

Conceptualizing Rater Judgments and Rating Processes for Rater-Mediated Assessments

Jue Wang

University of Miami

George Engelhard, Jr.

University of Georgia

Rater-mediated assessments exhibit scoring challenges due to the involvement of human raters. The quality of human ratings largely determines the reliability, validity, and fairness of the assessment process. Our research recommends that the evaluation of ratings should be based on two aspects: a theoretical model of human judgment and an appropriate measurement model for evaluating these judgments. In rater-mediated assessments, the underlying constructs and response processes may require the use of different rater judgment models and the application of different measurement models. We describe the use of Brunswik's lens model as an organizing theme for conceptualizing human judgments in rater-mediated assessments. The constructs vary depending on which distal variables are identified in the lens models for the underlying rater-mediated assessment. For example, one lens model can be developed to emphasize the measurement of student proficiency, while another lens model can stress the evaluation of rater accuracy. Next, we describe two measurement models that reflect different response processes (cumulative and unfolding) from raters: Rasch and hyperbolic cosine models. Future directions for the development and evaluation of rater-mediated assessments are suggested.

Rater-mediated assessments are widely used in a variety of contexts. They can be broadly defined as any assessment system that requires human scoring. Most performance assessments are rater-mediated and provide the opportunity to measure complex aspects of student learning. As pointed out by Lane (2016), performance assessments can “better reflect students’ competencies in applying knowledge and skills to solve educationally meaningful tasks” (p. 356). Writing assessment is a popular type of performance assessments in educational settings and largely relies on rater scoring. We focus on writing assessments in this study for our illustrations.

Behizadeh and Pang (2016) indicated that 46 out of 50 states (92%) in the United States have essay items in their state writing assessments, and that 45 states fully depend on hand-scoring on student essays by human raters. One other state uses a combination of machine-scoring and human scoring in writing assessment. We believe the next wave of writing assessments will be based on automated assessment systems across the globe, but these systems still rely on human scoring for training of the scoring engine (Wind, Wolfe, Engelhard, Foltz, & Rosenstein, 2018). We also believe the development of automated scoring engines is “not to replace humans but to give them an important tool” (Hammond, 1996; Lopes & Oden, 1991, p. 201). Due to the involvement of human raters, it is important to develop systems for the evaluation of rater scoring behaviors and quality of their ratings.

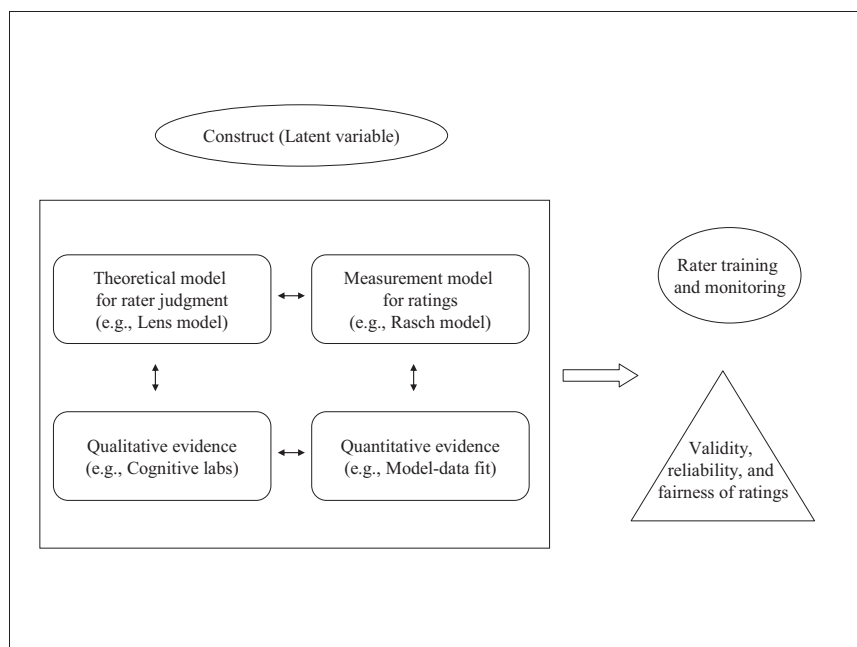


Figure 1. Conceptualizing rater-mediated assessments.

Figure 1 shows our conceptualization of selected aspects of rater-mediated assessments. As with all assessment systems, it is essential to begin with an idea of the intended construct. We argue that rater-mediated assessments require a theoretical model for rater judgments, as well as the identification of a congruent measurement model. Both quantitative and qualitative evidence provide support for validity arguments underlying intended uses of ratings in rater-mediated assessment systems. The qualitative information can further improve our understanding of rater cognitive processes and facilitate the development of human judgment models for rater-mediated assessments. These components combine to inform rater training and monitoring of the scoring processes. The ultimate goal is to assure that the rater-mediated assessments provide valid, reliable and fair ratings that reflect the meaning of the construct and support the intended uses of the ratings.

In the first section, we describe the use of lens models (Brunswik, 1955) for evaluating rater judgments. Next, we introduce two measurement models for analyzing ratings (i.e., Rasch model and hyperbolic cosine model). The third and fourth sections focus on the use of lens models for measuring two different underlying constructs—student writing proficiency and rater accuracy, as well as the applications of Rasch and hyperbolic cosine models for different response processes (cumulative and unfolding).

Theoretical Models for Rater Judgment in Scoring

Crisp (2012) defined scoring as “a process of comparing the representation of the meaning built from the student’s text with a mental representation built from reading

the scoring guidance or with an existing mental representation of what an ideal project should be like” (p. 11). The cognitive processes of raters can be evaluated with a variety of judgment models (Cooksey, 1996). Following Crisp’s definition, we have found Brunswik’s (1952) lens model to be a useful conceptual framework for rater-mediated assessments. The lens model approach is embedded within probabilistic functionalism framework (Athanasou & Kaufmann, 2015; Postman & Tolman, 1959), and this is an important aspect of Brunswik’s research (Hammond, 1955; Postman & Tolman, 1959). A lens model includes three components: an ecological system, a judgmental system, and a set of cues. As depicted in Panel A of Figure 2, the ecological system defines a distal variable that serves as the criteria. The researcher specifies the relationship between the distal variable and a set of cues that are selected to be indicators of the construct in the ecological system. A central response that is intended to empirically reflect the measurement of a distal variable is evaluated through the utilization of the cues in a judgmental system. Hammond (1955) adapted a lens model framework to evaluate clinical judgments, and since that time lens models have been widely applied in research studies on human judgments in the social sciences (Kaufmann, Reips, & Wittmann, 2013).

A key idea underlying lens models is the comparison of the utilization of cues in both the ecological system and judgmental system (Brunswik, 1952; Cooksey, 1996; Hammond, Hursch, & Todd, 1964; Hursch, Hammond, & Hursch, 1964; Tucker, 1964). For instance, in rater-mediated assessments, the mapping of features in text and the descriptions in rating rubrics may reflect rater cognitive processes in scoring tasks (Cumming, 1990; Milanovic, Saville, & Shuhong, 1996; Vaughn, 1991). The various interpretation of the cues may lead to discrepancies between two systems and further threaten the validity argument. As indicated by Brunswik’s (1952) judgment theory, whether the cues are noticeable by judges as well as the quality and importance of cues may affect empirical judgments (Crisp, 2012).

Cooksey, Freebody, and Davidson (1986) provided an informative example of using lens models for evaluating teacher judgments toward student reading proficiency. The scores of a standardized reading test were treated as the criteria of student reading proficiency in the ecological system. The ratings assigned by teachers based on their perceptions toward the students were the empirical judgments in the judgmental system. They used a set of cues including socioeconomic status, reading ability, and oral language ability. The relationship between criteria and the cues reflects ecological validity. The utilization of cues was examined using teachers’ ratings in the judgmental system. The correspondence between two systems indicated judgment accuracy of each teacher. By comparing these two systems, they found construct-irrelevant factors that influenced teacher judgments. Cooksey et al.’s study (1986) motivated us to use lens model design for evaluating rater judgment in rater-mediated assessments. Lens model designs help us depict rater judgment and decision making process in scoring activities. Multiple regression technique has been widely used in examining the usage of cues, and it assesses each individual separately (Hammond, 1996). We want to propose alternative methodological tools for quantitatively defining a lens model. The measurement models (Rasch models

