to 4 response to detect four stimuli (events).

The distances d between the perceptual distributions, shown in Figure 2, reflect a rater's ability to detect (discriminate) the latent categories, and so d provides a measure of rater detection or *rater precision*, exactly as in traditional SDT. Note that the approach is the same as that used by Clogg and Manning (1996) to define nonparametric measures of reliability for latent class models; for additional discussion and references, see DeCarlo (2002, 2005). Figure 2 also illustrates a simplifying assumption, which is that the distance d between the perceptual distributions is the same across categories, which was referred to as the *equal distance* model in DeCarlo (2002). One can also allow for unequal distances, but the equal distance model is more parsimonious and prior research has found that relative fit indices tend to select the equal distance model over the unrestricted model (DeCarlo, 2002, Table 4; also DeCarlo, 2005, p. 57); effects (which appear to be small) of relaxing this assumption are also being examined in current research.

A second basic idea of SDT is that the raters arrive at a decision by using their perceptions of essay quality together with *response criteria* that divide the decision space into the four response categories; the three criteria are shown as vertical lines in Figure 2. Thus, if a rater's perception of a particular essay is below the first criterion, c_1, the rater responds "1"; if it is between the first and second criteria, the rater responds "2," and so on. A useful aspect of the interpretation in terms of SDT is that it suggests reference points that obtained criteria can be compared to (e.g., the reference points reflect various decision aspects, such as attempting to maximize correct classifications, or basing decisions on likelihood ratios, and so on; see Egan, 1975). For example, Figure 2 shows criteria located at the intersection points of adjacent distributions (which is relevant when the number of latent categories is equal to the number of response categories and is where the likelihood ratios of adjacent distributions are unity), which is relevant to earlier discussions in SDT (Egan, 1975; Wickens, 2002) and to results found in recent studies of criteria locations in large-scale assessments (DeCarlo, 2008a).

The latent class signal detection model follows from Figure 2 and the assumptions discussed above and can be written as

$$p(Y_{jl} \leq k|\eta_l = \eta) = F(c_{jkl} - d_{jl}\eta_l), \tag{2}$$

where Y_{jl} is the response of jth rater to the lth item, where the response is a discrete score k with K categories, η_l is a latent categorical variable for the lth item that takes on M values η from 0 to $M - 1$ (the values implement the equal distance restriction on d_{jl}, see Figure 2), F is a cumulative distribution function (for a location-family of distributions, such as the logistic or normal), d_{jl} is a detection parameter for the jth rater and lth item, c_{jkl} are $K - 1$ strictly ordered response criteria, $c_{j1l} < c_{j2l} < \ldots < c_{j,K-1,l,}$ for the jth rater, lth item, and kth response category, with $c_{j0l} = -\infty$ and $c_{jKl} = \infty$. Note that, for the logistic model, d_{jl} and c_{jkl} are scaled with respect to the square root of the variance of the logistic distribution, $\pi^2/3$.

By allowing for (category-specific) response criteria, the latent class SDT model of (2) can handle a variety of *rater effects*. Rater effects refer to the observation that raters can have tendencies to be lenient or strict, to not use end categories, to restrict the range, and so on. For example, central tendency (see Myford & Wolfe, 2004) refers to the observation that some raters tend to not use (or underuse) the end categories (e.g., tending not to use 1 and 6 on a 1 to 6 scale). In SDT, this effect occurs if raters locate their highest criterion far to the right and their lowest criterion far to the left (and so the probabilities of using the end categories are small); this pattern was recently found in a latent class SDT analysis of a large-scale assessment (DeCarlo, 2008a, Figure 4). Note that central tendency cannot be reflected in a simple way by ϕ_{jl} in the rater model of Patz et al. (2002), because ϕ_{jl} only allows for changes in the *overall* level of severity, and not for differential severity across the response categories (as found below).

As another example, raters sometime show restriction of the range, in that they may not use all of the response categories (e.g., they may only give responses of 2 through 6 for a 1–6 scale); some examples of this are shown below. Again, in SDT, this type of effect simply means that the rater has a low criterion for the first category (c_{j1l}), and so they rarely or never give a response of "1" (and vice-versa if they only give responses of 1–5). However, this appears as higher (or lower) overall severity (as measured by ϕ_{jl}) in the model of Patz et al. (2002), in spite of the fact that the severity may be for only one of the response categories, and not overall. This and other results are studied below by comparing estimates of c_{jkl} to estimates of ϕ_{jl} in a large-scale assessment.

Level 2: Item Response Theory with Latent Indicators

The second level of the HRM treats the latent categorical variable η_l for each item (i.e., the true categories) as ordinal indicators of examinee proficiency θ (note that the Level 2 model is the same for both of the Level 1 models discussed above). For example, Patz et al. (2002) used the partial credit model (Masters, 1982), whereas we use the generalized partial credit model (Muraki, 1992). Both models use *adjacent category logits* (Agresti, 2002) and, in particular, the generalized partial credit model can be written as[3]

$$\log\left[\frac{p(\eta_l = \eta{+}1|\theta)}{p(\eta_l = \eta|\theta)}\right] = a_l\theta - b_{lm}, \tag{3}$$

where η_l is a latent categorical variable for item l that takes on values η from 0 to M – 1 (it is assumed that the number of latent classes, M, is the same as the number of response categories given in the scoring rubric, K, as discussed above, but this need not be assumed), θ is a latent continuous variable (examinee proficiency) assumed to be $N(0, 1)$, a_l is an item discrimination parameter for the lth item, and b_{lm} are M – 1 "item step" parameters (Masters, 1982), with $m = \eta + 1$ (so that the step parameters are b_{l1}, b_{l2}, and so on); the step parameters are also sometimes referred to as transition parameters (e.g., de Ayala, 2009). Equation 3 models the log of the ratio of a probability of a response in Category 1 (i.e., $\eta + 1$) versus Category 0 for $\eta = 0$, Category 2 versus Category 1 for $\eta = 1$, and so on. Using other transforms in the above, such as cumulative logits in lieu of adjacent category logits, gives other IRT models, such as the graded response model (Samejima, 1969). The partial credit model follows from (3) if a_l is set to be equal across items.

