Psychological Test and Assessment Modeling, Volume 60, 2018 (1), 33-52

A tale of two models: Psychometric and cognitive perspectives on rater-mediated assessments using accuracy ratings

George Engelhard, Jr[1], Jue Wang[2], & Stefanie A. Wind[3]

Abstract

The purpose of this study is to discuss two perspectives on rater-mediated assessments: psychometric and cognitive perspectives. In order to obtain high quality ratings in rater-mediated assessments, it is essential to be guided by both perspectives. It is also important that the specific models selected are congruent and complementary across perspectives. We discuss two measurement models based on Rasch measurement theory (Rasch, 1960, 1980) to represent the psychometric perspective, and we emphasize the Rater Accuracy Model (Engelhard, 1996, 2013). We build specific judgment models to reflect the cognitive perspective of rater scoring processes based on Brunswik's Lens model framework. We focus on differential rater functioning in our illustrative analyses. Raters who possess inconsistent perceptions may provide different ratings, and this may cause various types of inaccuracy. We use a data set that consists of the ratings of 20 operational raters and three experts of 100 essays written by Grade 7 students. Student essays were scored using an analytic rating rubric for two domains: (1) idea, development, organization, and cohesion; as well as (2) language usage and convention. Explicit consideration of both psychometric and cognitive perspectives has important implications for rater training and maintaining the quality of ratings obtained from human raters.

Keywords: Rater-mediated assessments, Rasch measurement theory, Lens model, Rater judgment, Rater accuracy

---

[1]*Correspondence concerning this article should be addressed to:* George Engelhard, Jr., Ph.D., Professor of Educational Measurement and Policy, Quantitative Methodology Program, Department of Educational Psychology, 325W Aderhold Hall, The University of Georgia, Athens, Georgia 30602, U.S.A. email: gengelh@uga.edu

[2]The University of Georgia

[3]The University of Alabama

Rater-mediated performance assessments are used in many countries around the world to measure student achievement in a variety of contexts. For example, Lane (2016) has noted: "performance assessments that measure critical thinking skills are considered to be a valuable policy tool for improving instruction and student learning in the 21st century" (p. 369). Performance assessments have been used to measure proficiency in writing (Wind & Engelhard, 2013), first and second languages (Eckes, 2005; Wind & Peterson 2017), teaching (Engelhard & Myford, 2010), and student achievement in many other areas, such as music education (Wesolowski, Wind, & Engelhard, 2016).

A unique feature of performance assessments is that they require human raters to interpret the quality of a performance using a well-developed rating scale. Performance assessments can be meaningfully viewed as rater-mediated assessments because the ratings modeled in our psychometric analyses are directly obtained from human judges (Engelhard, 2002). One of the critical concerns for rater-mediated assessments is how to evaluate the quality of judgments obtained from raters. Raters may bring a variety of potential systematic biases and random errors to the judgmental tasks that may unfairly influence the assignment of ratings. As pointed out by Guilford (1936), "Raters are human and they are therefore subject to all of the errors to which humankind must plead guilty" (p. 272). However, good quality control and rater training can minimize the biases and errors.

In this study, we argue that two complementary perspectives are needed in order to evaluate the quality of rater judgments: (1) a measurement model and (2) a model of human judgment and cognition. Focusing on the role of these perspectives, we consider the following questions:

- What psychometric perspectives can be used to evaluate ratings in rater-mediated assessments?
- What cognitive perspectives can provide guidance on how to model judgments obtained in rater-mediated assessments?
- How can we connect these two theoretical perspectives to improve rater-mediated assessments?

Figure 1 provides a conceptual model representing our view of the connections between psychometric and cognitive perspectives on rater-mediated assessments. The psychometric and cognitive perspectives provide the base of a triangle that supports the development and maintenance of rater-mediated assessments. It is our view that the vertices in this triangle should be viewed together, and a major thesis of this study is that current research on raters and judgments do not go far enough in explicitly considering these connections.

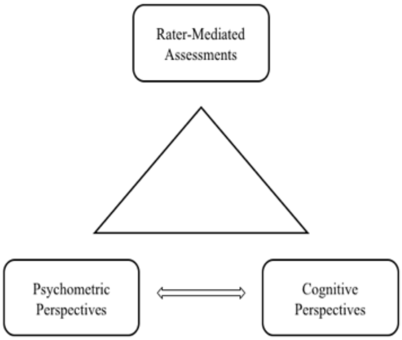

Figure 1:
Conceptual model for rater-mediated assessments

What psychometric perspectives can be used to evaluate ratings in rater-mediated assessments?

