Incorporating Criterion Ratings Into Model-Based Rater Monitoring Procedures Using Latent-Class Signal Detection Theory

Applied Psychological Measurement
2017, Vol. 41(6) 472–491
© The Author(s) 2017
Reprints and permissions:
sagepub.com/journalsPermissions.nav
DOI: 10.1177/0146621617698452
journals.sagepub.com/home/apm

Brian F. Patterson¹, Stefanie A. Wind², and George Engelhard Jr.³

Abstract

This study presents a new criterion-referenced approach for exploring rating quality within the framework of latent-class signal detection theory (LC-SDT) that goes beyond commonly used reliability indices, and provides substantively meaningful indicators of rater accuracy that can be used to inform rater training and monitoring at the individual rater level. Specifically, this study illustrates a flexible application of restricted LC-SDT modeling, in which restrictions can be specified for the true latent classification to reflect the unique characteristics of a particular assessment context. While the LC-SDT modeling framework provides immediately useful characterizations of raters' behavior, the restricted LC-SDT offers complementary evidence to further support the monitoring of rater behavior by bringing criterion ratings to bear. This study uses ratings from a large-scale writing assessment, and findings suggest that the criterion (i.e., restricted) LC-SDT provides useful information about rating quality for operational raters relative to criterion ratings, which may ultimately inform rater training and monitoring procedures.

Keywords

rater effects, signal detection theory, latent-class model, expert rater, writing assessment

It is common practice in large-scale rater-mediated assessment systems to employ expert raters, and to use their ratings as criteria to evaluate the quality of ratings assigned by operational raters (Engelhard, 1996, 2013; Johnson, Penny, & Gordon, 2009; Wolfe, Jiao, & Song, 2016). Often, these expert raters have extensive experience with the operational assessment system, and their judgments can serve as meaningful criteria for evaluating ratings of operational raters during training exercises or ongoing rater monitoring. Using this criterion-referenced perspective on rating quality, several scholars have described the match between expert and operational

¹Questar Assessment, Inc., Apple Valley, MN, USA

²The University of Alabama, Tuscaloosa, AL, USA

³The University of Georgia, Athens, GA, USA

Corresponding Author:

Brian F. Patterson, Senior Psychometrician, Questar Assessment, Inc., 5550 Upper 147th Street West, Apple Valley, MN 55124, USA

Email: bfp2103@columbia.edu

ratings as evidence of *rater accuracy*, as distinct from norm-referenced rating quality indices, in which operational ratings are compared with an overall group of raters, such as rater agreement or reliability indices (Engelhard, 1996; Sulsky & Balzer, 1988; Woehr & Huffcutt, 1994). These criterion-referenced rater accuracy indices are useful in the sense that they provide a substantively meaningful indicator of rating quality at the individual rater level that can be used to inform rater training and monitoring procedures in operational assessment settings.

Criterion-referenced indicators of rater accuracy have been incorporated into a variety of methods for evaluating rating quality. In addition to agreement statistics between operational and criterion raters (e.g., Johnson et al., 2009), accuracy indices based on number-correct scores include the use of ANOVA techniques to explore a variety of aspects of the difference between operational and criterion ratings (Cronbach, 1955; Sulsky & Balzer, 1988). Modern measurement techniques have also been adapted to include a criterion-referenced perspective on rater accuracy. For example, Engelhard (1996) and Wind and Engelhard (2013) demonstrated the use of Rasch measurement theory models in which the match between operational and criterion ratings is modeled as a latent variable on which individual rater locations are estimated as evidence of rater accuracy. Similarly, Wang, Engelhard, and Wolfe (2015) modeled the distance between operational and criterion ratings using an unfolding approach for rater accuracy.

In contrast, although previous research has demonstrated the utility of models based on latent-class signal detection theory (LC-SDT) in the context of rater-mediated assessments (e.g., DeCarlo, 2002; DeCarlo, Kim, & Johnson, 2011), the use of criterion ratings to evaluate rater accuracy has not been explored within a signal detection theory framework. Specifically, previous research on the use of LC-SDT models to evaluate rating quality suggests the diagnostic value of this approach for exploring differences in individual raters' discrimination and classification accuracy. However, the use of these LC-SDT indicators has not been fully explored within a criterion-referenced framework for exploring rating quality.

The purpose of this study was to demonstrate the use of a criterion-referenced approach to expand and complement the current applications of LC-SDT modeling for the purpose of monitoring rater behavior. Similar to previous applications of LC-SDT in the context of rater-mediated assessments, this approach to monitoring rating quality is grounded in a psychological theory of how raters arrive at their ratings. Specifically, this study demonstrates a flexible approach to a criterion LC-SDT model in which restrictions can be specified to reflect the unique characteristics of a particular assessment context including criterion ratings. It is important to note that just as previous work has been applied to the qualification and monitoring of operational raters, so too should this proposed method be viewed in that vein, as opposed to an approach that would lead to improved prediction of examinees' latent classification. In other words, while the interpretation of parameter estimates arising from existing LC-SDT methods may immediately aid an operational testing program, the criterion LC-SDT model results bring criterion ratings to bear and provide a different yet complementary perspective on rater behavior. Despite the availability of rating quality indicators based on signal detection theory approaches and other modern measurement techniques, the use of aggregate indicators of rating quality, such as interrater reliability, remains prevalent in research on rater-mediated assessments (Wind & Peterson, *in press*) as well as in practical assessment settings (Johnson et al., 2009).

This study is structured as follows: First, a set of descriptive analyses is presented to identify patterns in ratings for a large-scale state writing assessment, and the extent to which the operational constructed response (CR) ratings agree with those of a criterion rater is characterized. Next, LC-SDT rater model (DeCarlo, 2002) analyses of ratings in four analytic domains (Ideas, Style, Organization, and Conventions) used in a large-scale writing assessment are presented to demonstrate the immediate usefulness of existing LC-SDT procedures. The authors of the

Figure 1. Depiction of a hypothetical signal detection theory rater task in which the true latent classification of 2 is represented by the solid conditional distribution, with the dashed conditional distributions representing incorrect classifications as Categories 1, 3, and 4.