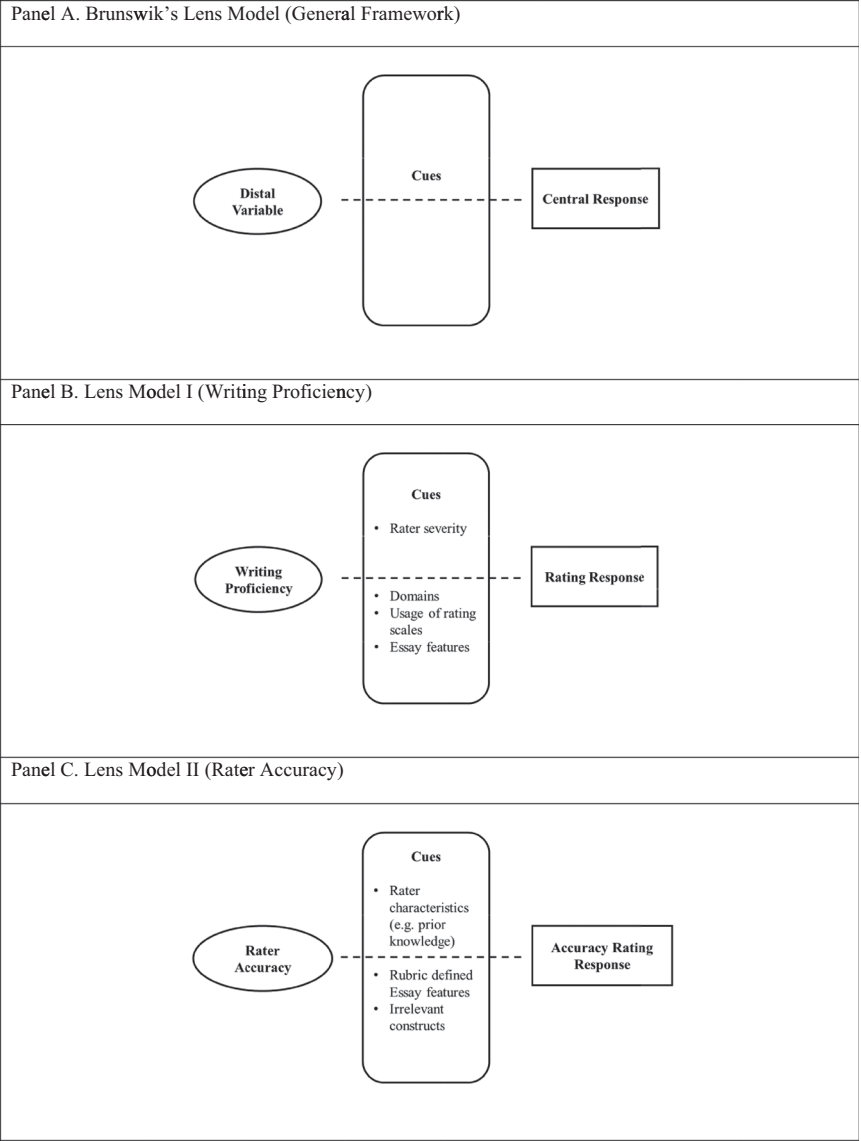


Figure 2. Lens models.

Note. Circle represents latent variable and square indicates observed responses. Cues, originally named as proximal peripheral cues by Brunswik (1955), mediate the measurement of construct.

and unfolding models) presented here focus on item-level and person-level analysis, which are good choices as methodological approaches for lens model studies.

Engelhard (2013) described two lens models for conceptualizing different constructs in rater-mediated assessments. Different from Cooksey et al. (1986) and the usual uses of lens model in social science, Engelhard (2013) specified a latent

variable (i.e., construct) as the distal variable in lens models. The first lens model focuses on the measurement of *student proficiency*, and it uses observed ratings that are assigned by human raters (Engelhard, 1992, 1994, 2013; Engelhard, Wang, & Wind, 2018). In this study, we label it as Lens Model I. Panel B of Figure 2 depicts this lens model within the context of writing assessments. Student writing proficiency is represented by observed ratings through a set of cues that may affect rater's empirical judgments, such as scoring severity, interpretation of analytic scoring rubrics, and category usage. Rater scoring severity is presented as a rater consistently assigns higher ratings than the criterion ratings (Wolfe, 2014). When rater effects exist, observed ratings may not accurately reflect student writing proficiency. Therefore, Lens Model I can be used to depict this scoring process, and to explore the factors that may influence a rater's scoring decisions including systematic rating errors and biases.

The second lens model (Engelhard, 1996, 2013) stresses the evaluation of rater accuracy by examining accuracy ratings (Figure 2, Panel C). Accuracy ratings directly reflect the distances between criterion ratings and observed ratings, specifically whether a rater scores a performance accurately and how accurate this rater is in the scoring activities. With the use of accuracy ratings, we can assess rater accuracy based on a continuum of latent scoring accuracy measures. Criterion ratings can be defined in different ways. One approach is to use the average scores across all raters as criterion scores (e.g., Wolfe & McVay, 2012). This can also be called the "Wisdom of the Crowds" approach. The other way is to have a group of expert raters resolve criterion scores through panel discussions (e.g., Engelhard, 1996). Engelhard and Wind (2017) defined expert raters as "highly trained raters who have deep understanding of the assessment system" (p. 80). The expert raters may also be in charge of the training for operational raters and monitoring of the scoring procedures. A lens model for rater accuracy is shown in Panel C of Figure 2, and we name it as Lens Model II. The underlying construct is *rater accuracy*, which is empirically reflected by observed accuracy ratings through a set of rater-specific factors such as prior scoring experiences, perception toward essay features, and individual uses of rating categories.

J. Wang and Engelhard (2017) also introduced a bifocal lens model in which the expert raters' ratings were viewed as the criteria in the ecological system and ratings of operational raters were empirical judgments. This lens model was formulated to examine the consistency in ratings between operational and expert raters, and it can further guide quantitative study for evaluating the discrepancies between mental models of expert raters and operational raters. There can be many variations of lens models based on the rater cognitive processes underlying the scoring decisions. The understanding and interpretation of cognitive processes of rater judgment require a selection and implementation of appropriate measurement models.

Measurement Models for Ratings in Rater-Mediated Assessments

As pointed out by Andrich and Luo (2017), Thurstone (1927a, 1927b, 1928) made an important distinction between cumulative and unfolding response processes. The Rasch model has been used frequently for modeling rater responses as a cumulative process, while unfolding response processes offer an approach for examining

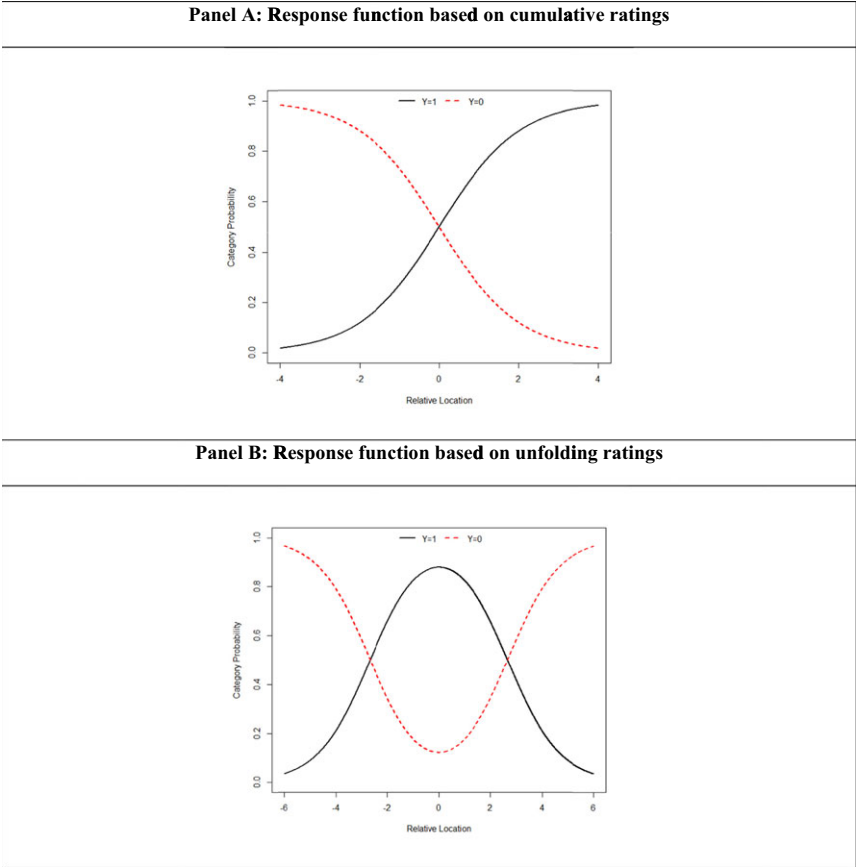


Figure 3. Response functions for cumulative and unfolding ratings. (Color figure can be viewed at wileyonlinelibrary.com)
Note. Relative location reflects the distance between essay and rater locations.

non-cumulative ratings. Figure 3 displays the response function curves for cumulative (Panel A) and unfolding (Panel B) response processes. Cumulative response function (Panel A) is monotonically increasing ($Y = 1$) or decreasing ($Y = 0$), while the unfolding function for category of one ($Y = 1$) is single-peaked (Panel B).