In evaluating the quality of ratings, there have been several general perspectives. These psychometric perspectives can be broadly classified into test score and scaling traditions (Engelhard, 2013). Many of the current indices used in operational testing to evaluate ratings are based on the test score tradition; for example, rater agreement indices, intraclass correlations, kappa coefficients, and generalizability coefficients (Cronbach, Gleser, Nanda, & Rajaratnam, 1972; Johnson, Penny & Gordon, 2009; von Eye, & Mun, 2005). It is safe to say that most operational performance assessment systems report the percentage of exact and adjacent category usage for operational raters. All of these models within the test score tradition treat the observed ratings as having categories with equal width. In other words, the ratings are modeled as equal intervals by using sum scores.

Ratings can also be evaluated using measurement models based on the scaling tradition (Engelhard, 2013). In the scaling tradition, the structure of rating categories is parameterized with category coefficients (i.e., thresholds). Thresholds that define rating categories are not necessarily of equal width (Engelhard & Wind, 2013). The most common IRT models for rating scale analysis include the Partial Credit Model (Masters, 1982), the Rating Scale Model (Andrich, 1978), the Generalized Partial Credit Model (Muraki, 1992), and the Graded Response Model (Samejima, 1969). The Many-Facet Rasch model (Linacre, 1989) specifically adds a rater parameter, and this model is widely used in the detection of rater effects. The Many-Facet Rasch model is a generalized form of the Rasch model that was specifically designed for rater-mediated assessments (Eckes, 2015). There are also several other rater models, such as the hierarchical rater model (Casabiaca, Junker, & Patz, 2016), that have been proposed. It is beyond the scope of this study to describe in detail other models for ratings, and we recommend Nering and Ostini (2010) for interested readers.

All of the psychometric perspectives described up to this point model the observed ratings assigned by raters. Engelhard (1996) proposed another approach based on accuracy ratings (Wolfe, Jiao, & Song, 2014). Accuracy ratings represent the distances between criterion ratings and operational ratings. For instance, criterion ratings are assigned by an expert rater or a group of expert raters. The observed ratings assigned by well-trained operational raters are referred to as operational ratings. The differences between these operational ratings and criterion ratings reflect the accuracy of operational rater judgments on each performance. Engelhard (1996) put forward an equation for calculating accuracy ratings. Since accuracy ratings reflect the distance between operational ratings and criterion ratings, we call them *direct measures* of rater accuracy. On the other hand, observed operational ratings are viewed as *indirect measures* for rater accuracy. Due to this difference, we use the term *Rater Accuracy Model*s (RAM) to label the Rasch models that examine accuracy ratings as the dependent variable on which individual raters, performances, and other facets can be measured. We present two lens models for observed operational ratings and accuracy ratings correspondingly.

Scholars have used the term *rater accuracy* in numerous ways to describe a variety of rating characteristics, including agreement, reliability, and model-data fit (Wolfe & McVay, 2012). In these applications, rater accuracy is used as a synonym for ratings with desirable psychometric properties. RAM provides a criterion-referenced perspective on rating quality that can be used to directly describe and compare individual raters, performances, and other facets in the assessment system with a focus on rater accuracy. From criterion-referenced perspective, the RAM provides a more specific definition and clear interpretation of rater accuracy. Furthermore, the criterion-referenced approach emphasizes the evaluation of rater accuracy using accuracy ratings as direct measures. These accuracy ratings can be coupled with a lens model to guide rater training and diagnostic activities during scoring.