Note. The rater has equally spaced conditional distributions, and his latent thresholds occur at the intersection point of the conditional distributions, giving rise to maximum classification accuracy.

present study then consider the utility of using criterion ratings to restrict the LC-SDT model classification and comparison of parameters estimated under the two models. Finally, they consider the practical applications of these techniques related to rater monitoring and training for implementation in both small research and large, high-stakes assessment settings.

LC-SDT Models for Rater-Mediated Assessments

The LC-SDT rater model has been applied in several settings outside the realm of educational achievement tests, including medical diagnostic imaging (e.g., DeCarlo, 2002). DeCarlo and his colleagues (DeCarlo, 2005, 2008; DeCarlo et al., 2011) have argued for the applicability of LC-SDT in evaluating ratings from rater-mediated assessments.

Figure 1 demonstrates the applicability of this method to rater-mediated assessments. Consider the task of a particular rater assigning a rating on a 1-to-4 scale for an essay in terms of a single analytic domain. The true latent classification of the writing sample is Category 2 as represented by the solid conditional probability distribution. The horizontal axis represents the rater's continuous perception of the extent to which the writing sample matches the ideals of the analytic domain—this perceptual component of the model is related to rater discrimination represented by the d parameter. The better the rater discriminates (i.e., the greater the d parameter), the further apart the conditional distributions will be, and the better the rater will be able to differentiate between adjacent latent classes. The judgmental part of the process is where the rater maps his or her continuous perception onto the categorical rating scale based on a well-defined rubric for assigning ratings. The judgmental process is represented in the model as the rater's latent threshold for rating scale category k (denoted as the c_k parameters). In this figure, the rater has equally spaced latent thresholds placed at the intersection points of the conditional distributions (DeCarlo, 2008), which maximizes classification accuracy. These latent thresholds are also referred to as *criteria* in existing research on LC-SDT. The authors of the present study use the term *latent thresholds* in this study to avoid confusion with the terms *criterion rater* and *criterion ratings*.

The LC-SDT model may be represented in terms of the conditional probability of a rater assigning a particular score as follows:

$$Pr(Y_{ij} = k | \eta_i = \eta) = F(c_{ijk} - d_{ij}\eta_i) \quad k \in [1, K] \quad \eta \in [1, H], \quad (1)$$

where Y_{ij} is rater j 's rating of the examinee's mastery of domain i ; η_i is the examinee's latent classification for domain i ; $F(\cdot)$ is a cumulative distribution function; and d_{ij} and c_{ijk} are rater j 's discrimination and $H - 1$ latent thresholds, respectively, for domain i .¹ It is convenient to consider LC-SDT models in which as many latent classes are predicted as there are manifest rating categories (i.e., $H = K$), but that need not be the case in general. Where a logistic cumulative distribution function is used, the d_{ij} parameters may be interpreted as the natural log of the odds of detecting a latent class η member for a true latent class η member, relative to a true latent class $\eta - 1$ member. Put another way, the rater has an odds of detecting a true class η member of $e^{d_{ij}}$ times greater than for a true class $\eta - 1$ member (DeCarlo, 2008).

Method

Latent Classification Restriction via Criterion Ratings

To illustrate how criterion ratings can be incorporated into LC-SDT models to provide additional insight into rating quality within a signal detection theory framework, a general class of restrictions of the latent classification resulting from the LC-SDT model analysis is presented. Specifically, how restrictions can be placed on latent classifications is demonstrated to reflect operational raters' alignment with criterion ratings, or more generally, to reflect the assessment context to which the ratings are applied.

Prior to determining the particular form of the restriction, it must be made clear what precisely one wants to restrict. In a certain scenario, one may wish to include the criterion ratings in the estimation of the model along with the operational raters. This would be most reasonable in the case of a criterion rater who does not define the construct per se, but rather maps potentially error-laden latent classifications onto the rating scale. In contrast, in the current study, the criterion rater was intimately involved in the development of the rating rubrics and training materials, and in the oversight of the raters. To that end, he had a role in defining the construct. Rather than modeling his ratings, the authors of this study instead used his idiosyncratic criterion ratings as a means of restricting the true latent classification, as opposed to constraining the latent classification to equal those somewhat fallible ratings.

Typically, one may consider exact agreement between criterion ratings and predicted latent classification as a reasonable restriction, but such a conservative rule would yield little meaningful information about the behavior of the operational raters relative to the criteria against which their performance is being judged. Instead, it may be more informative to define a restriction based on criteria that match the unique assessment context. For example, some operational assessment systems use a critical value of ± 1 rating scale point to identify essays for rescoring or adjudication by a third, independent rater. In this case, a potential restriction may be to require that the latent classification exhibit adjacent agreement—that is, the latent classification must be within 1 score point of the criterion rating. Under such a rule, the predicted latent classification from the criterion LC-SDT model would be confined to the range of ± 1 rating scale point around the criterion rating.

It is important to note that these restrictions that are conditional on the criterion rating are different from reducing the number of latent classes to be predicted. To illustrate this distinction, consider one essay response to which the criterion rater assigns a rating of 3 and a second essay

response to which that criterion rater assigns a rating of 2. Further, suppose that the restricted latent classification must be within the ± 1 rating category of the criterion rating. The true latent classification in this study would be restricted to either 2, 3, or 4 in the former case, and to either 1, 2, or 3 in the latter case. It is important to note that, while any given response may only be classified into one of three categories, the number of unique latent classes into which responses may be classified is still four. In other words, the restriction is not to reduce the total number of latent classes being predicted from four to three but rather to conditionally restrict the latent classification of responses according to the criterion rating.