Tables 1 and 2 compare the response patterns that two measurement models are designed to analyze. Panel A in Table 1 shows a cumulative response pattern. Essays are ordered based on an underlying construct (writing proficiency) with Essay 1 reflecting the highest level of writing proficiency. It shows a perfect cumulative pattern (Guttman pattern), which follows an assumption that a rater who provides a score of zero for a lower ordered essay would not provide a score of one for those located above it on the scale. Since perfect Guttman patterns are not estimable, we added a dummy coded essay (i.e., Essay 7) and a dummy coded rater (Rater H) to reverse the Guttman pattern as suggested by Linacre (2018). Specifically, Rater H has a score of one for Essay 7 and zero for the actual six essays. Essay 7 receives a

Table 1
Analyze Dichotomous Cumulative Ratings Using the Rasch Model

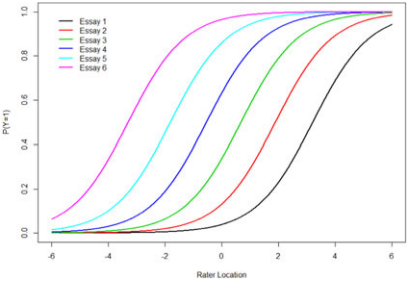
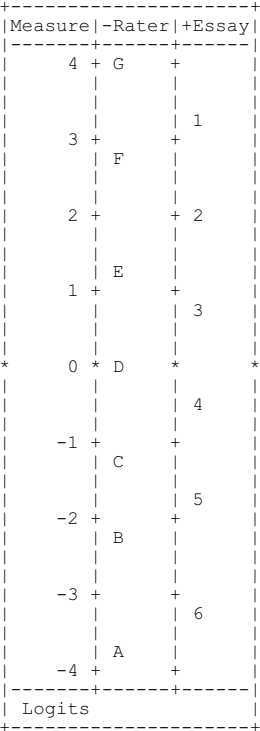
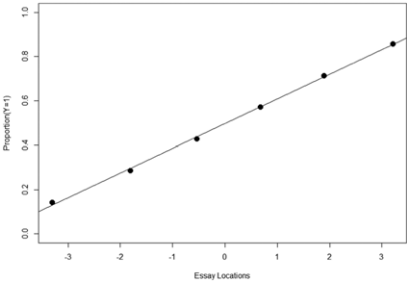
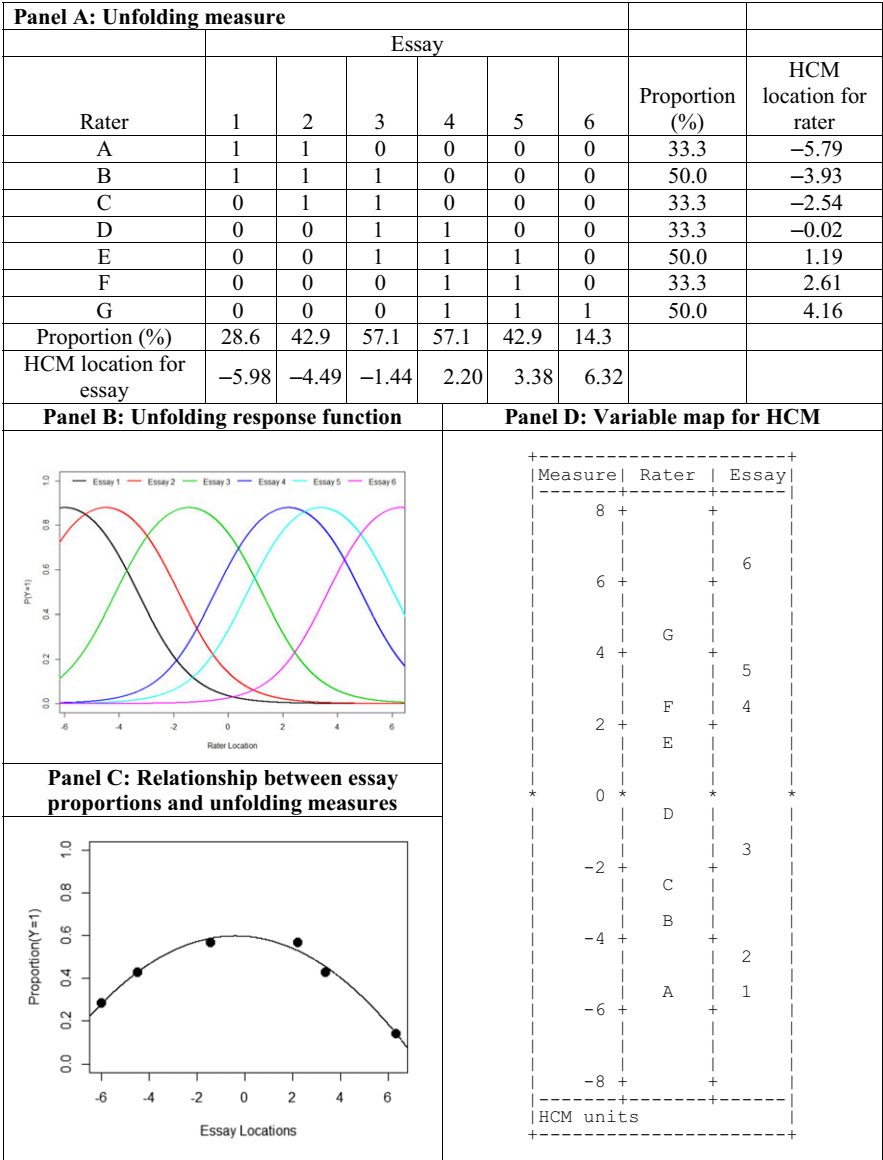
Panel A: Rasch measure								
	Essay							
Raters	1	2	3	4	5	6	Proportion (%)	Rasch location for rater
A	1	1	1	1	1	1	100.00	−3.66
B	1	1	1	1	1	0	83.33	−2.34
C	1	1	1	1	0	0	66.67	−1.13
D	1	1	1	0	0	0	50.00	.09
E	1	1	0	0	0	0	33.33	1.36
F	1	0	0	0	0	0	16.67	2.85
G	0	0	0	0	0	0	.00	4.54
Proportion (%)	85.71	71.43	57.14	42.86	28.57	14.29		
Rasch location for essay	3.21	1.89	.68	−.54	−1.81	−3.30		
Panel B: Cumulative response function					Panel D: Variable map for Rasch model			
								
Panel C: Relationship between essay proportions and Rasch measures								
								

Table 2
Analyze Unfolding Dichotomous Ratings Using Hyperbolic Cosine Model (HCM)



score of one from Rater H only but zero from actual raters. Panel B displays typical essay response functions based on the dichotomous Rasch model. The probability of getting a score of 1 is a monotonically increasing function. Panel C shows the relationship between the observed proportions and Rasch measures for essays, and it is linear. Panel D provides a variable map (a.k.a. Wright Map) showing the relative locations of raters and essays.

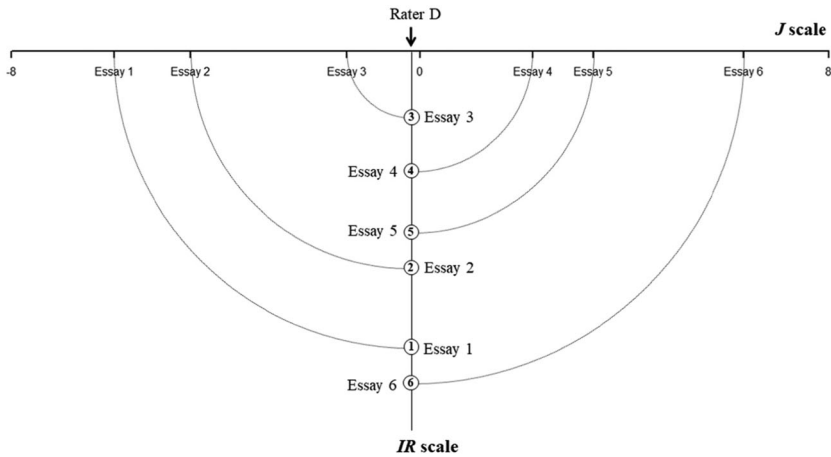


Figure 4. The joint (*J*) scale with individual rater (*IR*) scale for Rater *D*.

Panel A in Table 2 shows an unfolding response pattern that appears as a parallelogram. This pattern represents an underlying unfolding scale that assumes each rater assigns a score of one to a subset of the essays located near the rater's location on the unfolding scale. Panel B presents essay response functions based on the hyperbolic cosine model (HCM; Andrich, 1988; Andrich & Luo, 1993) for dichotomous responses. The relative distances between raters and an essay determines the probability of endorsing (i.e., having a score of 1) this essay. This probability is the highest when the absolute distance between them is the smallest, and it decreases as the distances increase. These models are sometimes referred to as ideal-point item response theory models with single-peaked response functions (Maydeu-Olivares, Hernández, & McDonald, 2006). Panel C shows the relationship between essay proportions and HCM essay measures. The second-order polynomial curve reflects the distinctive unfolding aspect of the HCM. Panel D gives the variable map with ordered essays and raters based on HCM measures. It reflects the preference proximity of essays to the raters. For instance, Rater G assigned a score of 1 to Essays 4, 5, and 6 so that Rater G locates approximately at the mid-point of these three essays' locations.

In addition to the variable map for the HCM, it is informative to examine the preference ordering of an individual rater. The variable map displayed in Panel D of Figure 2 can also serve as an empirical representation of the *joint (J) scale*. Figure 4 depicts the *J scale* and shows the locations of ordered essays and raters together with the formation of the *individual rater (IR) scale* for Rater D. As can be seen, The *IR scale* is created by folding *J scale* at Rater D's location (i.e., ideal point). Figure 5 further includes the *IR scales* for Rater A and Rater G. An *IR scale* shows a rater's preference ordering of essays as summarized in the box at the bottom. The *IR scales* can be different for raters at different locations on the *J scale*. For instance, Rater A had the highest probability of endorsing Essay 1 and lowest for Essay 6. Conversely, Rater G had higher probability of endorsing Essay 6 over Essay 1.