Adjacent category logits are used in the generalized (and partial) credit model because they were motivated by Masters (1982) in terms of "step" or transition probabilities between adjacent scores. For example, the first step parameter b_{l1} determines, for item l, the probability of going from a score of zero to a score of one; the second step parameter b_{l2} gives the probability of going from a score of one to a score of two, and so on. This is explicitly shown by (3). The model is also often written in terms of probabilities, in which case (3) can be rewritten as

$$p(\eta_l = \eta|\theta) = \frac{e^{\sum_{m=0}^{\eta}(a_l\theta - b_{lm})}}{\sum_{v=0}^{M-1} e^{\sum_{g=0}^{v} a_l\theta - b_{lg}}},$$

where $\sum\limits_{m=0}^{0}(a_l\theta - b_{lm}) \equiv 0$ (cf. Masters, 1982). Note that estimates of the marginal latent class sizes (given below), $p(\eta_l)$, can be obtained by computing the product of $p(\eta_l\ |\theta)$ and the node weights, w_q, at each quadrature point θ_q (used in Gaussian quadrature, see the estimation section below) and summing over all nodes.

The HRM-SDT Model

Figure 3 illustrates the complete HRM-SDT model with a latent class SDT model as the first level model and an IRT model as the second level model. As before, curved arrows are used to indicate nonlinear relations. In the first level, the raters' responses Y_{jl} (actually, the response probabilities) have a nonlinear relation to the raters' perceptions Ψ_{jl} of essay quality, with the response probabilities depending on the raters' criteria locations c_{jkl}, as shown in Figure 2. The location of the raters' perceptions, Ψ_{jl}, in turn depends on the true essay category, η_l; the straight arrows from η_l to Ψ_{jl} in Figure 3 indicate a linear relation, and in particular the mean of Ψ_{jl}

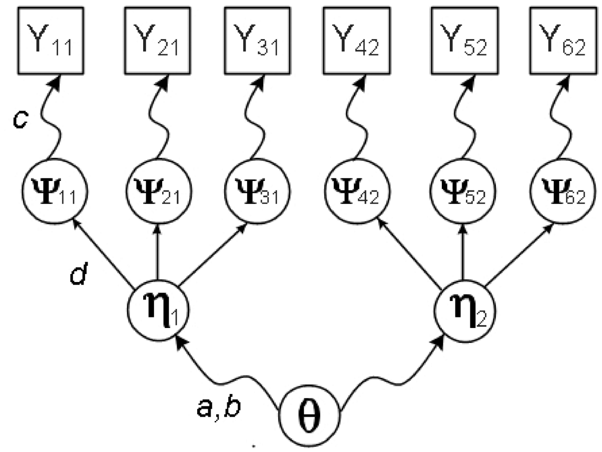

Figure 3. The HRM with a latent class signal detection model at Level 1, referred to here as the HRM-SDT model.

is shifted by d_{jl} as the latent category η_l increases by one. As shown in Figure 2, d_{jl} indicates the distance between the rater's perceptual distributions and is a measure of rater precision. In the second level, the true categories η_l serve as indicators of examinee proficiency θ, with discrimination and item step parameters a_l and b_{lm}, as shown in Figure 3.

Figure 3 shows that the HRM-SDT (and the HRM) is a type of higher-order factor model (see Bollen, 1989), with a latent class SDT model for the first level and an IRT model for the second level. Relations of the latent class SDT model to discrete factor models and discrete IRT models, as well as other models (e.g., located latent class models), are noted in DeCarlo (2002, 2005, 2008a); relations of the HRM to the Rasch model and generalizability theory models are discussed in Patz (1996) and Patz et al. (2002).

The complete HRM-SDT model includes both the Level 1 and 2 components given above. Let $\mathbf{Y}$ denote the vector of response variables for examinees, that is, $\mathbf{Y} = (Y_{11}, Y_{12},\ldots,Y_{1L}, Y_{21},\ldots, Y_{2L},\ldots,Y_{J1},\ldots, Y_{JL})$, where Y_{jl} is the response variable (which varies over examinees) for rater j and item l. Note that, for situations that commonly arise in practice (such as for the large-scale assessment examined here), some vectors of this matrix will be missing, in that the raters score only some of the examinees and also score only one of the items, for example.[4] Let $\boldsymbol{\eta} = (\eta_1, \eta_2,\ldots, \eta_L)$ indicate the L latent categorical variables for the CR items, each with M categories. The HRM-SDT model is

$$p(\mathbf{Y}) = \sum_{\eta} \int_{\theta} p(\mathbf{Y}|\boldsymbol{\eta}, \theta)\, p(\boldsymbol{\eta}|\theta)\; p(\theta)d\theta, \tag{4}$$

where $p(\mathbf{Y}|\boldsymbol{\eta}, \theta)$ is the rater component of the model (i.e., the first level) and $p(\boldsymbol{\eta}|\theta)$ is the model for the CR items (the second level). Two important assumptions are made in Level 1. The first is that, conditional on the latent variables $\boldsymbol{\eta}$, the observed ratings are independent of ability θ. The second is that, conditional on the vector of latent variables $\boldsymbol{\eta}$, the ratings are independent. These two assumptions simplify the rater component of the model as follows:

$$p(\mathbf{Y}|\boldsymbol{\eta}, \theta) = p(\mathbf{Y}|\boldsymbol{\eta}) = \prod_{jl} p(Y_{jl}|\boldsymbol{\eta}), \tag{5}$$

where j indicates the rater and l indicates the CR item. The conditional probabilities $p(Y_{jl}\,|\boldsymbol{\eta})$ are obtained from (2) by differencing the cumulative probabilities.

For Level 2, an assumption of conditional independence of the L latent variables given θ is made,

$$p(\boldsymbol{\eta}\mid\theta) = \prod_{l} p(\eta_l|\theta), \tag{6}$$

where the conditional probabilities $p(\eta_l\,|\theta)$ are obtained from (3), rewritten in terms of probabilities, as shown above. Substituting (5) and (6) into (4), and using the differenced form of (2) for the response probabilities and the probability form of (3) for the latent class probabilities, gives the complete HRM-SDT model.

Posterior Mode Estimation and Bayes' Constants

The HRM has previously been fit using a fully Bayesian approach implemented via Markov chain Monte Carlo (MCMC) methods (Patz et al., 2002). Here we present some notes on fitting the HRM-SDT model using maximum likelihood estimation (MLE) or posterior mode estimation (PME), the latter of which is a partly Bayesian approach that is useful for dealing with boundary problems, as discussed below. The approach can easily be implemented in current software.