The criterion LC-SDT model itself and the precise restriction may be written as Expression 2 and Expression 3, respectively:

$$Pr(Y_{ij} = k | \tilde{\eta}_i = \eta) = F(c_{ijk} - d_{ij}\tilde{\eta}_i) \quad k \in [1, K] \quad \eta \in [1, H], \quad (2)$$

$$\tilde{\eta}_i \in [Y_{i0} - \delta, Y_{i0} + \delta] \quad \delta \in [0, H], \quad (3)$$

where $\tilde{\eta}_i$ is the restricted latent classification and δ is some integer that is inversely related to how restrictive the latent classification ought to be around the criterion rating (Y_{i0}). In other words, where Expression 2 represents the criterion LC-SDT model, Expression 3 specifies the form of the restriction that is placed on the latent classification, according to the criterion rating (Y_{i0}). For example, if the model were to restrict latent classification to adjacent agreement with the criterion rating, that would imply $\delta = 1$. The choice of δ should be influenced by a few factors, including the strength of belief in the criterion ratings (it should be lower when the belief is stronger), the number of available score points on the rating scale (it may be higher with greater numbers of available score points), and other factors. Indeed, the whole form of restriction may need revision when, for example, the number of latent classes that best characterize the rating data does not equal the number of rating categories, or if the raters demonstrate idiosyncratic behavior that would be better served with alternative restrictions.

Data

This study is a secondary analysis of data from Wind and Engelhard's (2013) study of the Georgia Middle Grades Writing Assessment (MGWA). In the original study, a sample of essay ratings was obtained from the assessment system that reflected the authentic operational scoring design, in which two independent raters scored each essay using an analytic rubric. Furthermore, it was necessary to obtain a sample of essay ratings for which raters scored common performances that provided connectivity among the raters and essays. As a result, 2,120 essays were identified for analysis that were scored by a sample of 10 operational raters from the larger pool of raters hired to score essays for this administration of the assessment. Each of the raters completed training procedures specific to the 2011 administration of the MGWA, and passed a qualifying examination prior to operational scoring. Within the sample of essays, each rater scored at least 48 essays in common with each of the other nine raters in the sample, leading to an incomplete but connected rating design (Eckes, 2015; Engelhard, 1997). Because the focus of the current study is on the use of LC-SDT techniques to evaluate rating quality at the individual rater level, the sample of raters explored in this secondary analysis provides a manageable sample size with which to illustrate the investigation of psychometric properties of ratings in detail for each rater.

Students' 90-min writing samples were rated on a 1-to-5 scale (Georgia Department of Education, 2011). Due to infrequent usage of the highest category (<1% of ratings), ratings of 5 were combined with ratings of 4. The four domains were Ideas, Style, Organization, and

Conventions. From the pool of 10 operational raters considered in this study, two provided operational ratings on the four domains each, while a criterion rater gave ratings for all four domains to all 2,120 writing samples. In other words, for each student's writing sample, eight operational ratings (two raters by four domains) and four criterion ratings (one criterion rater by four domains) are observed.

Procedure

The data analysis procedure for this study included three major steps: First, to compare the utility of the LC-SDT model with traditional empirical approaches, a series of descriptive analyses are conducted. The authors of this study first inspected the frequency of each assigned rating category and other patterns in how the raters used the rating scale. Traditional statistics (such as exact and adjacent agreement, as well as quadratic-weighted kappa) were also prepared to compare and contrast with the model-based means of comparing operational and criterion judgments.

Next, two LC-SDT models were applied and compared using information criteria and rater parameter estimates. Specifically, the second step in the data analysis procedure included the application of the unrestricted LC-SDT model (Expression 1; henceforth simply, the "LC-SDT model") that included both the criterion and operational raters. This model was used to obtain indicators of rater discrimination and relative latent thresholds for the criterion rater and each of the operational raters. Finally, a restricted LC-SDT model (henceforth, the "criterion LC-SDT model") was applied to the set of operational raters in which a restriction was specified to reflect patterns observed in the criterion ratings. The resulting rater parameters from the LC-SDT model were compared with those from the criterion LC-SDT model.

All LC-SDT models in this study were estimated using Latent GOLD 4.0 (Magidson & Vermunt, 2005). Online Appendix A provides an example of Latent GOLD syntax for the criterion LC-SDT model; a comment indicates the modifications needed to produce the unrestricted solution.

Results

Rating Category Use and Descriptive Comparison With Criterion Ratings

Table B1 in Online Appendix B summarizes the distribution of observed ratings across categories, along with agreement statistics between each of the operational raters and the criterion rater. These results revealed several notable patterns in the use of the rating categories that are particularly relevant to the LC-SDT model results. First, 89.1% of all ratings across raters and domains were either a 2 or a 3. This suggests that across domains, raters generally exhibited a tendency toward centrality as described by Wolfe and McVay (2012). When looking at results for each domain, ratings of Organization were the most central with 90.4% of ratings being either 2 or 3. And ratings of students' essays on the Conventions domain tended to be lower than the other domains with 7.1% and 49.1% rated as 1 and 2, respectively, compared with 4.4% and 43.8%, respectively, across the other three domains.

The results in Table B1 of Online Appendix B also indicate heterogeneity in terms of the patterns of ratings by rater. For example, Rater 8 used Categories 2 and 3 more frequently (93.1%) relative to other raters (88.7%), indicating that she or he exhibited greater centrality. Furthermore, inspection of the pattern of ratings given by Rater 9 indicates potentially greater leniency than the other raters, where ratings in Category 1 were assigned less frequently (3.5%) relative to other raters (5.2%), and ratings in Category 4 were assigned more frequently (9.3%)

Figure 2. Rater discrimination for the four domains (d_{ij}), with the criterion rater denoted as Rater 0.

Note. Posterior mode estimates of the LC-SDT model with Bayes constants of 1 for the four-latent-class solution for ratings of each domain modeled separately. LC-SDT = latent-class signal detection theory.

relative to other raters (5.4%). When comparing the agreement measures for each operational with the criterion ratings, Rater 6 demonstrated the greatest exact agreement (83.0%), while Rater 9 agreed least often with the criterion ratings (72.6%). After correcting for chance agreement using quadratic-weighted kappa (κ), those inferences about the operational raters with the greatest and least agreement with the criterion ratings changed. Specifically, across domains, Rater 5 exhibited the greatest agreement [$M(\kappa_{i5}) = .881$], while Rater 8 demonstrated the least agreement [$M(\kappa_{i8}) = .780$].