We summarize five sources of inaccuracy due to differences among rater judgments in Table 1. First, we view *rater inaccuracy* as a tendency to consistently provide biased ratings. Second, *halo inaccuracy* or domain inaccuracy refers to the situations that raters fail to distinguish among different domains on an analytic scoring rubric when evaluating student performances. Wang, Engelhard, Raczynski, Song, and Wolfe (2017) observed this phenomenon that some raters tended to provide adjacent scores for two distinct domains of writing. Third, when raters use the rating scale in an idiosyncratic fashion, it leads to *response set inaccuracy* such that ratings are not consistent toward the *benchmarks* used as the basis for criterion ratings. Specifically, person benchmarks refer to the pre-calibrated performances (e.g., students' essays) that are used to evaluate raters' scoring proficiency. Fourth, *score range inaccuracy* occurs when ratings have less or more variation than expected based on the measurement model. Lastly, if raters interpret other facets differentially, *interaction effects* may appear in rater inaccuracy. It should also be noted that the focus (i.e., individual raters versus rater groups) yields different questions and conclusions related to rater inaccuracy. These sources of rater inaccuracy can guide researchers in identifying possible sources of rater inaccuracy with the use of RAM or other psychometric models.

Table 1:

Sources of Rater Inaccuracy

Definitions	Focus	
	Individual Raters	Rater Group
1. Rater Inaccuracy: The tendency on the part of raters to consistently provide higher or lower ratings than warranted based on known person benchmarks.	How accurate is each rater? Where is the rater located on the Wright map for accuracy?	Are the differences in rater accuracy significant? Can the raters be considered of equivalent accuracy?
2. Halo inaccuracy (domain inaccuracy): Rater fails to distinguish between conceptually distinct and independent domains on person benchmarks.	Is the rater distinguishing between conceptually distinct domains?	Are the raters distinguishing among the domains?
3. Response set inaccuracy: Rater interprets and uses rating scale categories in an idiosyncratic fashion.	Is the rater using the rating scale as intended?	Are the raters using the rating scales as intended?
4. Score Range Inaccuracy: More or less variation in accuracy ratings of benchmarks. Raters do not differentiate between person benchmarks on the latent variable.	How well did each rater differentiate among the benchmarks?	Did the assessment system lead to the identification of meaningful differences between the benchmarks?
5. Inaccuracy interaction effects: Facets in the measurement model are not interpreted additively.	Is the rater interpreting and using the facets accurately?	Are the facets invariant across raters?

Note. Person benchmarks represent the criterion performances (e.g., essays with ratings assigned by experts) used to evaluate rater accuracy.

There have been several recent applications of the RAM that reflect different measurement frameworks and contexts. For example, Engelhard (1996) adapted the Rasch model for examining rater accuracy in a writing assessment. Wesolowski and Wind (in press) as well as Bergin, Wind, Grajeda, and Tsai (2017) used the distance between operational and expert ratings as the dependent variable in a Many-Facet Rasch model to evaluate rater accuracy in music assessments and teacher evaluations, respectively. Another example is Patterson, Wind, and Engelhard (2017) who incorporated criterion ratings into signal detection theory for evaluating rating quality. Finally, Wang, Engelhard, and Wolfe (2016) have used accuracy ratings with an unfolding model to examine rater accuracy.

What cognitive perspectives can provide guidance on how to model judgments obtained in rater-mediated assessments?

The simple beauty of Brunswik's lens model lies in recognizing that the person's judgment and the criterion being predicted can be thought of as two separate functions of cues available in the environment of the decision.

(Karelaia and Hogarth, 2008, p. 404)

Cognitive psychology (Barsalou, 1992) offers a variety of options for considering judgment and decision-making tasks related to rater-mediated assessments. Cooksey (1996) describes 14 theoretical perspectives on judgment and decision making that can be potential models for examining the quality of judgments in rater-mediated assessments. Within educational settings, there was a special issue of *Educational Measurement: Issues and Practice* devoted to rater cognition (Leighton, 2012). There are many promising areas for future research on rater cognition and rater judgments (Lane, 2016; Myford, 2012; Wolfe, 2014).

Although there are numerous potential models of human judgment that may be useful guides for monitoring rating quality, the underlying model of judgmental processes used here is based on Brunswik's (1952) lens model. Lens models have been used extensively used across social science research contexts to examine human judgments. For example, there are two important meta-analyses of research organized around lens models. First, Karelaia and Hogarth (2008) conducted a meta-analysis of five decades of lens model studies (N=249) that included a variety of task environments. More recently, Kaufmann, Reips, and Wittmann (2013) conducted a meta-analysis based on 31 lens model studies including applications from medicine, business, education, and psychology. An important resource for recent work on lens models is the website of the Brunswik Society (http://www.brunswik.org/), which provides yearly abstracts of current research utilizing a lens model framework.