The log likelihood function for the HRM-SDT is,

$$\log L = \sum_{n_y} \log f(y) = \sum_{n_y} \log \sum_{\eta} p(\mathrm{y}|\eta)\, p(\eta),$$

$$= \sum_{n_y} \log \sum_{\eta} p(\mathrm{y}|\eta) \int_{\theta} p(\eta|\theta)\, p(\theta) d\theta,$$

$$= \sum_{n_y} \log \sum_{\eta} \prod_{jl} p(\mathrm{y}_{jl}|\eta) \int_{\theta} \prod_{l} p(\eta_l|\theta)\, p(\theta)\, d\theta,$$

where the sum over all patterns of $\mathbf{y}$ and n_y is the number of observations for pattern $\mathbf{y}$. MLE can be performed using the expectation-maximization algorithm (Dempster, Laird, & Rubin, 1977). Most implementations, including the one utilized here, approximate the integral over θ with Gaussian quadrature. The result is that the integral is replaced by a summation over Q quadrature points $\theta_1, \theta_2, \ldots, \theta_Q$. Implementation of the approach in latent class software such as LEM or Latent Gold has been discussed in Vermunt (1997) and Vermunt and Magidson (2005).

A limitation of the approach via MLE is that *boundary problems* often occur, as has long been recognized in latent class analysis (e.g., Clogg & Eliason, 1987; Maris, 1999). Boundary problems occur when one or more of the parameter estimates are close to the boundary, such as obtaining an estimate of a latent class size of zero or unity, or obtaining a large or indeterminate estimate of detection (with a large or indeterminate standard error).

A number of authors have discussed the use of PME as a simple way to deal with boundary problems (e.g., Galindo-Garre & Vermunt, 2006; Gelman, Carlin, Stern, & Rubin, 1995; Maris, 1999; Schafer, 1997; Vermunt & Magidson, 2005). In PME, rather than maximizing the log likelihood, the log posterior function is maximized. Note that the posterior is related to the likelihood and prior as follows:

$$\text{posterior} \propto \text{prior} \times \text{likelihood},$$

that is, the posterior is proportional to the prior times the likelihood; in PME, the prior in essence acts as a penalty for solutions that are close to the boundary.

One approach is to place priors on the conditional probabilities for responses and latent classes, instead of placing priors (directly) on the model parameters (e.g., see Agresti, 2002; Galindo-Garre, Vermunt, & Bergsma, 2004). Let $\pi_{y_{jl}/\eta}$ and $\pi_{\eta_l|\theta}$ denote the response and latent class probabilities, respectively, as given by (5) and (6). Using the Dirichlet distribution as a prior gives

$$p\left(\pi_{y_{jl}/\eta}\right) \propto \prod_{\eta} \prod_{k} p(y_{jl} = k|\eta_{jl} = \eta)^{\alpha_1 - 1},$$

$$p\left(\pi_{\eta_l|\theta}\right) \propto \prod_{\eta} p(\eta_l = \eta|\theta)^{\alpha_2 - 1},$$

for the conditional response and latent class probabilities, respectively, where

$$\alpha_1 = 1 + \frac{B_1 \widehat{\pi}_{jkl}}{M^L},$$

$$\alpha_2 = 1 + \frac{B_2 W_q}{M^L}$$

(see Vermunt & Magidson, 2005), where the *Bayes' constants*, B, determine the strength of the prior distribution and can be thought of as the number of observations added to cells of the (complete) data frequency table; w_q are (scaled) weights used in Gaussian quadrature (i.e., the observations are in essence distributed across the quadrature points according to the node weights), M is the number of latent class categories, L is the number of CR items, and $\hat{\pi}_{jkl}$ are the observed marginal proportions for Y_{jl}. For the latent class probabilities, the Bayes' constants are in essence B_2 pseudo-observations that are split across the M^L latent class patterns and weighted by w_q for the Q nodes. For the response probabilities, B_1 is split across the M^L latent class patterns and weighted by the observed marginal probabilities of Y_{jl} for the K response categories (instead of simply split across the K response categories), as discussed by Clogg, Rubin, Schenker, Schultz, and Wiedman (1991) and Vermunt and Magidson (2005).

The important aspect to note is that, for the Level 2 generalized partial credit (GPC) model, a larger value of B_2 penalizes the likelihood more and so the fitted data is smoothed toward an independence model; thus, a_l of the IRT-part of the HRM-SDT is smoothed toward zero and b_{lm} is smoothed toward locations that give equal probabilities. For the Level 1 rater model (latent class SDT), larger values of B_1 smooth d_{jl} toward zero and, because of the smoothing toward the observed margins (under independence), there is only a small effect on c_{jkl} (it is smoothed toward the marginal probabilities and not toward equal probability locations, which one can argue is a more sensible approach; see in particular Clogg et al., 1991). PME with Bayes' constants used in the manner described above has been implemented in Latent Gold (Vermunt & Magidson, 2005), which was used here to fit the HRM-SDT model.

The HRM model of Patz et al. (2002) was fit using MCMC in WinBUGS (Lunn, Spiegelhalter, Thomas, & Best, 2009). The approach was the same as described by Patz et al. (2002), except for a minor reparameterization, which is that the discretized normal density shown in (1) was used, but with a slope parameter $\tau_{jl} = \frac{1}{2\Psi_{jl}^2}$, with a Gamma(1,1) prior for τ and a $N(0,4)$ prior for ϕ.

Parameter Recovery for the HRM-SDT

Parameter recovery for the HRM-SDT using MLE or PME as described above has been previously investigated in several simulation studies (DeCarlo, 2008b, 2010; DeCarlo & Kim, 2009; Kim, 2009). For example, results for simulations of fully crossed (across-rater) designs (i.e., all raters score all essays) were presented by DeCarlo (2008b, 2010) and by Kim (2009). The main findings were that the rater parameters for the SDT model (first level) were generally well recovered. For the second level (IRT model), the item parameters were poorly recovered for the generalized partial credit model when there were only two items, whereas recovery of the item parameters was greatly improved when a third item was added. Boundary problems also occurred; however, it was shown that using PME with Bayes' constants of unity led to good parameter recovery.