LC-SDT Rater Models

LC-SDT model including criterion ratings and operational raters. Turning to the LC-SDT model-based framework described above, for each domain separately, estimates were prepared for the LC-SDT model that includes both the criterion rater and operational raters. In Figure 2 (also see Table B2 in Online Appendix B), the authors of this study showed the estimated rater discrimination by rater and domain. These results highlight the diagnostic value of the LC-SDT approach to exploring rating quality at the individual rater level. For example, it is possible to compare discrimination estimates within individual raters across domains to identify differences in rater judgment related to various aspects of an assessment system. Furthermore, the operational raters can be compared with the criterion rater to identify differences in discrimination that may warrant further attention.

Examination of the results in Figure 2 reveals several interesting patterns related to individual raters. First, the estimated discrimination for the criterion rater (labeled “Rater 0”) has the highest discrimination across the four domains, when compared with the 10 operational raters. There was also heterogeneity in the mean discrimination across operational raters, with Rater 3 having the highest mean discrimination [$M(d_{i3}) = 4.84$] and Rater 9 having the lowest mean discrimination [$M(d_{i9}) = 2.91$]. The figure also shows that there were some differences across

Figure 3. Rater relative latent thresholds for Ideas domain (c_{ijk}/d_{ij}), with the criterion rater denoted as Rater 0.

Note. Posterior mode estimates of the LC-SDT model with Bayes constants of 1 for the four-latent-class solution. LC-SDT = latent-class signal detection theory.

domains, with raters typically having higher discrimination for the Conventions domain [$M(d_{4j}) = 5.01$] when compared with the other three domains [$M(d_{1j}, d_{2j}, d_{3j}) = 3.56$]. It is worth noting—as Online Appendix B demonstrates—that the standard errors of the d_{ij} parameters are greater, relative to their corresponding estimates for the Conventions domain, when compared with the other three domains. This is unsurprising, given that previous studies have also found that in the presence of larger estimates of d , standard errors tend to be larger (DeCarlo, 2005).

Results related to the raters' three estimated latent thresholds for distinguishing between each of the four latent classes also revealed interesting diagnostic information at the individual rater level; these results are summarized in Figures 3 through 6. These figures present each rater's estimated relative thresholds within each domain. Horizontal reference lines are included that indicate the optimal relative threshold for maximizing classification accuracy under certain conditions, including the assumption of a common d and equally spaced conditional distributions for latent classification. The optimal relative threshold (DeCarlo, 2005) can be calculated as follows:

$$RT_k^{\text{Opt}} = [(2k-1)/2][1/K], \quad (4)$$

where k indexes the threshold, and K is the total number of thresholds. By this definition, the optimal criteria are restricted to be between 0 and 1. In the case of the current study, three optimal thresholds are relevant to distinguish between four latent classes. These optimal thresholds can be calculated as follows:

$$RT_1^{\text{Opt}} = (1/2)(1/3) = 1/6, \quad (5)$$

Figure 4. Rater relative latent thresholds for Style domain (c_{2jk}/d_{2j}), with the criterion rater denoted as Rater 0.

Note. Posterior mode estimates of the LC-SDT model with Bayes constants of 1 for the four-latent-class solution. LC-SDT = latent-class signal detection theory.

Figure 5. Rater relative latent thresholds for Organization domain (c_{3jk}/d_{3j}), with the criterion rater denoted as Rater 0.

Note. Posterior mode estimates of the LC-SDT model with Bayes constants of 1 for the four-latent-class solution. LC-SDT = latent-class signal detection theory.

Figure 6. Rater relative latent thresholds for Conventions domain (c_{4jk}/d_{4j}), with the criterion rater denoted as Rater 0.

Note. Posterior mode estimates of the LC-SDT model with Bayes constants of 1 for the four-latent-class solution. LC-SDT = latent-class signal detection theory.

$$RT_2^{Opt} = (3/2)(1/3) = 3/6, \tag{6}$$

and

$$RT_3^{Opt} = (5/2)(1/3) = 5/6. \tag{7}$$

First, the raters’ estimated relative thresholds for separating the first and second latent classes ($=c_{ij1}/d_{ij}$ or the “1/2” latent threshold on the figures) and for separating the third and fourth latent classes ($=c_{ij3}/d_{ij}$ or the “3/4” latent threshold on the figures) also supported the inference that raters exhibited centrality. In other words, across domains and raters, the first relative latent threshold was lower than what would be optimal in terms of classification accuracy (the lowest dashed horizontal reference line at 1/6), and the third relative latent threshold was higher than what would be optimal in terms of classification accuracy (see the highest reference line at 5/6).

Interestingly, the criterion rater (denoted as Rater “0”) also exhibited centrality, though perhaps not as much as the operational raters, and it is worth noting that his second relative criterion was generally very close to the optimal level of 1/2. The exception to that finding was that both the criterion rater and operational raters’ latent threshold for the Conventions domain tended to be shifted up, indicating that they rated writing samples more strictly for Conventions (a) than on the other three domains; and (b) than would have been optimal in terms of classification accuracy on the Conventions domain. The other raters who stand out relative to their operational peers are Rater 8—who exhibited the greatest centrality—and Rater 9—who was the most lenient across domains. From this point forward, the criterion rater’s rating data were not explicitly included in the estimation of the models. The same (unrestricted) LC-SDT model

Table 1. Comparison of Information Criteria for Criterion and LC-SDT Models.

Domain	BIC		AIC	
	LC-SDT	Crit.	LC-SDT	Crit.
Ideas	7,609	9,015	7,382	8,789
Style	7,787	9,107	7,561	8,881
Organization	7,581	9,025	7,354	8,798
Conventions	7,917	9,407	7,690	9,181

Note. “LC-SDT” indicates the posterior mode estimation results of the LC-SDT (i.e., unrestricted) model with Bayes constants of 1 for the four-latent-class solution for ratings of each domain, while “Crit.” refers to the criterion LC-SDT model results. LC-SDT = latent-class signal detection theory.

was estimated without the criterion rater, to facilitate meaningful comparison with the estimates arising from the criterion LC-SDT (see Table B3 in Online Appendix B).