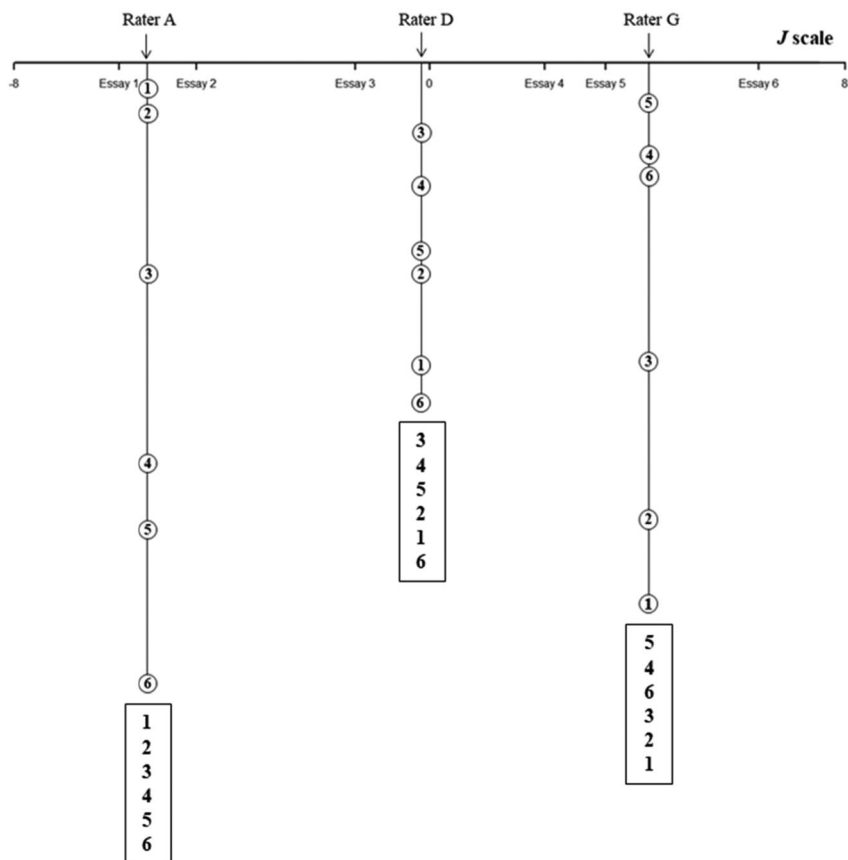


Figure 5. The joint (J) scale with individual rater (IR) scale for Raters A, D, and G.

Assessing Student Writing Proficiency

Lens Model I, as depicted in Panel B of Figure 2, focuses on the measurement of student writing proficiency. In writing assessments, student writing proficiency is a latent variable that cannot be observed directly. The observed rating responses assigned by human raters are empirical judgments in the lens model. The quality of judgments may be affected by a set of cues such as rater severity, centrality, essay features, domains, genres, and rating categories. Lens Model I conceptualizes writing proficiency as the underlying construct (distal variable) for student or essay facet. The underlying rating response process determines the choice of an appropriate measurement model.

The underlying construct for the rater facet can be different based on cumulative and unfolding response processes. Impersonal judgments are typically reflected in a cumulative response process, and these can be evaluated using a cumulative model such as Rasch measurement models. The unfolding response process may reflect personal preference and requires the use of an unfolding model (e.g., HCM) for

scaling purposes. Andrich and Luo (2017, p. 2) illustrated the distinction between impersonal judgments and personal preferences using an example as follows:

Consider the stimuli to be cups of coffee identical in all respects except for fine gradations of the amount of sugar in them. Two different instructions can be given for making comparative selections. Instruction I: Select the cup in each pair which has more sugar; Instruction II: Select the cup in each pair that you prefer.

The rater-mediated assessments seem to follow the first type of instruction given an established scoring rubric. We expect raters share consistent understanding of different levels of writing proficiency reflected in essays and provide congruent impersonal judgments in the decision-making process. Impersonal judgments refer to empirical judgments made by raters according to a well-established rubric. It is clear that impersonal judgments are desired in rater scoring activities; however, human raters may still be influenced by their own characteristics and unique prior experiences so that personal preferences may affect their ratings.

A cumulative response process may be assumed when raters are thoroughly trained to be highly competent and consistent in scoring student performances. In other words, an underlying cumulative response pattern (e.g., Guttman pattern) is assumed (Table 1, Panel A), and individuals that deviate from an ideal Guttman pattern are treated as misfit raters or essays. In rater-mediated assessments, a variety of measurement models assuming cumulative response patterns have been used for detecting rater effects, such as the many-facet Rasch model (Linacre, 1989), the hierarchical rater model (Casabianca, Junker, & Patz, 2016; Patz, Junker, Johnson, & Mariano, 2002), the rater bundle model (Wilson & Hoskens, 2001), and the generalized rater model (W. C. Wang, Su, & Qiu, 2014). Wolfe, Jiao, and Song, (2015) provide a summary and comparison on the performance of Rasch-based measurement models on evaluating rater effects. Wolfe (2014) especially described the use of partial credit model (PCM; Masters, 1982) with a unique rater threshold parameter in detecting rater effects such as severity/leniency and centrality/extremity. The model specification can be shown as follows for rater-mediated writing assessments:

$$\ln \left(\frac{\pi_{ij,k}}{\pi_{ij,(k-1)}} \right) = \theta_i - \lambda_j - \tau_{jk}, \quad (1)$$

where

$\pi_{ij,k}$ = probability of getting a score of k for essay i assigned by rater j ;
 $\pi_{ij,(k-1)}$ = probability of getting a score of $k - 1$ for essay i assigned by rater j ;
 θ_i = proficiency level reflected by essay i ;
 λ_j = severity of rater j ;
 τ_{jk} = rater threshold parameter for rater j .

The essay estimates are latent scores representing student writing proficiency. Rasch estimates for raters are on a continuum of harshness in scoring. In addition, the standard deviation of rater threshold estimates reflects the degree of centrality/extremity.

Unfolding models (Coombs, 1964) are mainly applied in psychological studies, especially in the measurement of attitudes and preferences. The application of

unfolding models in rater-mediated assessments are novel and promising (J. Wang, Engelhard, & Wolfe, 2015; J. Wang & Engelhard, 2016, 2019). Unfolding models are designed to analyze data responses that exhibit an unfolding pattern (Table 2, Panel A). The evaluation of personal preferences provides information on individual differences in their scoring decisions. The purpose of rater training and monitoring, in this case, can be viewed as reducing the impact of personal preferences in rater scoring decisions. J. Wang and Engelhard (2019) proposed the use of a hyperbolic cosine model for polytomous responses (HCM-P; Andrich, 1996; Luo, 2001) for this process, and HCM-P can be expressed as follows:

$$P(X_{ij} = k) = \frac{[\cosh(\theta_i - \lambda_j)]^{m-k} \prod_{l=1}^k \cosh(\rho_{jl})}{\sum_{k=0}^m [\cosh(\theta_i - \lambda_j)]^{m-k} \prod_{l=1}^k \cosh(\rho_{jl})}, \quad (2)$$

when, $k = 0$, $\prod_{l=1}^k \cosh(\rho_{il}) \equiv 1$;
where

$k = 0, \dots, m$, and m is the number of rating categories;

X_{ij} = observed rating score given by rater j to student essay i ;

θ_i = proficiency level reflected by essay i ;

λ_j = preference location of rater j ;

ρ_{jl} = rater threshold parameter; these threshold parameters are constrained to be equally distanced across categories, and this distance is called as *latitude of preference* for rater j .

A J scale can show ordered essays and raters based on their HCM-P estimates. Ratets are located closer to the essays that they prefer and have assigned higher scores. Furthermore, we can construct IR scales that reflect individual rater preference toward student essays (e.g., Figure 4). Each rater may have a unique ordering of essays since unfolding models reveal individual differences (e.g., Figure 5). It is worth noting that the threshold parameter is associated with rater facet so that we can explore a *latitude of preference* toward the essays for each rater. Due to unfolding process, the relationship between HCM-P essay measures and essay proportions would appear as a second-order polynomial curve (e.g., Table 2, Panel C) when good model-data fit is met. That said, student essays that reflect high writing proficiency levels are located in the middle and those showing lower proficiency levels are separated into two directions. This allows us to further explore different reasons that those essays received low scores.

With the use of unfolding models, we can detect whether personal preference is involved or not. If IR scales show very similar preference orderings of essays, then it indicates minimal personal preferences on scoring decisions across raters. If unfolding models suggest very different IR scales among raters, personal preferences of raters should not be ignored. By using Rasch measurement models, we assume that a common IR scale can be applied for all the raters and examine idiosyncratic response patterns that deviated from Guttman's triangular pattern using model fit indices. With the use of unfolding models, we allow the existence of individual differences among raters in scoring essays and compare how different or similar of their unique preference orderings toward student essays that are reflected by IR scales.

Table 3
An Illustration of Computation Procedure for Accuracy Ratings

Ratings	Essays				
	1	2	3	4	5
Observed	2	1	3	0	2
Criterion	2	0	1	3	3
Difference ($O_{ij} - C_i$)	0	1	2	-3	-1
Absolute difference ($ O_{ij} - C_i $)	0	1	2	3	1
Accuracy	3	2	1	0	2

Evaluating Rater Accuracy

Lens Model II emphasizes the evaluation of rater accuracy. Rater accuracy is the distal variable (construct) in the lens model, and accuracy ratings are central responses (Figure 2, Panel C). Engelhard (1996, 2013) defined rater accuracy as a latent variable that can be measured objectively using latent trait models with accuracy ratings. Accuracy ratings directly represent the difference between operational ratings and criterion ratings (Engelhard, 1996). We recommend having a group of expert raters to provide resolved scores as criterion ratings (e.g., Engelhard, 1996).

Next, we show the calculation of accuracy ratings with a hypothetical example. The accuracy ratings can be computed using the following formula (Engelhard, 1996):

$$A_{ij} = \max(|O_{ij} - C_i|) - |O_{ij} - C_i|, \quad (3)$$

where

A_{ij} = accuracy rating for operational rater j on essay i ;
 O_{ij} = Observed rating of operational rater j on essay i ;
 C_i = Criterion rating for essay i .