Kim (2009) presented simulations that examined parameter recovery in the HRM-SDT for incomplete designs (across raters), as obtained in many real-world assessments (i.e., each rater scores only a subset of examinees, and so there are missing values). PME with Bayes' constants of unity was used. Kim found that, for the SDT part of the model, parameter recovery for the HRM-SDT was good for a range of detection (d) that is found in practice (e.g., about 1 to 6; see DeCarlo, 2005, 2008a, 2008b, 2010). Estimation of the item parameters was marginal for the generalized partial credit model when there were only two items, and was improved when a third item was added, as was also found for fully crossed designs. Kim (2009) and DeCarlo and Kim (2009) also noted that, when the HRM-SDT was extended by using multiple choice items as direct indicators of θ in Level 2, estimation of the CR item parameters (Level 2) was greatly improved, even with only one or two CR items. Another interesting result was that estimation of the rater parameters at Level 1 also appeared to be slightly improved when multiple choice items were added in Level 2.

In sum, simulation studies have shown that the rater parameters are generally well recovered for the HRM-SDT model (at least for PME with Bayes' constants of unity, as also used here). The item parameters also appear to be adequately recovered, although when the generalized partial credit is used as the Level 2 model, more than two items or other additional information (i.e., such as that provided by multiple choice items) appear to be needed for adequate parameter recovery, though this requires further study.

The next section applies both the HRM-SDT model and the HRM of Patz et al. (2002) to data from a large-scale language assessment. Particular attention is paid to results for the Level 1 model, the rater model, given that this is where the models differ, although results for Level 2 are also examined.

Application to a Large-Scale Assessment

The real-world data are from a large-scale language assessment where each examinee wrote two essays. The data consist of essays from 2,350 examinees obtained on one test day, with each examinee answering the same two CR items. Each essay was scored by 2 raters out of a pool of 54 raters; 34 of the raters scored the first item, whereas 20 different raters scored the second item; in addition, 13 raters who scored the first item also scored the second item (but for a different examinee) and so the first item was scored by 34 raters and the second item was scored by 33 raters (with 13 common raters). The scoring rubric consisted of a 1 to 5 rating scale. Each rater scored anywhere from 7 to 484 essays, with a mean of 174 essays per rater (median of 168).

Results

Level 1: Rater Models

Detection. Figure 4 shows, separately for each item, the distribution of estimates of the rater detection parameter, d_{jl}, obtained for a fit of the HRM-SDT model. The estimates of d_{jl} are generally within a range of 1 to 6, as also found in previous studies (e.g., DeCarlo, 2002, 2005, 2008a; Kim, 2009), and are approximately normally distributed. The mean d_{jl} for the first item (3.8) is higher than for the second item (3.1), which suggests that the raters were better at detecting the true categories for the first item than for the second item. However, this result could also reflect that a (mostly) different set of raters scored the second item. For the subset of 13 raters who scored both items, the mean of d_{jl} was 3.6 for the first item and 3.0 for the second item, which again suggests that rater detection was better for the first item. With

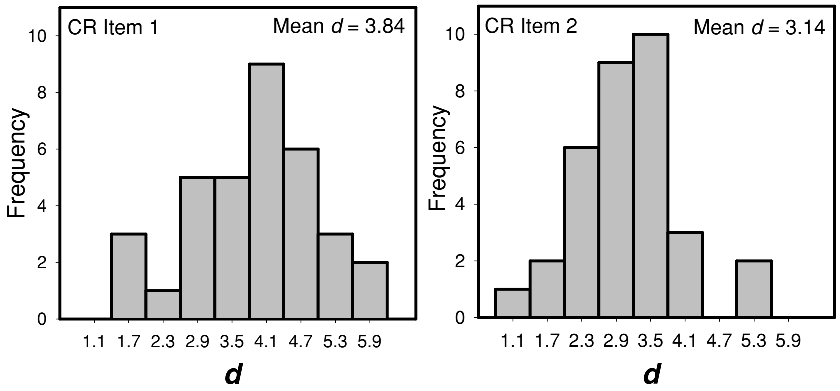

Figure 4. Frequency plots of the estimates of detection (d) for the HRM-SDT model for the first and second items.

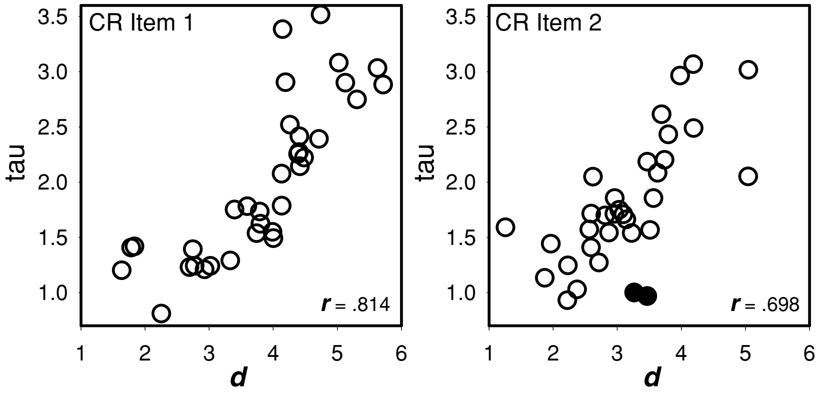

Figure 5. Plots of estimates of the rater precision parameter τ for the HRM model of Patz et al. (2002) against estimates of the precision parameter d for the HRM-SDT model, separately for each item.

respect to estimates of the rater precision parameter τ_{jl} for the HRM model, the mean was 2.0 for the first item and 1.8 for the second item.

Figure 5 shows a plot of estimates of the rater precision parameter τ_{jl}, where $\tau_{jl} = \frac{1}{2\Psi_{jl}^2}$, for a fit of the HRM of Patz et al. (2002), against estimates of d_{jl} for a fit of the HRM-SDT. The figure shows that the precision estimates are generally similar across the two models; however, there are some differences. For example, in the right panel for item 2, the two filled circles near to $d_{jl} = 3.5$ suggest that the two raters show adequate precision; however, they would be tagged by the HRM of Patz et al. (2002) as performing poorly, given that the estimates of τ_{jl} are among the lowest; differences of this sort could potentially have practical implications (e.g., in terms of monitoring rater performance).