Criterion LC-SDT model including only operational raters. First and given the effect that the larger context of this assessment system had on raters’ usage of the rating scale, the proposed method for restricting latent classification given in Expression 3 was not suitable for these data. Instead, the following restriction was used:

$$\tilde{\eta}_i \in \left\{ \begin{array}{l} \{1, 2\} \\ \{3, 4\} \end{array} \right\} \text{ where } Y_{i0} \in \left\{ \begin{array}{l} \{1, 2\} \\ \{3, 4\} \end{array} \right\}. \tag{8}$$

In other words, the latent classification was restricted to either 1 or 2, when the criterion rating was either a 1 or 2; similarly, when the criterion rating was either a 3 or 4, the predicted latent classification was restricted to a 3 or 4. This restriction is used because it was in keeping with how raters used the rating scale, given their prior knowledge of the consequences of scoring at or above a 3 (i.e., passing). It is important to note that there were still a total of four latent classes being predicted; the restriction was placed on the true latent classification according to the criterion rating, but latent classes of 1, 2, 3, and 4 were considered, although any single response was restricted to two of those four.

Beyond comparing the predictions of the latent-class model, both the traditional information criteria and model classification statistics are evaluated. In Table 1, the authors of the present study presented the model Akaike information criterion (AIC) and Bayesian information criterion (BIC) values for the LC-SDT and criterion LC-SDT models (for the full model, see Table B4 in Online Appendix B). Across the four domains, the criterion model information criteria were consistently larger (at least 1,320 units) than the LC-SDT model. This indicates that no matter the selected information criterion, the LC-SDT model characterized the data better across domains, as indicated by smaller information criteria than the criterion model.

Comparing LC-SDT and criterion LC-SDT model discrimination. Perhaps more importantly than comparing the LC-SDT and criterion LC-SDT models in terms of information criteria or model fit indices is to evaluate the impact that the restriction had on the estimation of each operational rater’s discrimination parameter. As discrimination is most closely tied to the accuracy with which raters classify responses, it is of paramount importance for large-scale testing programs’ ability to monitor raters’ behavior and ensure that the best possible quality ratings are assigned by operational raters. Consider Figure 7, which shows the estimated discrimination on the Ideas domain by the operational raters under either the LC-SDT or the criterion LC-SDT model (also compare Tables B2 and B3 in Online Appendix B). Rater 10 exhibits roughly similar but

Figure 7. Rater discrimination for Ideas domain (d_{ij}) for the LC-SDT and criterion LC-SDT models.

Note. Posterior mode estimates of the LC-SDT model with Bayes constants of 1 for the four-latent-class solution. LC-SDT = latent-class signal detection theory.

relatively low discrimination under either the LC-SDT or the criterion LC-SDT model. This rater may be contrasted with Rater 5, whose discrimination under the criterion LC-SDT model is substantially lower than under the LC-SDT model.

It is these sorts of relative comparisons that represent the incremental utility of the criterion LC-SDT model over and above the already useful LC-SDT model. It is important to note that while Rater 10's discrimination appears to be relatively unaffected by the restriction of latent classification, his or her discrimination is low, relative to his or her peers. In other words, this similarity in the estimate of discrimination does not in and of itself indicate that Rater 10 is good in general, but that his relatively inaccurate ratings give rise to predicted latent classifications that are consistent with those of the criterion rater. Absent criterion ratings, such comparisons would not be possible, hence the usefulness of applying both the LC-SDT and the criterion LC-SDT models. Figures 8 through 10 present the same comparison of estimated discrimination for the remaining three domains: Style, Organization, and Conventions.

Discussion

Rater-mediated assessments with operational and criterion ratings are common in educational measurement. The present study illustrates the use of the LC-SDT model as applied to the task of incorporating expert (i.e., criterion) ratings as criteria for evaluating operational raters. One of the most commonly used statistics for evaluating ratings is kappa. Kappa is used to summarize agreement between a set of criterion ratings and operational ratings after a simple correction for chance agreement. The correction for chance agreement itself depends heavily on the marginal distributions of ratings (Agresti, 2002), which may yield quite different conclusions, depending upon how raters use the rating scale. Indeed, this summary statistic, while easy to calculate and widely used, does not give researchers or practitioners much in the way of diagnostic information to apply toward improving rating quality.

Figure 8. Rater discrimination for Style domain (d_{2j}) for the LC-SDT and criterion LC-SDT models.
Note. Posterior mode estimates of the LC-SDT model with Bayes constants of 1 for the four-latent-class solution. LC-SDT = latent-class signal detection theory.

Figure 9. Rater discrimination for Organization domain (d_{3j}) for the LC-SDT and criterion LC-SDT models.
Note. Posterior mode estimates of the LC-SDT model with Bayes constants of 1 for the four-latent-class solution. LC-SDT = latent-class signal detection theory.

Measures such as quadratic-weighted kappa characterize the relationship between criterion and operational ratings focus on agreement, when what ought to be the focus is the correct classification of CRs. If criterion and operational ratings both exhibit accurate classification, then

Figure 10. Rater discrimination for Conventions domain (d_{4j}) for the LC-SDT and criterion LC-SDT models.

Note. Posterior mode estimates of the LC-SDT model with Bayes constants of 1 for the four-latent-class solution. LC-SDT = latent-class signal detection theory.

they may also exhibit association (not necessarily agreement). However, association or agreement among criterion and operational raters should not be a goal in and of itself. Instead, one should endeavor to select a model that can incorporate ratings that, even if not exhibiting extensive agreement, may still give rise to accurate latent classification.