For example, a rater scored five essays using categories 0 to 3, and the maximum absolute difference between observed and criterion ratings across all raters and essays is 3. We first compute the difference between observed ratings and criterion ratings for this rater on each essay and then take the absolute values. Finally, the accuracy ratings can be obtained by subtracting each absolute difference value from the maximum absolute difference (i.e., 3). A higher accuracy rating reflects higher degree of empirical accuracy. This procedure is illustrated in Table 3.

In operational settings, we may not observe large difference between observed and criterion ratings. A study on evaluating the effectiveness of rater training methods conducted by Raczynski, Cohen, Engelhard, and Lu (2015) obtained accuracy ratings using this method, and they found 99% of the accuracy ratings were either zero or one. In practice, dichotomous accuracy ratings can be more common especially for well-trained raters. A short-cut for obtaining dichotomous ratings is to directly match observed and criterion ratings. If matched, a value of one indicating accurate is assigned; otherwise, a code of zero representing inaccurate is given.

In addition to the underlying construct, with accuracy ratings, Lens Model II can be used to consider rater characteristics such as language background and educational training (Caban, 2003), prior scoring experience (Davis, 2016), rater personality (Carrell, 1995), and individual differences in the effects of training (Lumley & McNamara, 1995), as well as rubric-defined and irrelevant essay features as potential cues that may affect rating accuracy. In particular, J. Wang, Engelhard, Raczynski, Song, and Wolfe (2017) investigated rater perception toward difficult-to-score essays and found that rater accuracy in scoring activities is affected by certain essay features such as amount of textual borrowing and essay length.

The response process of accuracy ratings needs to be explored. A cumulative response process has been commonly assumed in evaluating rater accuracy. In addition, J. Wang et al. (2015) indicated the possibility of an unfolding response process underlying rater accuracy. The choice of appropriate measurement models for analyzing accuracy ratings depends on the response pattern. Cumulative models assume a Guttman scale so that the essay order based on their difficulty for raters to score accurately is the same for every rater. Unfolding models are based on Coombs's unfolding scale, which reflects individual differences in their scoring accuracy toward different essays. In other words, unfolding models assume that the difficulty in ordering of essays may vary across raters. For instance, the mechanical errors in writing may affect a few raters' scoring accuracy but not others on the meaning/content of an essay (J. Wang et al., 2017); therefore, essays with mechanical errors might be more difficult for some raters to score accurately than others.

Next, we present a representative cumulative model and a recommended unfolding model for analyzing accuracy ratings, and then illustrate their uses through an empirical data analysis.

To model cumulative response processes, the many-facet Rasch model (Linacre, 1989) is proposed by Engelhard (1996) to examine rater accuracy. This model is also named the rater accuracy model (RAM; Wolfe, Song, & Jiao, 2016), because it provides rater accuracy estimates on a latent continuum. The RAM can be specified as follows:

$$\ln \left(\frac{\pi_{ij,k}}{\pi_{ij,(k-1)}} \right) = \delta_i - \lambda_j - \tau_k, \quad (4)$$

where

$\pi_{ij,k}$ = probability of receiving an accuracy rating k on essay i for rater j ;

$\pi_{ij,(k-1)}$ = probability of receiving an accuracy rating $k - 1$ on essay i for rater j ;

δ_i = difficulty of essay i to be scored accurately;

λ_j = accuracy of rater j ;

τ_k = difficulty of reaching category k relative to category $k - 1$ of accuracy ratings.

For dichotomous accuracy ratings (0, inaccurate; 1, accurate), the RAM is reduced to be a dichotomous Rasch model (Rasch, 1960/1980) in the following form:

$$\ln \left(\frac{\pi_{ij}}{1 - \pi_{ij}} \right) = \delta_i - \lambda_j. \quad (5)$$

Based on the RAM, the essays are ordered from least difficult to most difficult for raters to score accurately, and raters are ordered from least accurate to most accurate in scoring. Engelhard (1996) suggested fit indices for evaluating the performance of the RAM in analyzing accuracy ratings. For instance, the mean square and standardized infit and outfit statistics for rater facet can be used to diagnose an individual rater who has an irregular response pattern. The reliability of separation and chi-square test can examine whether there are significant differences among rater accuracy measures.

Unfolding models can be used when the rank orderings of essays based on difficulty-to-score are assumed to be different between raters. J. Wang et al. (2015) suggested using the hyperbolic cosine model with an essay threshold parameter for examining accuracy ratings. In this study, we call it as the hyperbolic cosine accuracy model (HCAM):

$$P(X_{ij} = k) = \frac{[\cosh(\delta_i - \lambda_j)]^{m-k} \prod_{l=1}^k \cosh(\rho_{il})}{\sum_{k=0}^m [\cosh(\delta_i - \lambda_j)]^{m-k} \prod_{l=1}^k \cosh(\rho_{il})}, \quad (6)$$

when $k = 0$, $\prod_{l=1}^k \cosh(\rho_{il}) \equiv 1$;

where

$k = 0, \dots, m$, and m is the number of categories of accuracy ratings;

X_{ij} = observed accuracy rating received by rater j on essay i ;

δ_i = difficulty of essay i to score accurately;

λ_j = accuracy location of rater j ;

ρ_{il} = essay threshold parameter, these threshold parameters are constrained to be equally distanced across accuracy rating categories; the equal distance between any two categories is viewed as *zone of accuracy* for essay i .

With dichotomous accuracy ratings, the HCAM is reduced to be a dichotomous hyperbolic cosine model (Andrich & Luo, 1993), which can be specified as follows:

$$P(X_{ij} = 1) = \frac{\cosh(\rho_i)}{\cosh(\rho_i) + \cosh(\delta_i - \lambda_j)}. \quad (7)$$

The unfolding scale (i.e., J scale) orders essays and raters based on their HCAM measures. Raters are located near the essays that they scored more accurately. To evaluate the fit between accuracy data and HCAM, a Pearson chi-square statistic can be used. It quantifies the discrepancies between observed proportions and model-based probabilities by dividing raters into intervals. A nonsignificant result indicates acceptable overall fit. A likelihood ratio test may also be employed to examine whether essay threshold parameters should be constrained as equal across all essays. A significant result of likelihood ratio test suggests estimating unique essay threshold parameters.

Empirical Example Analyzing Accuracy Ratings Using the RAM and HCAM

J. Wang et al. (2015) conducted an empirical data analysis based on a writing assessment and illustrated the use of HCAM with dichotomous accuracy ratings. The

Table 4
Summary of Statistics for the Rater Accuracy Model (RAM)

	Essay	Rater
Location measures		
<i>M</i>	.00	.78
<i>SD</i>	1.20	.36
<i>N</i>	50	20
Infit MnSq		
<i>M</i>	1.00	1.00
<i>SD</i>	.08	.14
Outfit MnSq		
<i>M</i>	.98	.98
<i>SD</i>	.17	.24
Reliability of separation		
	.75	.54
Chi-square test	$\chi^2(49) = 157.4, p < .05$	$\chi^2(19) = 39.8, p < .05$
% of variance explained	22.59%	

Note: MnSq, mean square fit statistic.

writing data was based on a large-scale state writing assessment and was originally used in Gyagenda and Engelhard (2010). Fifty essays written by eighth-grade students were scored by a random group of twenty raters. An analytic rating scale was applied to score four domains, and criterion ratings were obtained from a panel of experts. The dichotomous accuracy ratings were computed for the domain Style and were analyzed using a dichotomous HCAM which is shown in Equation 7. In this study, we would like to stress the key findings and illustrate the use of both RAM and HCAM for analyzing cumulative and unfolding response processes in rater-mediated assessments. The analysis of RAM is conducted in the FACETS computer program (Linacre, 2018), and HCAM is fitted using the RateFOLD computer program (Luo & Andrich, 2003).

Table 4 shows Rasch measures obtained with RAM. The infit and outfit mean square statistics are generally acceptable. The reliability of separation for essays is .75 and for rater facet is .54. According to the significant chi-square test results, essays cannot be assumed to be equally difficult, and raters shall not be viewed as equally accurate. The variable map for the constructed Rasch scale is displayed in Panel A of Figure 6. Essays are ordered from the least to most difficult to score. Correspondingly, raters are ordered from the least to most accurate in scoring. Raters are relatively accurate in scoring this set of essays. The least accurate rater, Rater 19, has scored 48% of essays accurately.

Table 5 displays the HCAM measures for essays and raters. Based on the chi-square test, the overall fit of the model is acceptable. The likelihood ratio test is not significant, suggesting a more parsimonious model with a common essay threshold parameter. Panel B in Figure 6 shows a variable map for unfolding scale. Each rater is located closer to the essays that (s)he scored more accurately. Figure 7 displays the relationship between essay proportions of accurate ratings (i.e., accuracy rates) and HCAM essay measures, which appears to be a second-order polynomial curve

Table 5
Summary of Statistics for Hyperbolic Cosine Accuracy Model (HCAM)

	Essay	Rater
Location measures		
<i>M</i>	.00	−.85
<i>SD</i>	2.86	.62
<i>N</i>	50	20
Threshold estimate		
<i>M</i>	3.51	NA
<i>SD</i>	.00	NA
Likelihood ratio test for common thresholds	$\chi^2(48) = 9.19, p > .99$	NA
Pearson chi-square test for overall fit	$\chi^2(949) = 927.78, p = .68$	

Note. NA, not applicable.