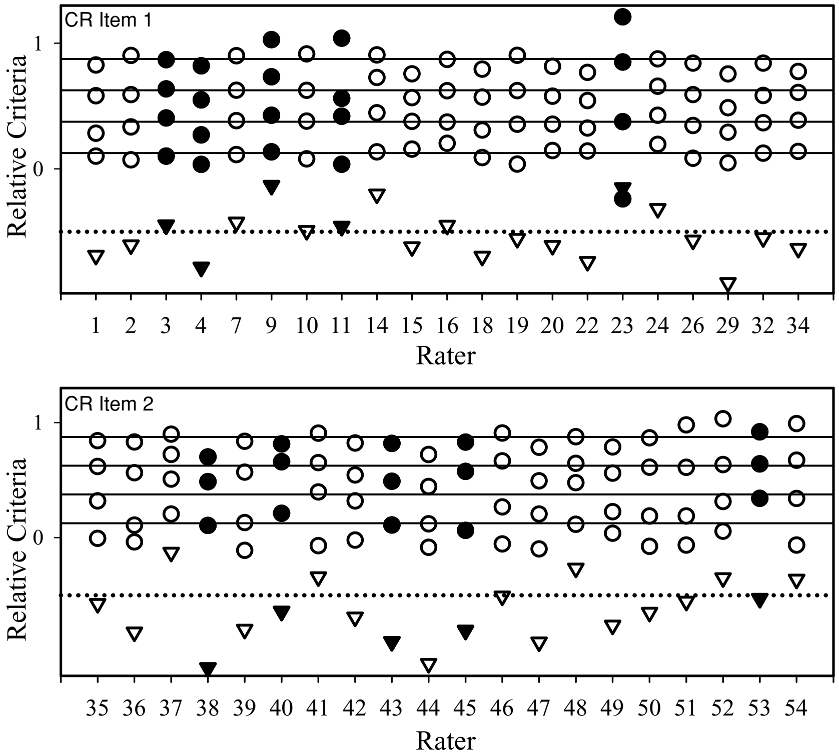

Figure 6. Plots of the relative criteria locations (circles) for HRM-SDT for 41 raters, shown separately for each item. The solid horizontal lines show intersection points for the underlying distributions (see the text). Estimates of the severity parameter (ϕ) for Patz et al.'s (2002) model are also shown (as inverted triangles); the dotted line shows $\phi = 0$ (note that the zero and inverted triangles are shifted down by –.5 for visual clarity). The filled circles show examples that are discussed in the text.

Response criteria and rater effects. Figure 6 shows, for fits of the HRM-SDT model, estimates of the (rescaled) response criteria c_{jkl} locations for 21 raters who scored only the first item (top graph) and 20 raters who scored only the second item (bottom graph; the remaining 13 raters who scored both items are discussed below). Note that it is difficult to compare the absolute criteria locations across raters when d_{jl} differs across raters (e.g., because the intersection points then differ across raters), whereas it is more informative to examine locations of the criteria rescaled so that the underlying distributions all have the same relative locations for all of the raters, with the lowest distribution at 0 and the highest at 1; these were referred to as *relative* criteria locations by DeCarlo (2005, 2008a). In particular, d_{jl} is rescaled to be one for each rater, with c_{jkl} also rescaled (by dividing it by the estimate of d_{jl} times one minus the number of latent classes). This allows the criteria locations to be compared across raters, in that they indicate where the rater locates his or her criteria relative to the intersection points of the underlying distributions. In particular, the four horizontal lines between 0 and 1 in Figure 6 show the four points at which the five underlying distributions (for the five latent categories) intersect.

The top plot of Figure 6 shows that, for the first item, the relative criteria locations for the 21 raters (shown as circles; the triangles are discussed below) are generally remarkably close to the intersection point locations, which shows that the intersection points serve as useful reference points. The bottom plot of Figure 6 shows that, for the second item, the third and fourth criteria (top two circles) are generally close to the intersection points; however, the first and second criteria (bottom two open circles, the filled circles are discussed below) tend to fall below the first and second lines. This means that the raters gave fewer scores of 1 and 2 for the second item (as compared to the first item), and so they were more lenient on the lower end for the second item. Note, however, that they were *not* more lenient with respect to giving scores of 4 or 5, and so the leniency is only with respect to low scores, and not high scores (note that "leniency" and "severity" are being used here only in a relative way, that is, relative to other observers).

Figure 6 also shows several other rater effects. For purposes of comparison, Figure 6 also includes, for a fit of the HRM of Patz et al. (2002), estimates of the severity parameter ϕ_j, which are shown as inverted triangles; note that the severity estimates are all shifted downward in Figure 6 by –.5 for visual clarity (the dotted line shows the zero point for ϕ_j). The first result to note in Figure 6 is that, when the relative criteria estimates lie on or close to the horizontal lines, as for Raters 3 and 32, for example, the estimates of ϕ_j are close to zero (i.e., the inverted triangles are close to the dotted line). Thus, Figure 6 shows that, when a rater's criteria in the HRM-SDT model are close to the intersection points, the severity parameter in Patz et al.'s model tends to be near zero.

Figure 6 also shows several cases where estimates of the criteria (c_{jk}) are all simply shifted downward, indicating a more lenient rater, as for Raters 4 and 29 (the circles are all below their corresponding lines). In these cases, the estimates of ϕ_j are also shifted downward, also indicating a more lenient rater. Similarly, when the criteria estimates are (mostly) simply shifted upward, as for Rater 9, then the estimate of ϕ_j is also shifted upward, which indicates that the rater is relatively more severe. Thus, results for fits of the HRM-SDT and HRM show that ϕ_j adequately captures simple upward or downward shifts in the response criteria c_{jk} (as it should), which indicates whether a rater is overall more severe or lenient.

However, when the response criteria show patterns other than a simple upward or downward shift, differences between c_{jk} and ϕ_j appear. Rater 11 in the top plot of Figure 6 provides an example. The criteria estimates in this case show that the rater's first criterion (circle) is below the first horizontal line and the fourth criterion is above the last line, which suggest that the rater tends to under-use the end categories of 1 and 5; for additional examples where this tendency appears, see DeCarlo (2008a) and Kim (2009). As discussed above, this effect is referred to in the measurement literature as central tendency, because the rater tends to primarily use the middle categories (Myford & Wolfe, 2004). It is also interesting to note that the middle two criteria for Rater 11 are close together, which indicates that the rater gives scores of 2 and 4 but tends to not give a score of 3. Unfortunately, this detail is lost with respect to the estimate of ϕ_j for the HRM, shown as a filled triangle, which is simply close to zero for Rater 11. Thus, in this case the rater severity parameter of the HRM of Patz et al. (2002) does not reveal a rater effect, namely central tendency, whereas the HRM-SDT does. This occurs because ϕ_j represents an average effect for each rater, and so it cannot capture the under-use of end categories shown by Rater 11 (particularly when the under-usage is fairly symmetrical).