LC-SDT Model Including Criterion and Operational Ratets

The LC-SDT models were first estimated separately for each of the four analytic domains including both the criterion and operational raters. Regardless of the authors' philosophical stance on the genesis of the "expert" ratings, this model enabled them to confirm that the criterion rater was in fact superior to the operational raters in terms of his ability to discriminate between the four latent classes. This substantially higher d parameter for criterion or "expert" raters relative to operational raters has been seen in previous studies, such as DeCarlo (2002, p. 439). That perceptual part of the rating task (i.e., d_{ij}) varied both across raters and when comparing the Conventions domain with the other three, but the criterion rater's estimated discrimination was at least as great as that of any operational rater.

Where the criterion rater did not distinguish himself was in the mapping of those perceptions onto the rating scale. Specifically, the pattern of criteria (i.e., c_{ijk}) for this rater did not deviate much from the patterns exhibited by the operational raters; all raters gave ratings that tended to be very central (i.e., Category 2 or 3). Again, this finding is nothing new, as DeCarlo et al. (2011) noted that even expert raters displayed some level of centrality in their use of rating categories.

The results from this judgmental component of the LC-SDT model may simply reflect the fact that the rater who provided the criterion ratings trained many of the operational raters, and they were calibrated to his rating idiosyncrasies. This model was estimated as a means of

comparing the relative performance of the expert and operational raters and should not be compared with the subsequently estimated models, as they rely on different observed data, and hence are incommensurate when it comes to traditional information criteria.

Criterion LC-SDT Model Including Only Operational Raters

The authors' stated goal when considering criterion LC-SDT models was to determine their utility for monitoring rating quality when criterion ratings were present. Taking the analyses together, the authors of this study concluded based on Table 1 that the LC-SDT model better characterized the operational raters' behavior than the model that was restricted via the criterion rater. Most relevant to proposed application of this research are the comparisons presented in Figures 7 through 10. These may be interpreted as the accuracy with which raters assign rating categories when latent classification is either unrestricted (i.e., the LC-SDT model) or restricted to certain ranges conditional on the criterion rating (i.e., the criterion LC-SDT model).

The difference in the estimated discrimination across raters and domains indicates that the restriction has a different relative influence on the estimate of operational raters' accuracy. It is also important to note that there are cases where a rater may have high discrimination, when one ignores the criterion rating (consider the LC-SDT model estimate of the same rater's discrimination). When the true latent classification is grounded in the expert judgment of the criterion rating, the estimated discrimination may be substantially reduced (consider the criterion LC-SDT model estimate of Rater 3's Organization discrimination as shown in Figure 9).

Another key difference in the results when comparing the discrimination parameter for the LC-SDT model with those of the criterion LC-SDT model is the magnitude of the standard errors. Indeed, across domains and raters, the mean standard error for the discrimination parameter is 0.71 for the LC-SDT model, and it is 0.25 for the criterion LC-SDT model (compare the same cells in Tables B2 and B3 in Online Appendix B). This result calls for further investigation via a simulation study, which is beyond the scope of this article, but it should be taken into account when comparing the discrimination parameters across the LC-SDT and criterion LC-SDT models.

Comparison With Wind and Engelhard's Rasch Modeling Framework Results

As was pointed out earlier, these data were originally collected by Wind and Engelhard (2013), and analyzed using a Many-Facet Rasch (MFR) model approach (Linacre, 1989). Specifically, the original analysis included the use of a MFR model in which observed ratings were used as a dependent variable, and estimated rater locations reflected varying levels of rater severity/leniency. Table 2 summarizes a MFR model for rater accuracy, in which the comparison between operational and criterion raters served as the dependent variable, and estimated rater locations reflected varying levels of rater accuracy.

Of particular interest, Wind and Engelhard (2013) found that Rater 3 tended to be the most lenient (0.41 logits; see Column A in Table 2) and Rater 9 the most strict (-1.07). In terms of the results based on LC-SDT, the authors of this study also observed that these two raters have relative latent thresholds in Figures 3 through 6 that appear shifted. In particular, the more lenient Rater 3 had criteria that were, on average, greater than those of the remaining raters, particularly for Style and Conventions. The stricter Rater 9 had criteria that are marginally lower than his or her peers.

Wind and Engelhard (2013) also reported on rater accuracy, which corresponds to d_{ij} in the LC-SDT modeling framework. Raters 6 and 3 were identified as the most accurate operational raters (0.40 and 0.34 logits, respectively; see Column B in Table 2), while Rater 9 had the

Table 2. Abbreviated Results From Wind and Engelhard.

Rater	Rater severity	Rater accuracy
	Calibration (logits)	Calibration (logits)
Crit.	0.00	n/a
1	0.09	−0.15
2	0.21	−0.08
3	0.41	0.34
4	−0.30	0.27
5	−0.47	0.16
6	0.10	0.40
7	−0.06	−0.06
8	−0.46	−0.11
9	−1.07	−0.50
10	−0.14	−0.26

Source. Wind and Engelhard (2013).
Note. “Crit.” refers to the criterion rater. Rater severity calibrations describe rater severity/leniency on the Logit scale, where higher values indicate more severe raters and lower values indicate more lenient raters. Rater accuracy calibrations indicate rater accuracy in terms of the match between operational and criterion ratings, where higher values indicate exact agreement more often and lower values indicate exact agreement less often.

lowest estimated accuracy (−0.50 logits) among operational raters. In general, as shown in Figure 2, the two models corroborate the finding that Rater 9 was the least accurate operational rater (in terms of MFR model results) and least discriminating (in terms of LC-SDT results; $M(d_{i9})=2.91$). The comparison with the most accurate rater in terms of MFR model results was less cut-and-dried as with respect to the LC-SDT results. The most discriminating operational rater in terms of LC-SDT results was Rater 3 [$M(d_{i3})=4.84$], and this rater was the second most accurate in terms of MFR results.

Implications for Rater Monitoring and Training

The LC-SDT model results presented reinforce that model’s usefulness for monitoring rater behavior and appropriateness for identifying raters who need additional training. As DeCarlo (2002) notes, the best ways to increase classification accuracy are to increase (a) raters’ detection and (b) the number of raters with nonzero detection. Given that research funding or operational demands may constrain the number of raters tasked with rating each response, the researcher or practitioner interested in increasing classification accuracy would do well to focus rater training on increasing raters’ abilities to detect key features of the CRs that are outlined in the scoring rubric and that serve to differentiate those which belong to one latent class from the adjacent lower class.