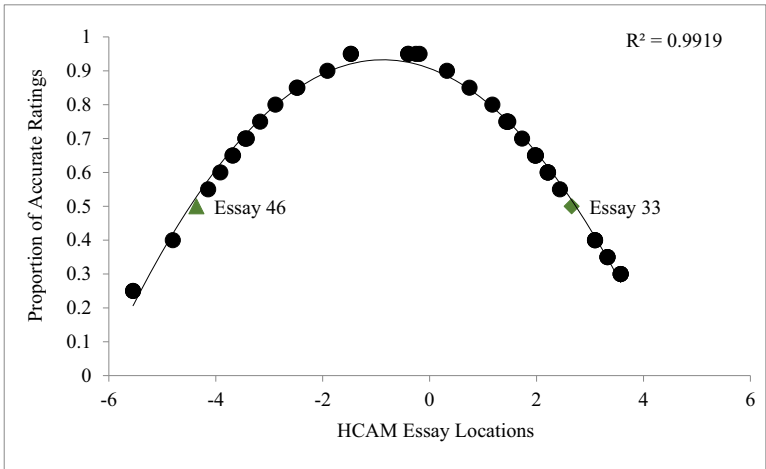
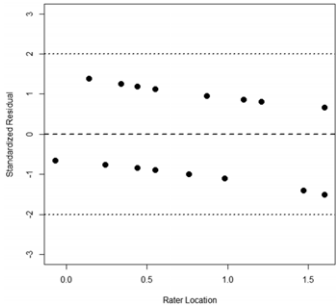
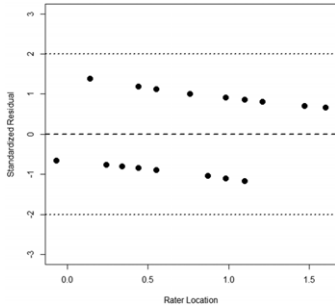
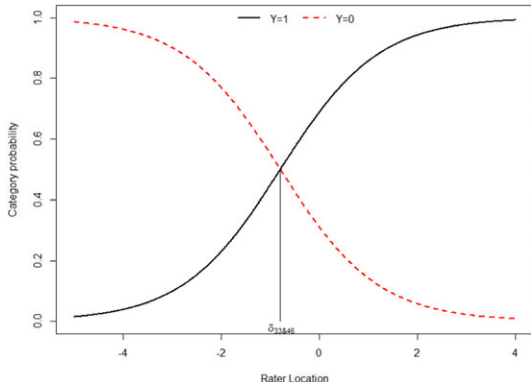


Figure 7. Relationship between accuracy rates and essay locations based on hyperbolic cosine accuracy model (HCAM). (Color figure can be viewed at wileyonlinelibrary.com)

As shown in Table 6, the Rasch location estimates and standard errors based on RAM for Essays 46 and 33 are the same. These two essays also share the same probability function curve. The infit and outfit mean square fit indices for these two essays are not the same, but they are all close to 1.00, indicating good fit to the Rasch scale. After ordering the Rasch measures of 20 raters from the least to most accurate, we obtained the response patterns for these two essays. As shown, the response patterns are quite different from each other and have some deviations from a perfect Guttman pattern. Standardized residuals are also plotted against RAM rater location estimates. The patterns are similar for both essays, and the values are all within ± 2 .

Based on HCAM, Essays 46 and 33 have different location estimates (Table 7). They both have good fit to the unfolding scale based on the individual chi-square

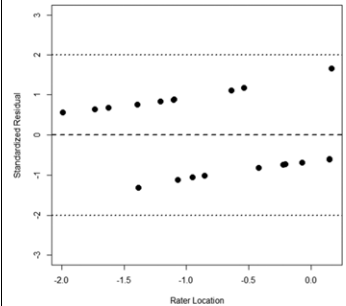
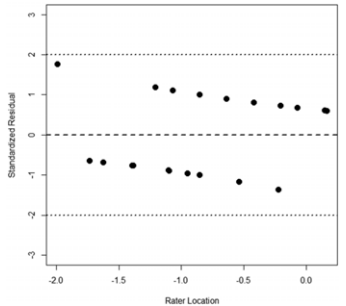
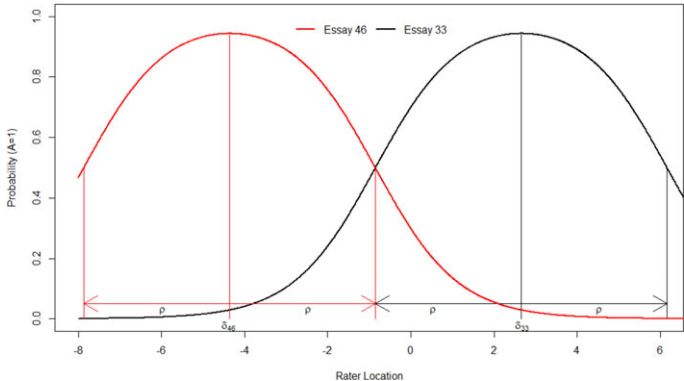
Table 6
Rater Accuracy Model (RAM) Measures for Essays 33 and 46

	Essay 46	Essay 33
Accuracy rate	.50	.50
Location measure	−0.79	−0.79
S.E.	0.46	0.46
Infit MnSq	1.05	0.91
Outfit MnSq	1.05	0.90
Response pattern	01001010101001111010	01000101010010011111
Standardized residuals		
<p>Probability function curves</p> 		

Note: Accuracy rate indicates the proportion of accurate ratings; the response pattern is ordered based on rater location measures of RAM.

test for each essay. The separated essay probability functions are plotted along the underlying unfolding scale (i.e., *J* scale). The zone of accuracy for each essay is calculated as essay location ± essay threshold parameter. Raters who locate within the zone of accuracy that have probabilities higher than .50 score this essay accurately. After ordering raters based on HCAM measures, we obtained two distinct essay response patterns. For Essay 46, more are on the left end, whereas for Essay 33 more are on the right end. We also observe deviations from ideal unfolding pattern. Plots

Table 7
Hyperbolic Cosine Accuracy Model (HCAM) Measures for Essays 33 and 46

	Essay 46	Essay 33
Accuracy rate	.50	.50
Location measure	−4.36	2.66
S.E.	0.23	0.23
Threshold estimate	3.51	3.51
Chi-square test	$\chi^2(19) = 18.00, p = .52$	$\chi^2(19) = 18.72, p = .47$
Response pattern	11110111000011000001	10000100100110101111
Standardized residuals		
<p>Probability function curves</p> 		

Note: Accuracy rate indicates the proportion of accurate ratings; the response pattern is ordered based on rater location measures of HCAM; A = 1 indicates accuracy rating is 1.

of standardized residuals are displayed showing that all values are within ± 2 and the patterns are different.

With the inclusion of external criteria such as rater characteristics and essay feature indices, we can explore the substantive interpretation of either a Rasch or an unfolding scale for accuracy ratings. This facilitates the development of Lens Model II for improving the evaluation of rater accuracy.

Discussion, Implications, and Future Directions

This study provides a framework for building theoretical rater judgment models to conceptualize an underlying construct, and selecting an appropriate measurement model to reflect rater response process. Within the context of rater-mediated assessments, two theoretical rater judgment models based on Brunswik's lens model are specified for different underlying constructs. In terms of psychometric models, Engelhard et al. (2018) discussed the use of Rasch measurement models for both the measurement of student writing proficiency and rater accuracy. The current study introduces unfolding models (i.e., hyperbolic cosine models) into this framework for evaluating the rating responses. In addition, McIver and Carmines (1981) recommend that the selection of a measurement model needs to consider the type of data and intended use of scores, as well as substantive theory of the decision-making process of the rater scoring activities.

Table 8 provides a summary of key concepts and measurement models described in this study. The illustrative articles that have used these models are also presented. Lens Model I emphasizes the measurement of student writing proficiency and analyzes observed ratings. The PCM can be used for a cumulative response process to provide objective measures for student writing proficiency as well as to evaluate impersonal judgments among raters. The underlying scale reflects impersonal judgment toward essays with essays ordered from the least to the most proficient in writing and raters ordered based on their scoring severity. A requirement for assuming cumulative response process (i.e., Guttman pattern) is that a single ordering of the essays is applied to all raters. Based on this requirement, three key research questions can be raised: (a) *Does the essay ordering based on raters' impersonal judgments conform to a cumulative scale?* (b) *Do raters vary in their scoring severity?* and (c) *What is a rater's usage of the rating categories?* It is worth emphasizing that a unique threshold structure for rater facet is included in the model estimation.