Another interesting example is Rater 23 (top plot of Figure 6). The criteria estimates from the HRM-SDT model show that this rater tends to not use scores of 1 and 5, as shown by the extreme locations of the first (bottom filled circle) and fourth (top filled circle) criteria, and so this rater can be said to also show central tendency. Note that the third and fourth criteria for Rater 23 (top two filled circles) are well above the third and fourth intersection locations (top two lines), whereas the first criterion (bottom circle) is well below the first intersection location (bottom line; and the second criterion is on the appropriate line). Put simply, Rater 23 is severe with respect to assigning scores of 4 and 5, in that the rater tends to not give these high scores, but is lenient with respect to low scores, in that the rater tends to give a score of 2 over a score of 1 (i.e., the rater tends to not give the lowest score). With respect to the HRM, Figure 6 shows that Rater 23's severity (filled inverted triangle) is well above zero, and so the rater is tagged by the HRM simply as being severe. This is a not an accurate summary of Rater 23's performance, however, in that the rater is only severe with respect to high scores but is lenient with respect to low scores, which is nicely shown by the relative criteria estimates of the HRM-SDT.

The bottom plot of Figure 6 shows five cases, shown as filled circles, where the raters did not use all of the response categories. In all cases, the raters only gave scores of 2 through 5, and so the lowest filled circle is c_2 (and not c_1). Note that Rater 53's three relative criteria (for the HRM-SDT, circles) are located at the appropriate intersection points (horizontal lines) and the estimate of ϕ_j (for the HRM, inverted triangle) is close to zero, as was also found in the top plot of Figure 6. The other four raters in the bottom plot (filled circles) show relative criteria that tend to be close to or below the intersection points and these raters are tagged as "lenient" by ϕ_j (i.e., the filled inverted triangles are low). However, the simple conclusion of leniency is again not accurate. For example, for Rater 45, the second criterion is clearly below the intersection point (i.e., the lowest filled circle, which is c_2, is on the first line rather than on the second line); however, the third and fourth criteria are close to their corresponding intersection point locations (they are close to the top two lines). Thus, Rater 45 is lenient only with respect to tending to give a score of 2 over a score of 1; however, the rater is *not* lenient with respect to assigning scores of 4 or 5, and so the simple conclusion of leniency suggested by ϕ_j is not accurate. The examples reflect (expected) limitations of ϕ_j with respect to dealing with situations where not all of the response categories are used or the categories are used differentially across the scale.

Raters who scored both items. Figures 7 and 8 present results for the 13 raters who scored both items. Figure 7 shows that estimates of the detection parameter d_j of the HRM-SDT show a degree of consistency across the two items; the Pearson

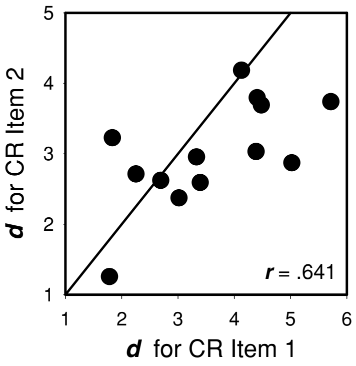

Figure 7. A plot of the estimates of *d* (HRM-SDT model) for the second item against estimates of *d* for the first item for 13 raters who scored both items.

correlation is $r = .64$ with $p = .018$. There was a somewhat weaker relation between τ_j for these raters, with a correlation of $r = .45$ with $p = .123$.

Figure 8 shows estimates of the relative criteria locations and ϕ_j for the 13 raters who scored both items. Differences in score category usage, as discussed above for Figure 6, again appear. Figure 8 shows that, in general, the raters' first two criteria locations tend to be lower for the second item as compared to the first item (i.e., the lower two circles are close to the lower two solid lines in the upper plot but tend to be below the lines in the lower plot). This means that the raters tended to give fewer scores of 1 or 2 for the second item as compared to the first item. This result was also found for raters who only scored one item (Figure 6), and so this result appears both between and within raters. Figure 8 also shows that, as for the raters in Figure 6, the raters only tend to be lenient with respect to not assigning low scores, but this is not the case for high scores (4 or 5, given that the top two circles tend to be close to the lines). Again, these details are not picked up by estimates of ϕ_j for the HRM.

In summary, the results show that various rater effects such as central tendency, restriction of the range, and other idiosyncrasies with respect to score usage, appear in a large-scale assessment. The criteria estimates of the HRM-SDT model provide useful information about these effects, whereas estimates of the severity parameter ϕ_j of the original HRM do not always correctly reflect the effects. The detection parameter d_j of the HRM-SDT model also provides useful information about rater precision; for the data examined here, the mean of d_j (3–4) indicates good detection and was consistent with values found in other (similar) studies.

Level 2: CR Item Model

HRM-SDT. The top part of Table 1 shows results for the second level of the HRM-SDT model, which is the IRT model (i.e., the generalized partial credit model). The estimates of the item discrimination parameters a_l are 2.97 and 4.92 for the first