Brunswik (1952, 1955a, 1955b, 1956) proposed a new perspective in psychology called probabilistic functionalism (Athanasou & Kaufmann, 2015; Postman & Tolman, 1959). An important aspect of Brunswik's research was the concept of a lens model (Hammond, 1955; Postman & Tolman, 1959). The structure of Brunswik's lens models varied over time and application areas. Figure 2 presents a lens model for perception proposed by Brunswik (1955a). In this case, a person utilizes a set of cues (i.e., proximal-peripheral cues) to generate a response (i.e., central response). The accuracy of a person's response can be evaluated by its relationship to the distal variable, which is called functional validity. Ecological validities represent the relationships between the distal variable and the cues, while utilization validities reflect the relationship between the cues and the central response. In both cases, higher values of correspondence are viewed as evidence of validity. It is labeled a *lens model* because it resembles the way light passes through a lens defined by cues.

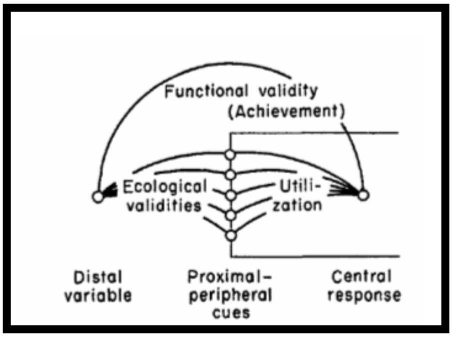

Figure 2:

Lens model for perception constancy (Adopted from Brunswik (1955a, p. 206)

In rater-mediated assessments, the accuracy of a rater's response (i.e., observed rating) is evaluated by its correspondence to or relationship with the latent variable (i.e., distal variable). Engelhard (1992, 1994, 2013) adapted the lens model as a conceptual framework for rater judgments in writing assessment. Figure 3 provides a bifocal perspective on rater accuracy in measuring writing competence. We refer to Figure 3 as *Lens Model I,* where the basic idea is that the latent variable — writing competence — is made visible through a set of cues or intervening variables (e.g., essay features, domains, and rating scale usages) that are interpreted separately by experts and operational raters. Our goal in this case is to have a close correspondence between the measurement of the latent variable (i.e., writing competence) between expert and operational raters. Judgmental accuracy in Lens Model I refers to the closeness between rater's operational ratings and experts' criterion ratings of student performances including their interpretations of the cues. Wang and Engelhard (2017) applied Lens model I to evaluate rating quality in writing assessments.

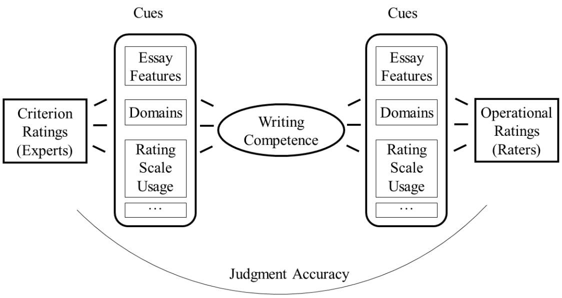

Figure 3:

Lens model I (bifocal model) for measuring writing competence

In contrast to Lens Model I, the current study focuses on a slightly different definition of a lens model. Specifically, we focus on *Lens Model II* (see Figure 4). In Lens Model II, the latent variable is rater accuracy instead of writing competence in the assessment system. The goal is to evaluate accuracy ratings (i.e., differences between observed and criterion ratings) as responses of raters in the judgmental system. These accuracy ratings can be distinguished from the ratings modeled separately for expert and operational raters in Lens Model I.

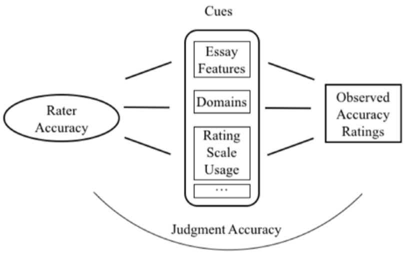

Figure 4:

Lens model II for measuring rater accuracy

As pointed out in the opening quote for this section, a defining feature of lens models is that they include two separate functions reflecting judgment and criterion systems. Brunswik (1952) primarily used correlational analyses to examine judgmental data. Multiple regression analyses are currently the most widely used method for examining data from lens-model studies of judgments (Cooksey, 1996). It is interesting to note that Hammond (1996) suggested that lens-model research may have overemphasized the role of multiple regression techniques, and that the "lens model is indifferent — a priori — to which organizing principle is employed in which task under which circumstances; it considers that to be an empirical matter" (p. 245). In our study, we suggest using psychometric models based on Rasch measurement theory and invariant measurement as an organizing principle (Engelhard, 2013). As pointed out earlier, the majority of analyses conducted with lens models are regression-based analyses. Lens Model I reflects this perspective very closely with the Rasch model substituted for multiple regression analyses.