The HCM is suggested for unfolding response process to obtain a joint preference ordering of the essays (i.e., *J* scale) among raters. Essay locations on the *J* scale shows writing proficiency based on rater preference toward essays. Raters are located near the essays that they have assigned high scores (i.e., essays that they prefer). Furthermore, we can construct *IR* scales to reflect each individual rater's preference in order of essays. A requirement of constructing an unfolding scale is that *IR* scales reflecting individual preferences toward essays can be unfolded to conform to a *J* scale. In particular, three key research questions can be addressed with the use of HCM: (a) *Can essay orderings reflecting personal preferences be unfolded to construct a J scale?* (b) *What is the ideal point for a rater to have the highest preference?* and (c) *What is a rater's latitude of preference toward essays?* The

Table 8
Summary of Key Concepts

	Construct (Distal Variable)			
	Writing proficiency (Lens Model I: Observed ratings)		Rater accuracy (Lens Model II: Accuracy ratings)	
Measurement models	Partial credit model (PCM)	Hyperbolic cosine model (HCM)	Rater accuracy model (RAM)	Hyperbolic cosine accuracy model (HCAM)
Underlying response process	Cumulative	Unfolding	Cumulative	Unfolding
Underlying scale for essay facet	Writing proficiency (judgment toward essays)	Writing proficiency (preference toward essays)	Difficulty to score accurately (applicable to all raters)	Difficulty to score accurately (relative to each rater)
Underlying scale for rater facet	Impersonal judgment (rater severity)	Personal preference	Rater accuracy (based on consistent essay difficulty ordering)	Rater accuracy (based on individual preference in essay difficulty ordering)
Research question for essay facet	Does the essay ordering based on impersonal judgment conform to a cumulative scale?	Can essay orderings reflecting personal preference be unfolded to construct a J scale?	Does an essay difficulty ordering that is consistent across all raters conform to a cumulative scale?	Can essay difficulty orderings that vary across raters be unfolded to construct a J scale?
Research question for rater facet	Do raters vary in their scoring severity?	What is the ideal point for a rater to have the highest preference?	Do raters vary in their scoring accuracy?	What is the ideal point for a rater to be the most accurate?
Research question for essay threshold	NA	NA	What is the difficulty of using a category relative to an adjacent category?	What is the accuracy zone for each essay?

(Continued)

Table 8
Continued

	Construct (Distal Variable)			
	Writing proficiency (Lens Model I: Observed ratings)		Rater accuracy (Lens Model II: Accuracy ratings)	
Research question for rater threshold	What is a rater's usage of rating categories?	What is a rater's latitude of preference toward essays?	NA	NA
Illustrative article	Wolfe (2014)	J. Wang and Engelhard (2019)	Engelhard (1996)	J. Wang, Engelhard, and Wolfe (2015)

Note. NA, not applicable.

latitude of preference for each rater is obtained through rater threshold parameter in HCM.

Lens Model II stresses the evaluation of rater accuracy and analyzes accuracy ratings. Accuracy ratings directly represent the distances between a rater's operational ratings and criterion ratings. The RAM is intended for the use with cumulative response process, and the essays are ordered based on their difficulties for raters to score accurately. Rater location estimates show scoring accuracy based on a cumulative scale of essay ordering. An important requirement of this scale is that the essay ordering based on scoring difficulty is applied to all raters. Three research questions are proposed based on this requirement: (a) *Does an essay difficulty ordering that is consistent across all raters conform to a cumulative scale?* (b) *Do raters vary in their scoring accuracy?* and (c) *What is the difficulty of using a category relative to an adjacent category?* An essay threshold parameter is included in the RAM showing the category difficulty for raters to move from one accuracy level to an adjacent higher level.

As to unfolding response process, HCAM is appropriate for measuring rater accuracy in scoring activities. Ratets may have unique essay orderings based on their scoring accuracy toward essays. This unique ordering of essays is represented by an *IR* scale. A requirement of using an unfolding scale is that the *IR* scales of all raters can be unfolded to construct a *J* scale. The underlying scale of essays reflects their scoring difficulty with the least difficult essays located in the middle and more difficult essays unfolded into two opposite directions (Figure 7). Ratets are located near the essays that they have scored more accurately (i.e., obtained higher accuracy scores). We attend to three research questions using HCAM: (a) *Can essay difficulty orderings that vary across raters be unfolded to construct a J scale?* (b) *What is the ideal point for a rater to be the most accurate?* and (c) *What is the accuracy zone for each essay?* The zone of accuracy for each essay is

estimated with the essay threshold parameter. Raters who locate within the zone have a probability of .50 or higher of scoring this essay accurately. Unfolding models in rater-mediated assessments provide a unique ordering of essays for each rater based on this rater's accuracy toward the essays on an *IR* scale. This information can be used for the development of an adaptive and personalized rater training program.

In summary, we suggest a conceptual framework for examining and improving rating quality in rater-mediated assessments (Figure 1). Rater-mediated assessments should consider the theoretical model for rater judgment in making scoring decisions as well as a psychometric model for evaluating the ratings. Quantitative evidence obtained with a psychometric model can be used together with external validation criteria to facilitate the collection of qualitative information such as rater perceptions toward essays and interpretations of scoring rubrics. The qualitative evidence is a valuable source for improving the theoretical rater judgment models. In operational scoring settings, these four components (i.e., theoretical model for rater judgments, measurement model for ratings, quantitative and qualitative evidence) can provide feedback to enhance rater training practices and monitoring procedures. Ultimately, our goal is to improve the reliability, validity, and fairness of ratings within the context of rater-mediated assessments. Engelhard and Wind (2017) summarized specific forms of validity, reliability, and fairness evidence for the interpretation and use of rater-mediated assessments based on the general specifications of testing in the *Standards for Educational and Psychological Testing* (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 2014).

Although in this study we focused on Brunswik's original lens model, which is also called double-systems design, other types of lens model designs are available such as single-system design, triple-system design, *N*-systems design, and the hierarchical design (Dhami & Mumpower, 2018; Hammond, 1972; Hammond, Stewart, Brehmer, & Steinmann, 1975). Future research on rater judgment and the decision-making process in scoring activities can explore the potential use of different lens model designs.

The use of unfolding models in rater-mediated assessments is promising so that some measurement problems that are common in rater-mediated assessments will be considered. The first issue is the prevalent use of incomplete rating designs in operational scoring activities. Modeling missing data is possible with unfolding models. Busing and De Rooij (2009) discussed the effects of incomplete data that are missing at random and a couple of imputation techniques on the estimation performance of multidimensional unfolding models. Liu and Wang (2016) proposed an approach to handle missing not at random situations. However, the existing research has a close focus on the Likert-type items. Incomplete rating designs within rater-mediated assessments can be a unique issue. It is worth exploring in future studies based on HCM. Second, measurement invariance of both items and raters between subpopulation groups is critically important in rater-mediated assessments. The traditional item response theory-based indices and methods for detecting differential item functioning (DIF) can be applied to unfolding models (Carter, 2011; Seybert, Stark, & Chernyshenko, 2014; W. C. Wang, Tay, & Drasgow, 2013).

For rater-mediated assessments, an investigation on differential rater functioning with the use of these indices based on HCM will be considered in future research.

References

- American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. Washington, DC: American Psychological Association.
- Andrich, D. (1988). The application of an unfolding model of the PIRT type to the measurement of attitude. *Applied Psychological Measurement*, 12, 33–51.
- Andrich, D. (1996). A hyperbolic cosine latent trait model for unfolding polytomous responses: Reconciling Thurstone and Likert methodologies. *British Journal of Mathematical and Statistical Psychology*, 49, 347–365.
- Andrich, D., & Luo, G. (1993). A hyperbolic cosine latent trait model for unfolding dichotomous single-stimulus responses. *Applied Psychological Measurement*, 17, 253–276.
- Andrich, D., & Luo, G. (2017). A law of comparative preference: Distinctions between models of personal preference and impersonal judgment in pair comparison designs. *Applied Psychological Measurement*, 43(3), 181–194.
- Athanasou, J. A., & Kaufmann, E. (2015). Probability of responding: A return to the original Brunswik. *Psychological Thought*, 8(1), 7–16.
- Behizadeh, N., & Pang, M. E. (2016). Awaiting a new wave: The status of state writing assessment in the United States. *Assessing Writing*, 29, 25–41.
- Brunswik, E. (1952). *The conceptual framework of psychology*. Chicago, IL: University of Chicago Press.
- Brunswik, E. (1955). Representative design and probabilistic theory in a functional psychology. *Psychological Review*, 62(3), 193–217.
- Busing, F. M., & De Rooij, M. (2009). Unfolding incomplete data: Guidelines for unfolding row-conditional rank order data with random missings. *Journal of Classification*, 26(3), 329–360.
- Caban, H. L. (2003). Rater group bias in the speaking assessment of four L1 Japanese ESL students. *Second Language Studies*, 21(2), 1–44.
- Carrell, P. L. (1995). The effect of writers' personalities and raters' personalities on the holistic evaluation of writing. *Assessing Writing*, 2(2), 153–190.
- Carter, N. T. (2011). *Applications of differential functioning methods to the generalized graded unfolding model*. Doctoral dissertation, Bowling Green State University.
- Casabianca, J. M., Junker, B. W., & Patz, R. (2016). The hierarchical rater model. In W. J. van der Linden (Ed.), *Handbook of modern item response theory* (Vol. 1, pp. 449–465). Boca Raton, FL: Chapman & Hall/CRC.
- Cooksey, R. W. (1996). The methodology of social judgement theory. *Thinking and Reasoning*, 2(2–3), 141–174.
- Cooksey, R. W., Freebody, P., & Davidson, G. R. (1986). Teachers' predictions of children's early reading achievement: An application of social judgment theory. *American Educational Research Journal*, 23(1), 41–64.
- Coombs, C. H. (1964). *A theory of data*. New York: John Wiley and Sons, Inc.
- Crisp, V. (2012). An investigation of rater cognition in the assessment of projects. *Educational Measurement: Issues and Practice*, 31(3), 10–20.
- Cumming, A. (1990). Expertise in evaluating second language compositions. *Language Testing*, 7(1), 31–51.
- Davis, L. (2016). The influence of training and experience on rater performance in scoring spoken language. *Language Testing*, 33(1), 117–135.