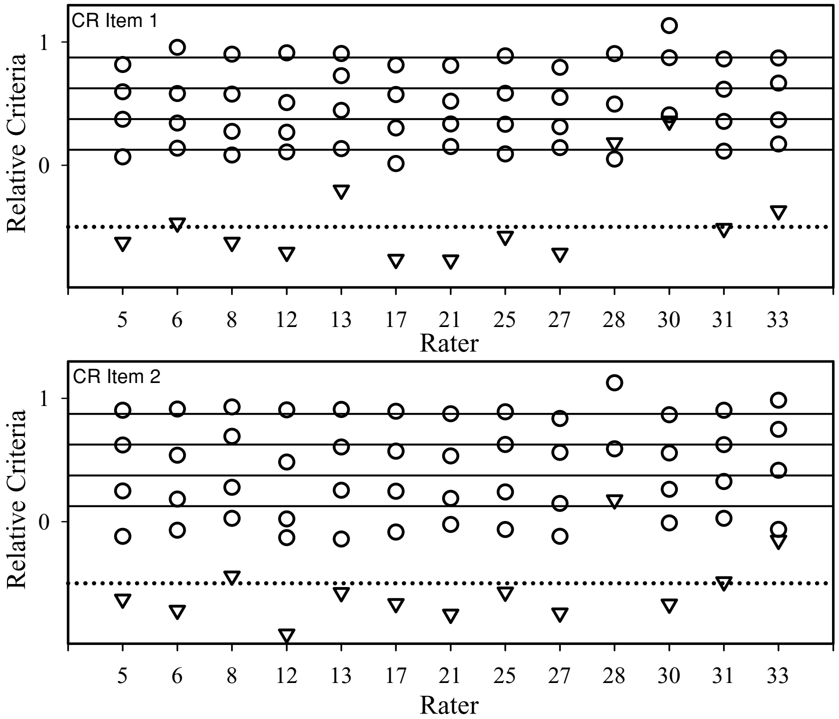

Figure 8. Plots of the relative criteria locations for the HRM-SDT model (circles) for 13 raters who scored both items, along with estimates of the severity parameter ϕ (inverted triangles, which are again shifted down by –.5 for visual clarity) for the HRM model of Patz et al. (2002).

and second items, respectively. Restricting the two a_l parameters to be equal gives a difference in the log posteriors of 1.64 with 1 *df* (which suggests that the restricted model is not rejected) with a common estimate of *a* of 3.73 (with a standard error of .24). In light of the simulations reviewed above, there is likely low power to detect differences in item discrimination when there are only 2 items (in which case the partial credit model can simply be used). Table 1 also shows that the first two item step parameters are lower for the second item as compared to the first, whereas the last step parameter is higher; the result is smaller latent class sizes for the corresponding categories, as shown next.

The middle part of Table 1 shows estimates of the latent class sizes, $p(\eta_l)$, where estimates of a_l and b_{lm} were used to get estimates of the conditional probabilities—the conditional probabilities are multiplied by the weights and densities of the 11 nodes used for Gauss-Hermite quadrature and then summed, as noted above. Table 1 shows that, for both items, Categories 3 and 4 have the largest estimated class sizes. For the second item, there tend to be fewer cases in the first category as compared to the first item, which is also indicated by the larger negative estimate of

Table 1
Results for Level 2, the IRT Model (GPC), of the HRM-SDT and the HRM for a Large-Scale Language Assessment

HRM-SDT		
Parameters	CR Item 1	CR Item 2
a_l	2.97 (.43)	4.92 (.92)
b_{l1}	–3.11 (.50)	–7.76 (1.40)
b_{l2}	–1.43 (.23)	–4.34 (.74)
b_{l3}	.85 (.24)	.81 (.32)
b_{l4}	3.57 (.82)	6.08 (1.05)
Estimates of the Latent Class Sizes		
Parameters	CR Item 1	CR Item 2
$p(\eta_{l1})$	.16 (.01)	.07 (.01)
$p(\eta_{l2})$	.18 (.01)	.15 (.02)
$p(\eta_{l3})$	.27 (.02)	.36 (.02)
$p(\eta_{l4})$	.25 (.02)	.31 (.02)
$p(\eta_{l5})$	.14 (.03)	.11 (.01)
HRM		
Parameters	CR Item 1	CR Item 2
a_l	2.31 (.28)	3.78 (.43)
b_{l1}	–2.26 (.32)	–8.04 (.88)
b_{l2}	–1.32 (.16)	–4.49 (.50)
b_{l3}	.57 (.14)	.90 (.32)
b_{l4}	2.65 (.35)	5.40 (.62)

b_{l1} for Item 2 shown in the top part of the table (i.e., the category step boundary is lower for going from the first to second category, and so there are fewer observations in the first category). Note that, for both items, the latent class sizes suggest a slightly negatively skewed distribution, which might reflect aspects of the language test; other tests have given a more normal distribution. An analysis of a large sample (>42,000) who took the language test also suggested negative skew (see DeCarlo, 2010, p. 26).

HRM of Patz et al. (2002). The lower part of Table 1 shows estimates (i.e., posterior means and standard deviations) of Level 2 parameters obtained for a fit of the HRM of Patz et al. (2002). The item discrimination and item step parameters, a_l and b_{lm}, respectively, are slightly smaller than those found for the HRM-SDT but show a similar pattern, in that discrimination a_l is higher for the second item (the posterior standard deviations also tend to be smaller than the standard errors shown in the top of the table). The item step parameter estimates b_{lm} for the HRM are close to those found for the HRM-SDT model (and so the latent class size estimates, not shown, are also very close to those shown in the middle of the table). Overall, it appears that using Patz et al.'s model as the Level 1 model (and MCMC) in lieu of the SDT model gives similar results with respect to the Level 2 parameters.

Discussion

The use of CR items is an integral part of educational assessment. Given that CR items require raters to score them, a model of rater behavior in CR scoring is needed. A latent class signal detection model provides a useful framework for understanding how raters score items and for monitoring and evaluating rater performance; here it is shown that the model can easily be incorporated into the HRM. The approach has advantages over the signal detection-like model used by Patz et al. (2002), in that the latent class SDT model can capture the "catalog of rater effects" (Myford & Wolfe, 2004) that appear in real-world data. In particular, a fit of the SDT model provides information about the locations of raters' response criteria, c_{jkl}, and the criteria locations in turn provide information about various rater effects. The SDT model also provides an estimate of rater detection, d_{jl}, which reflects rater precision. The approach via SDT also avoids problems that arise with the signal detection-like model of the original HRM, as discussed above. The HRM-SDT can easily be fit with standard software, given that it is within a family of generalized linear models with latent variables (e.g., Skrondal & Rabe-Hesketh, 2004).

As shown in Figures 1 and 3, the hierarchical model differs from the usual IRT approach to rater scoring in that the HRM includes a middle layer, η, whereas rater responses Y_{jl} load directly (and nonlinearly) on examinee proficiency θ in an IRT approach. As discussed above, the IRT approach raises problems with respect to increasing precision of estimates of examinee proficiency with increasing numbers of raters. Here it is noted that it also follows from the hierarchical model that an IRT analysis of CR items confounds rater effects (i.e., the response criteria and detection parameters) with item effects (i.e., the item step and discrimination parameters). For example, an item might appear to be more difficult for a different sample in an IRT approach not because of a change in item difficulty, but because of a change in rater severity. This type of problem with an IRT approach has previously been recognized and has led to suggestions as to how to link CR items (e.g., Tate, 1999).