The results from this study highlighted the utility of the LC-SDT approach for identifying differences in rating quality for individual raters using a criterion-referenced perspective. Specifically, this approach provides indicators of rater discrimination (e.g., Figure 2) that can be compared within individual raters and between operational and criterion raters. These results can be used to target rater remediation for individual raters or particular domains in which lower values of discrimination are observed. Along the same lines, LC-SDT model indicators of rater criteria for distinguishing among latent classes may inform rater training related to focusing on increasing the discrimination indices for each rater rather than standard agreement indices.

In addition, the data collection systems that many large-scale assessments currently use may readily be augmented by applying the LC-SDT model. No matter which measurement model is in place for the prediction of scores for a given assessment program, the LC-SDT model may be estimated for all operational raters, perhaps for every few 100 ratings they provide. By monitoring raters' estimated discrimination, one can identify large increases in rater discrimination—for rewards or leadership roles—or substantial decreases in rater discrimination—for retraining or identifying raters that are not suited to the rating task at hand.

In terms of the practical impact of this study, as the criterion LC-SDT grounds the true latent classification in the expert judgment of an external rater, testing organizations may use it to implement a richer system of rater monitoring. Absent criterion ratings, operational testing programs may apply existing LC-SDT methods, which themselves would represent a vast improvement over simple agreement statistics for monitoring rating quality. When criterion ratings are available, the criterion LC-SDT model proposed here enables testing organizations make additional comparisons with the LC-SDT model to further strengthen their rater monitoring procedures. Practitioners would do well to identify raters with highly discrepant detection parameter estimates, taking into account any differences in the relative precision of estimates, and more closely review their ratings relative to the criterion ratings.

One example of how the two models' results could be used operationally is to compare the rank ordering of raters' discrimination for the LC-SDT and the criterion LC-SDT models. In general and as shown in Figures 7 through 10, the rank ordering of discrimination estimates is fairly similar for the criterion LC-SDT when compared with the LC-SDT model. There were cases, in particular for the Style and Conventions domains where certain raters' discrimination ranks were fairly different. For example, in Figure 10, see Rater 1's discrimination for the Conventions domain. Under the LC-SDT that rater's discrimination parameter was estimated as 3.92 (third of 10 raters), while under the criterion LC-SDT it was 2.30 (eighth of 10 raters). First, the absolute magnitude of that rater's discrimination was much lower under the criterion LC-SDT, relative to the LC-SDT. Second and perhaps more importantly, his or her relative standing among operational raters is different. Hence, the action that an operational testing agency may take with respect to this rater could be different depending upon which model is estimated. Indeed, as the interest in this study is in comparing the operational with the criterion raters, one may be more inclined to rely on the criterion LC-SDT model results, relative to the LC-SDT model results.

Limitations and Direction for Future Research

This study represents an initial step in exploring the potential use of incorporating criterion ratings into LC-SDT models as a criterion-referenced technique for exploring rating quality. Additional research using both simulated and operational assessment data is needed to gain a more complete understanding of this technique, as well as its practical implications. In particular, several limitations and directions for future research are notable, and will inform the interpretation and use of this approach.

First, the use of restricted LC-SDT models for the task of incorporating criterion ratings requires additional development and application. In particular, simulation studies would do a great deal to identify how large a difference in terms of information criteria for the criterion and LC-SDT models indicate practically different predicted latent classification.

Along the same lines, research is needed to more fully understand the effects of simultaneously estimating parameters for operational and criterion raters. For example, the degree to which the estimated parameters for operational raters are influenced by the inclusion of criterion raters in general, along with characteristics of the criterion rater, is a topic that could

potentially be explored via simulation. Simulation studies could also be used to explore the degree to which differences between operational and criterion raters result in estimation bias.

Furthermore, while some research has investigated how sparse data affect parameter recovery in the LC-SDT model context (DeCarlo, 2010; Kim & Moses, 2013), it is not yet clear how rating design characteristics influence the recovery of rater parameters. In the present study, the criterion rater scored all of the essays, while the operational raters were systematically linked. Research related to the influence of connectivity among raters in general, as well as differences in rating designs across the criterion and operational raters in particular, will provide insight into the influence of data collection designs on the interpretation of results from LC-SDT models.

Future research is needed that expands the models presented in this study to include a comprehensive model for analytic scoring of CRs where those ratings need not be independent across domains. This study demonstrated the use of a restricted LC-SDT approach to evaluating rating quality within four separate domains. This analytic approach facilitated the exploration of diagnostic information about rating quality for individual raters within domains, and also demonstrated the application of this approach across multiple sets of ratings. Just as Nieto, Casabianca, and Junker (2016) extended Patz et al.'s hierarchical rater model framework to account for multidimensional ratings, DeCarlo et al.'s (2011) formulation of the hierarchical rater model could be extended to allow for the dependent structure that one cannot as a reasonable researcher deny may exist for a given rater's four domain-level ratings of the same CR. In other words, this would violate the model's second key assumption that "conditional on the vector of latent variables η , the ratings are independent" (p. 342), and a model that could relax this assumption would be a great contribution to this literature.

Finally, the relatively small sample of raters explored in this illustration could potentially limit the generalizability of the results. As noted above, this methodological illustration was focused on demonstrating the application of a new technique for incorporating criterion raters into a LC-SDT approach to rating quality analyses that resulted in diagnostic indices of rater accuracy at the individual rater level. Although the current rater sample size facilitated a concise illustration and the exploration of rating quality for each of the raters in the example dataset, additional research related to the use of restricted LC-SDT models should include larger samples of raters, along with the exploration of the utility of this approach in operational assessment settings.