- Dhami, M. K., & Mumpower, J. L. (2018). Kenneth R. Hammond's contributions to the study of judgment and decision making. *Judgment and Decision Making*, 13(1), 1–22.
- Engelhard, G., Jr. (1992). The measurement of writing ability with a many-faceted Rasch model. *Applied Measurement in Education*, 5(3), 171–191.
- Engelhard, G., Jr. (1994). Examining rater errors in the assessment of written composition with a many-faceted Rasch model. *Journal of Educational Measurement*, 31, 93–112.
- Engelhard, G., Jr. (1996). Evaluating rater accuracy in performance assessments. *Journal of Educational Measurement*, 33, 56–70.
- Engelhard, G., Jr. (2013). *Invariant measurement: Using Rasch models in the social, behavioral, and health sciences*. New York, NY: Routledge.
- Engelhard, G., Jr., Wang, J., & Wind, S. A. (2018). A tale of two models: Psychometric and cognitive perspectives on rater-mediated assessments using accuracy ratings. *Psychological Test and Assessment Modeling*, 60(1), 33–52.
- Engelhard, G., Jr., & Wind, S. (2017). *Invariant measurement with raters and rating scales: Rasch models for rater-mediated assessments*. New York, NY: Routledge.
- Gyagenda, I. S., & Engelhard, G., Jr. (2010). Rater, domain, and gender influences on the assessed quality of student writing. *Advances in Rasch Measurement*, 1, 398–429.
- Hammond, K. R. (1955). Probabilistic functioning and the clinical method. *Psychological Review*, 62(4), 255–262.
- Hammond, K. R. (1972). Inductive knowing. In J. Royce & W. Rozeboom (Eds.), *The psychology of knowing* (pp. 285–320). New York, NY: Gordon and Breach.
- Hammond, K. R. (1996). Upon reflection. *Thinking and Reasoning*, 2, 239–248.
- Hammond, K. R., Hursch, C. J., & Todd, F. J. (1964). Analyzing the components of clinical inference. *Psychological Review*, 71, 438–456.
- Hammond, K. R., Stewart, T. R., Brehmer, B., & Steinmann, D. O. (1975). Social judgment theory. In M. Kaplan & S. Schwartz (Eds.), *Human judgment and decision processes* (pp. 271–312). San Diego, CA: Academic Press.
- Hursch, C. J., Hammond, K. R., & Hursch, J. L. (1964). Some methodological considerations in multiple-cue probability studies. *Psychological Review*, 71, 42–60.
- Kaufmann, E., Reips, U. D., & Wittmann, W. W. (2013). A critical meta-analysis of lens model studies in human judgment and decision-making. *PLoS One*, 8(12), e83528.
- Lane, S. (2016). Performance assessment and accountability: Then and now. In C. Wells & M. Faulkner-Bond (Eds.), *Educational measurement: From foundations to future* (pp. 356–372). New York, NY: Guilford Press.
- Linacre, J. M. (1989). *Many-facet Rasch measurement*. Chicago, IL: MESA Press.
- Linacre, J. M. (2018). *Winsteps® Rasch measurement computer program user's guide*. Beaverton, OR: Winsteps.com.
- Liu, C. W., & Wang, W. C. (2016). Unfolding IRT models for Likert-type items with a don't know option. *Applied Psychological Measurement*, 40, 517–533.
- Lopes, L. L., & Oden, G. C. (1991). The rationality of intelligence. In E. Eells & T. Maruszewski (Eds.), *Probability and rationality: Studies on L. Jonathan Cohen's philosophy of science* (pp. 199–223). Amsterdam, The Netherlands: Rodopi.
- Lumley, T., & McNamara, T. F. (1995). Rater characteristics and rater bias: Implications for training. *Language Testing*, 12(1), 54–71.
- Luo, G. (2001). A class of probabilistic unfolding models for polytomous responses. *Journal of Mathematical Psychology*, 45(2), 224–248.
- Luo, G., & Andrich, D. (2003). *RateFOLD computer program*. Perth, Australia: Social Measurement Laboratory, School of Education, Murdoch University.
- Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika*, 47, 149–174.

- Maydeu-Olivares, A., Hernández, A., & McDonald, R. P. (2006). A multidimensional ideal point item response theory model for binary data. *Multivariate Behavioral Research, 41*, 445–472.
- McIver, J., & Carmines, E. G. (1981). *Unidimensional scaling*. Beverly Hills, CA: Sage.
- Milanovic, M., Saville, N., & Shuhong, S. (1996). A study of the decision-making behaviour of composition markers. *Performance Testing, Cognition and Assessment, 3*, 92–114.
- 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.
- Postman, L., & Tolman, E. C. (1959). Brunswik's probabilistic functionalism. In S. Koch (Ed.), *Psychology: The study of a science, Vol. 1: Sensory, perceptual, and physiological formulations*. New York, NY: McGraw-Hill.
- Raczynski, K. R., Cohen, A. S., Engelhard, G., Jr., & Lu, Z. (2015). Comparing the effectiveness of self-paced and collaborative frame-of-reference training on rater accuracy in a large-scale writing assessment. *Journal of Educational Measurement, 52*, 301–318.
- Rasch, G. (1960/1980). *Probabilistic models for some intelligence and attainment tests*. Copenhagen: Danish Institute for Educational Research. (Expanded edition, Chicago, IL: University of Chicago Press, 1980).
- Seybert, J., Stark, S., & Chernyshenko, O. S. (2014). Detecting DIF with ideal point models: A comparison of area and parameter difference methods. *Applied Psychological Measurement, 38*, 151–165.
- Thurstone, L. L. (1927a). A law of comparative judgement. *Psychological Review, 34*, 278–286.
- Thurstone, L. L. (1927b). The method of paired comparisons for social values. *Journal of Abnormal Social Psychology, 21*, 384–400.
- Thurstone, L. L. (1928). Attitudes can be measured. *American Journal of Sociology, 33*, 529–554.
- Tucker, L. R. (1964). A suggested alternative formulation in the developments by Hirsch, Hammond, and Hirsch, and by Hammond, Hirsch, and Todd. *Psychological Review, 71*, 528–530.
- Vaughn, C. (1991). Holistic assessment: What goes on in the rater mind? In L. Hamp-Lyons (Ed.), *Assessing second language writing in academic contexts* (pp. 111–125). Norwood, NJ: Ablex.
- Wang, J., & Engelhard, G., Jr. (2016). A hyperbolic cosine unfolding model for evaluating rater accuracy in writing assessments. In Q. Zhang (Ed.), *Pacific Rim Objective Measurement Symposium (PROMS) 2015 Conference Proceedings* (pp. 183–197). Berlin, Germany: Springer-Verlag.
- Wang, J., & Engelhard, G., Jr. (2017). Using a multifocal lens model and Rasch measurement theory to evaluate rating quality in writing assessments. *Pensamiento Educativo: Journal of Latin-American Educational Research, 54*(2), 1–16.
- Wang, J., & Engelhard, G., Jr. (2019). Exploring the impersonal judgments and personal preferences of raters in rater-mediated assessments with unfolding models. *Educational and Psychological Measurement*. Advance online publication. <https://doi.org/0013164419827345>
- Wang, J., Engelhard, G., Jr., Raczynski, K., Song, T., & Wolfe, E. W. (2017). Evaluating rater accuracy and perception for integrated writing assessments using a mixed-methods approach. *Assessing Writing, 33*, 36–47.
- Wang, J., Engelhard, G., Jr., & Wolfe, E. W. (2015). Evaluating rater accuracy in rater-mediated assessments using an unfolding model. *Educational and Psychological Measurement, 76*, 1005–1025.

- Wang, W. C., Su, C. M., & Qiu, X. L. (2014). Item response models for local dependence among multiple ratings. *Journal of Educational Measurement*, 51, 260–280.
- Wang, W. C., Tay, L., & Drasgow, F. (2013). Detecting differential item functioning of polytomous items for an ideal point response process. *Applied Psychological Measurement*, 37, 316–335.
- Wilson, M., & Hoskens, M. (2001). The rater bundle model. *Journal of Educational and Behavioral Statistics*, 26, 283–306.
- Wind, S. A., Wolfe, E. W., Engelhard, G., Jr., Foltz, P., & Rosenstein, M. (2018). The influence of rater effects in training sets on the psychometric quality of automated scoring for writing assessments. *International Journal of Testing*, 18(1), 27–49.
- Wolfe, E. W. (2014). *Methods for monitoring rating quality: Current practices and suggested changes*. Iowa City, IA: Pearson.
- Wolfe, E. W., Jiao, H., & Song, T. (2015). A family of rater accuracy models. *Journal of Applied Measurement*, 16(2), 153–160.
- 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.
- Wolfe, E. W., Song, T., & Jiao, H. (2016). Features of difficult-to-score essays. *Assessing Writing*, 27, 1–10.

Authors

JUE WANG is Assistant Professor in the Research, Measurement, and Evaluation Program, Department of Educational and Psychological Studies, School of Education and Human Development, University of Miami; jxw1389@miami.edu. Her primary research interests include latent variable modeling, hierarchical linear modeling, and evaluation of rater accuracy and judgment in rater-mediated assessments.

GEORGE ENGELHARD, Jr. is Professor in the Quantitative Methodology Program, Department of Educational Psychology, College of Education, University of Georgia, 325W Aderhold Hall, 110 Carlton Street, Athens, Georgia 30602; gengelh@uga.edu. His primary research interests include educational measurement and policy, Rasch measurement theory, and invariant measurement in the human sciences.