The HRM does not raise this problem because it separates rater effects (d_{jl} and c_{jkl}) from item effects (a_l and b_{lm}), and so one can evaluate aspects of the raters separately from aspects of the items (e.g., the particular essay question). For example, the results found above suggested that rater detection d_{jl} was, on average, lower for the second item than for the first, whereas item discrimination a_{jl} appeared to be higher for the second item. This means that the raters were less able to detect the correct categories for the second item (lower d_{jl}), but the true categories for the second item provided higher discrimination of examinee proficiency (higher a_{jl}). Put simply, the first result tells us how accurately the raters can determine the true categories of the essays, whereas the second result tells us how well a particular item (i.e., the essay question) functions with respect to discriminating between examinees' proficiency. The separation of rater and item aspects by the hierarchical model offers interesting possibilities with respect to item banking and has implications for the equating of CR items and the study of rater drift, some of which are being examined in current research.

Notes

[1] This is not to say that the rubric necessarily defines the correct number of categories. The SDT model is useful in this respect in that one can easily compare the relative fit of models with different numbers of latent categories (see DeCarlo, 2005). Suffice it to say that if empirical evidence for a different number of latent categories is consistently found, then the scoring rubric needs to be rethought and revised.

[2] Note that estimates of ψ of .05 and .06 give estimates of τ of $1/(2 \times .05^2) = 200$ and $\tau = 139$, whereas estimates of τ obtained here are in the range of 1 to 3.5, and those obtained by Patz et al. (2002) are generally around $\psi = .40$ which gives $\tau = 3.1$. Thus, in terms of τ, the estimates of 200 and 139 seem excessively large and could reflect boundary problems.

[3] To show the model that was actually fit, the parameterization is changed slightly from the usual $a_l\,(\theta - b_{lm})$.

[4] Bock, Brennan, and Muraki (2002) noted that a nested design, with different raters across two or more items, as illustrated in Figures 1 and 3, minimizes the effects of differences in rater severity, and so this type of design is commonly used in large-scale assessments.

References

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

Bock, R. D., Brennan, R. L., & Muraki, E. (2002). The information in multiple ratings. *Applied Psychological Measurement*, *26*, 364–375.

Bollen, K. A. (1989). *Structural equations with latent variables*. New York, NY: 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., & 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* (pp. 169–182). New York, NY: Academic Press.

Clogg, C. C., Rubin, D. B., Schenker, N., Schultz, B., & Wiedman, L. (1991). Multiple imputation of industry and occupation codes in census public-use samples using Bayesian logistic regression. *Journal of the American Statistical Association*, *86*, 68–78.

de Ayala, R. J. (2009). *The theory and practice of item response theory*. New York, NY: Guilford Press.

DeCarlo, L. T. (1998). Signal detection theory and generalized linear models. *Psychological Methods*, *3*, 186–295.

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

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

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

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

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: ETS.

DeCarlo, L. T. & Kim, Y. K. (2009, April). *On scoring constructed response items and multiple choice items: Incorporating signal detection and item response models into a hierar-*

chical rater model. Paper presented at the 2009 annual meeting of the National Council on Measurement in Education, San Diego, CA.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm (with discussion). *Journal of the Royal Statistical Society Series B*, *39*, 1–38.

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.

Egan, J. P. (1975). *Signal detection theory and ROC analysis*. New York, NY: Academic Press.

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

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

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

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

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

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

Lunn, D., Spiegelhalter, D., Thomas, A., & Best, N. (2009). The BUGS project: Evolution, critique, and future directions. *Statistics in Medicine*, *28*, 3049–3067.

Macmillan, N. A., & Creelman, C. D. (2005). *Detection theory: A users guide* (2nd ed.). New York, NY: Cambridge University Press.

Mariano, L. T. (2002). *Information accumulation, model selection and rater behavior in constructed response student assessments*. Doctoral dissertation, Carnegie Mellon University.

Mariano, L. T., & Junker, B. W. (2007). Covariates of the rating process in hierarchical models for multiple ratings of test items. *Journal of Educational and Behavioral Statistics*, *32*, 287–314.

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

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

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

Myford, C. M., & Wolfe, E. W. (2004). Detecting and measuring rater effects using many-facet Rasch measurement: Part I. In E. V. Smith, Jr. & R. M. Smith (Eds.), *Introduction to Rasch measurement* (pp. 460–517). Maple Grove, MN: JAM Press.

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

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

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

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

Skrondal, A., & Rabe-Hesketh, S. (2004). *Generalized latent variable modeling: Multilevel, longitudinal, and structural equation models*. New York, NY: Chapman & Hall/CRC.

Tate, R. L. (1999). A cautionary note on IRT-based linking of tests with polytomous items. *Journal of Educational Measurement*, *36*, 336–346.

Vermunt, J. K. (1997). LEM: A general program for the analysis of categorical data [Computer software and manual]. Retrieved from http://www.tilburguniversity.edu/nl/over-tilburg-university/schools/socialsciences/organisatie/departementen/mto/onderzoek/software/ Tilburg University

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

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

Wilson, M., & Hoskens, M. (2001). The rater bundle model. *Journal of Educational and Behavioral Statistics*, *26*, 283–306.

Authors

LAWRENCE T. DECARLO is an Associate Professor of Psychology and Education in the Department of Human Development, Teachers College, Columbia University, 525 West 120th Street, New York, NY 10027; decarlo@tc.edu. His primary research interests include statistical models in psychology and education.

YOUNGKOUNG KIM is an Assistant Psychometrician at The College Board, 45 Columbus Avenue, New York, NY 10023; rkim@collegeboard.com. Her primary research interests include latent variable modeling for educational data.

MATTHEW S. JOHNSON is an Associate Professor of Statistics and Education in the Department of Human Development, Teachers College, Columbia University, 525 West 120th Street, New York, NY 10027; johnson@tc.edu. His primary research interests are the development of statistical models for the analysis of educational data, especially data from large-scale educational assessments.