How can we connect these two perspectives to improve rater-mediated assessments?

Accuracy …refers to closeness of an observation to the quality intended to be observed

(Kendall & Buckland, 1957, p. 224)

Researchers have adopted several different statistical approaches for analyzing data for lens-model studies. First, the ratings have been modeled directly using correlational and multiple regression analyses (Brunswik 1952; Cooksey, 1996; Hammond, Hursch, and Todd, 1964; Hursch, Hammond, & Hursch, 1964; Tucker, 1964). Cooksey (1986) provided an informative example of using a lens model approach to examine teacher judgments of student reading achievement. In this study, student scores on standardized reading achievement tests define the ecological or criterion system with three cues (i.e., social economic status, reading ability, and oral language ability). In a similar fashion, the judgmental system was defined based on the relationship between teacher judgments and the same set of cues. Regression-based indices were used to compare the ecological and judgmental systems. Cooksey, Freebody, and Wyatt-Smith (2007) also applied a lens model to study teacher's judgments of writing achievement. The drawback of this methodology is that each person's judgment is compared against the criterion individually; that said, separate regression analyses are required for each judge.

A second approach is to use IRT models that are developed within the scaling tradition. Researchers can obtain individual-level estimates using various IRT models in one analysis instead of separate multiple-regression analyses. For example, Engelhard (2013) proposed the use of a Many-Facet Rasch Model to examine the lens model I for measuring writing proficiency.

Finally, it is possible to model the criterion and judgmental systems as the distances between the ratings from each system. The lens model for measuring rater accuracy based on this approach can be best represented by the RAM. RAM has been proposed and applied to evaluate rater accuracy in writing assessments (Engelhard, 1996, 2013; Wolfe, Jiao, & Song, 2014). We illustrate the correspondence between the Lens Model II and the RAM. Specifically, we use the distances between the ratings of expert raters and the operational raters to define accuracy ratings which are analyzed in the judgment system of Lens Model II. RAM analyzes the accuracy ratings that are direct measures of rater accuracy.

In addition, there are several advantages of using Rasch measurement theory over regression-based approaches for judgment studies. First of all, multiple regression analyses may lead to a piecemeal approach with an array of separate analyses. Cooksey (1996) provides ample illustrations of these types of analyses within the context of judgment studies. Our approach based on Rasch measurement theory provides a coherent view for analyzing rater-mediated assessments. Second, it is hard to substantively conceptualize the focal point (i.e., object of measurement) when a regression-based approach is used. In this study, we describe two Rasch-based approaches that focus on either students or raters as the object of measurement. Our approach offers the advantages of obtaining invariant indicators of rating quality under appropriate conditions. Lastly, we would like to stress the value of Wright Maps that define an underlying continuum, and provide the opportunity to visualize and understand rater-mediated measurement as a line representing the construct or latent variable of interest.

Illustrative data analyses

In this study, we use illustrative data analyses to highlight the use of the RAM and Brunswikian lens model as a promising way to bring together psychometric and cognitive perspectives related to evaluating rater judgments. Specifically, we conducted a secondary data analysis with the use of RAM to examine differential rater functioning as one of the sources causing inaccurate ratings through the lens. The data, which were originally collected and analyzed by Wang, Engelhard, Raczynski, Song, and Wolfe (2017), were part of a statewide writing assessment program for Grade 7 students in a southeastern state of the United States.

Participants

According to Wang et al. (2017)'s data collection procedure, twenty well-trained operational raters were randomly chosen from a larger rater pool. The group of raters scored a random sample of 100 essays. This set of essays was used as training essays to evaluate rater performance prior to the actual operational scoring. The design was fully crossed with all of the raters rating all of the essays. A panel of three experts who provided the training and picked the training essays assigned the criterion ratings for these 100 essays.