Conclusion

This study applies a model that is strongly grounded in the psychological theory that gives rise to CR ratings to the task of characterizing rater behavior. Modeling CR ratings with the LC-SDT model enables us to separate the perceptual component of the rating task (i.e., discrimination) from the decision-based components (i.e., latent thresholds). Indeed, the LC-SDT model parameters themselves would be easy to incorporate into (a) confirmation of criterion raters' expertise, (b) the selection of raters for leadership roles, and (c) rater training, all of which are part and parcel of monitoring rater performance. Overall, the most novel aspect of this study approach was to apply a restricted LC-SDT model to incorporate criterion ratings into that framework. Indeed, by comparing individual raters' discrimination parameter estimates between the criterion and LC-SDT models, the authors of this study were able to make overall comparisons of the quality of predicted latent classifications when either constrained or unconstrained by the criterion ratings. In doing so, they added another tool for the researcher or practitioner to use for making model-based comparisons of operational and criterion ratings.

Acknowledgment

The authors acknowledge helpful feedback provided by Dr. Lawrence DeCarlo in his role as discussant during the 2015 Annual Meeting of the National Council on Measurement in Education, as well as the thoughtful suggestions of three anonymous reviewers.

Declaration of Conflicting Interests

The author(s) declared the following potential conflicts of interest with respect to the research, authorship, and/or publication of this article: I work remotely from my home office in northern New Jersey, so please use my home address, mobile phone, and Columbia e-mail address for any communication. As far as a fax number, I can provide a number if something absolutely needs to be faxed, but perhaps printing, scanning, and e-mailing will be an acceptable substitute for faxing.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

Note

1. Expression 1 is expressed in terms of a cumulative distribution function. Figure 1 displays the same model as conditional (i.e., not cumulative) distributions for illustrative purposes.

References

- Agresti, A. (2002). *Wiley Series in Probability and Mathematical Statistics. Categorical data analysis* (2nd ed.). New York, NY: Wiley-Interscience. doi:10.1002/0471249688
- Cronbach, L. (1955). Processes affecting scores on "understanding of others" and "assumed similarity." *Psychological Bulletin*, 52(3), 177. doi:10.1037/h0044919
- DeCarlo, L. T. (2002). A latent class extension of signal detection theory, with applications. *Multivariate Behavioral Research*, 37, 423-451. doi:10.1207/S15327906MBR3704_01
- DeCarlo, L. T. (2005). A model of rater behavior in essay grading based on signal detection theory. *Journal of Educational Measurement*, 42, 53-76. doi:10.1111/j.0022-0655.2005.00004.x
- 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. Retrieved from <https://www.ets.org/Media/Research/pdf/RR-10-08.pdf>
- 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. Retrieved from <http://www.ets.org/Media/Research/pdf/RR-10-08.pdf>
- DeCarlo, L. T., Kim, Y.-K., & Johnson, M. S. (2011). A hierarchical rater model for constructed responses, with a signal detection rater model. *Journal of Educational Measurement*, 48, 333-356. doi:10.1111/j.1745-3984.2011.00143.x
- Eckes, T. (2015). *Introduction to many-facet Rasch measurement: Analyzing and evaluating rater-mediated assessments* (2nd ed.). Frankfurt am Main, Germany: Peter Lang.
- Engelhard, G. (1996). Evaluating rater accuracy in performance assessments. *Journal of Educational Measurement*, 33, 56-70. doi:10.1111/j.1745-3984.1996.tb00479.x
- Engelhard, G., Jr. (1997). Constructing rater and task banks for performance assessments. *Journal of Outcome Measurement*, 1(1), 19-33.
- Engelhard, G., Jr. (2013). *Invariant measurement: Using Rasch models in the social, behavioral, and health sciences*. New York, NY: Routledge.
- Georgia Department of Education. (2011). *Georgia grade 8 writing assessment*. Atlanta, GA: Scoring Rubric.

- Johnson, R. L., Penny, J. A., & Gordon, B. (2009). *Assessing performance: Designing, scoring, and validating performance tasks*. New York, NY: The Guilford Press.
- Kim, S., & Moses, T. (2013). Determining when single scoring for constructed-response items is as effective as double scoring in mixed-format licensure tests. *International Journal of Testing, 13*, 314-328. doi:10.1080/15305058.2013.776050
- Linacre, J. M. (1989). *Many-Facet Rasch measurement*. Chicago, IL: MESA Press.
- Magidson, J., & Vermunt, J. K. (2005). Latent GOLD 4.0 [Software]. Belmont, MA: Statistical Innovations. Available from <http://www.statisticalinnovations.com/>
- Nieto, R., Casabianca, J. M., & Junker, B. W. (2016, April). *The hierarchical rater model for multidimensional rating rubrics*. Paper presented at the 2016 annual meeting of the National Council on Measurement in Education, Washington, DC.
- Sulsky, L. M., & Balzer, W. K. (1988). Meaning and measurement of performance rating accuracy: Some methodological and theoretical concerns. *Journal of Applied Psychology, 73*, 497-506. doi: 10.1037/0021-9010.73.3.497
- Wang, J., Engelhard, G., Jr., & Wolfe, E. W. (2015). Evaluating rater accuracy in rater-mediated assessments using an unfolding model. *Educational and Psychological Measurement*. Advance online publication. doi:10.1177/0013164415621606
- Wind, S. A., & Engelhard, G. (2013). How invariant and accurate are domain ratings in writing assessment? *Assessing Writing, 18*, 278-299. doi:10.1016/j.asw.2013.09.002
- Wind, S. A., & Peterson, M. E. (in press). A systematic review of methods for evaluating rating quality in language assessment. *Language Testing*.
- Woehr, D., & Huffcutt, A. (1994). Rater training for performance appraisal: A quantitative review. *Journal of Occupational and Organizational Psychology, 67*, 189-205. doi:10.1111/j.2044-8325.1994.tb00562.x
- Wolfe, E. W., & McVay, A. (2012). Application of latent trait models to identifying substantively interesting raters. *Educational Measurement: Issues and Practice, 31*(3), 31-37. doi: 10.1111/j.1745-3992.2012.00241.x
- Wolfe, E. W., Song, T., & Jiao, H. (2016). Features of difficult-to-score essays. *Assessing Writing, 27*, 1-10.