Instrument

The writing assessment was document based, that is students were asked to write an essay based on a prompt. The essays were scored analytically in two domains: (a) idea development, organization, and coherence (IDOC Domain), and (b) language usage and conventions (LUC Domain). IDOC Domain was scored using a category of 0-4, and LUC domain was rated from 0-3. A higher score indicates better proficiency in a specific writing domain.

Procedures

In our study, exact matches between operational and criterion ratings from the panel of expert raters are assigned an accuracy rating of 1, while other discrepancies are assigned a 0. Higher scores reflect higher levels scoring accuracy for raters. In other words, accuracy ratings are dichotomized (0=inaccurate rating, 1=accurate ratings).

The RAM includes three facets: Raters, essays and domains. We used the Facets computer program (Linacre, 2015) to analyze the dichotomous accuracy ratings. The general RAM model can be expressed as follows:

$$\mathrm{Ln}[P_{nmik} / P_{nmik-1}] = \beta_n - \delta_m - \lambda_i - \tau_k \tag{1}$$

where

P_{nmik} = probability of rater n assigning an accurate rating to benchmark essay m for domain i,

P_{nmik-1} = probability of rater n assigning an inaccurate rating to benchmark essay m for domain i,

β_n = accuracy of rater n,

δ_m = difficulty of assigning an accurate rating to benchmark essay m,

λ_i = difficulty of assigning an accurate rating for domain i, and

τ_k = difficulty of accuracy-rating category k relative to category k-1.

Next, we examine an interaction effect between rater accuracy measures and domain facet using the model as below:

$$\mathrm{Ln}[P_{\mathrm{nmik}} / P_{\mathrm{nmik-1}}] = \beta_n - \delta_m - \lambda_i - \beta_n\lambda_i - \tau_k \qquad (2)$$

where $\beta_n\lambda_i$ represents the interaction effect between rater and domains.

The τ_k parameter is not estimated in this study because the accuracy ratings are dichotomous. However, we included it here because it is possible to apply this model to polytomous accuracy ratings, in which case the threshold parameter would be included.

Results

Summary statistics for the calibrated facets are shown in Table 2. The Wright Map is shown in Figure 5. The reliability of separation for rater accuracy is .47, and the Chi-square test for variation among raters is statistically significant ($\chi^2 = 35.6$, $df = 19$, $p < .05$). Table 3 shows the detailed analyses of accuracy for each rater. The mean accuracy measure for raters is .63 logits with a standard deviation for .22. Rater 2702 is the most accurate rater with a measure of 1.02 logits, and Rater 2696 is the least accurate rater with an accuracy measure of .55 logits. Based on the standardized Outfit and Infit values, Rater 2569 appears to be exhibiting misfit.

Table 2:

Summary statistics for Rater Accuracy Model

	Rater	Essays	Domains
Measure			
Mean	.63	.00	.00
SD	.22	.76	.62
N	20	100	2
Infit MSE			
Mean	1.00	1.00	1.00
SD	.06	.15	.00
Outfit MSE			
Mean	1.00	1.00	1.00
SD	.10	.20	.01
Separation statistics			
Reliability of separation	.47	.77	.99
Chi-square (χ^2)	35.6*	348.4*	154.1*
df	19	99	1

Note. MSE = mean square error, * $p < .05$.

Table 3:

Accuracy measures and fit statistics for raters

Rater ID	Accuracy (Prop.)	Measure (Logits)	S.E.	Infit MSE	Infit Z	Outfit MSE	Outfit Z	Slope
2702	0.70	1.02	0.17	1.07	1.01	1.10	0.86	0.84
2744	0.69	0.91	0.16	0.98	-0.21	0.92	-0.76	1.07
3051	0.67	0.83	0.16	1.05	0.78	1.16	1.53	0.84
3271	0.67	0.83	0.16	1.07	1.10	1.08	0.76	0.82
1714	0.66	0.81	0.16	0.95	-0.82	0.92	-0.79	1.14
2505	0.65	0.73	0.16	0.99	-0.19	0.97	-0.26	1.04
3076	0.65	0.73	0.16	0.99	-0.13	0.98	-0.16	1.03
3083	0.65	0.76	0.16	1.03	0.42	1.06	0.66	0.91
3372	0.65	0.73	0.16	1.04	0.59	1.00	-0.01	0.93
698	0.64	0.70	0.16	0.90	-1.76	0.84	-1.79	1.31
3153	0.64	0.70	0.16	0.91	-1.49	0.86	-1.52	1.26
2911	0.63	0.63	0.16	0.93	-1.26	0.89	-1.20	1.23
2423	0.61	0.53	0.16	0.99	-0.15	1.04	0.54	1.00
3084	0.60	0.48	0.16	0.97	-0.57	0.93	-0.87	1.13
2020	0.59	0.44	0.15	0.96	-0.81	0.93	-0.82	1.16
2905	0.58	0.41	0.15	0.98	-0.39	0.95	-0.59	1.09
730	0.57	0.36	0.15	1.08	1.53	1.10	1.24	0.70
2481	0.57	0.36	0.15	1.02	0.37	1.03	0.43	0.92
2569	0.57	0.34	0.15	1.13	2.39*	1.23	2.85*	0.48
2696	0.55	0.25	0.15	0.98	-0.44	0.96	-0.51	1.09

Note. Accuracy is the proportion of accurate ratings. Raters are ordered based on measures (logits). SE = standard error, MSE=mean square error, and * $p < .05$.

As shown in Table 2, the benchmark essays are centered at zero with a standard deviation of .76. Overall, the benchmark essay accuracy measures have relatively good fit to the model. Measures for domain accuracy are also centered at zero. Domain IDOC has a measure of .44 logits and Domain LUC has a measure of -.44 logits (Table 4). IDOC seems to be more difficult for raters to score accurately than LUC. The reliability of separation is .99, and the differences among the domain locations on the logit scale are statistically significant ($\chi^2 = 154.1$, $df = 1$, $p < .05$)

Table 4:

Summary statistics for Rater Accuracy Model by Domain

Domains	Accuracy	Measure	SE	Infit MSE	Infit Z	Outfit MSE	Outfit Z	Slope
IDOC	0.54	0.44	0.05	1.00	-0.04	1.00	0.07	1.00
LUC	0.72	-0.44	0.05	1.00	0.02	0.99	-0.15	1.00

Note. IDOC = idea, development, organization, and cohesion, LUC = language usage and convention, SE = standard error, and MSE = mean square error.

We also included an interaction term (i.e., domain by rater facets) in the model. We used *t*-tests to compare the differences of accuracy measures between domains for each rater. Results indicate that three raters have significantly different accuracy measures between the two domains (Table 5). Specifically, Raters 3271 and 2905 appear to be significantly more accurate in scoring Domain IDOC than Domain LUC. On the contrary, Rater 3084 seems to be significantly more accurate in Domain LUC than Domain IDOC.

In order to interpret these results in terms of their substantive implications, it is informative to relate these results to the five aspects of inaccuracy described in Table 1. Specifically, *rater inaccuracy* is the tendency on the part of raters to consistently provide higher or lower ratings overall. The illustrative data in this study suggest that the individual raters vary in their levels of inaccuracy. The Wright Map (Figure 5) provides a visual display of where each rater is located on the accuracy continuum. The raters are not equivalent in terms of accuracy rates. The data also provide evidence of domain variation in inaccuracy (halo inaccuracy). Some raters appear to vary in their accuracy rates as a function of domain. Overall, there were differences in rater accuracy between the two domains, where the IDOC domain was more difficult for raters to score accurately as compared to the LUC domain.

Next, *response set inaccuracy* implies that a rater interprets and uses rating scale categories in an idiosyncratic fashion. Because the accuracy data in this study are dichotomous, this issue is moot. Third, *score range inaccuracy* is observed in these data with the benchmark essays varying in difficulty to rate accurately as shown on the Wright Map (Figure 5). Further research is needed on why certain essays appear to be more accurately rated than other essays. Finally, there was evidence of an *inaccuracy interaction effect* between raters and domains. This result suggests that rater effects are not additive, and that the domain facet is not invariant across raters. In other words, the relative ordering of the domains in terms of the difficulty to assign accurate ratings was not the same for all of the raters.

Table 5:
Analysis of differential rater functioning across domains

	IDOC Domain		LUC Domain				
Rater	**Measure**	**SE**	**Measure**	**SE**	**Contrast**	**t-value**	**Prob**
3271	1.15	0.22	0.47	0.23	0.68	2.14*	0.03
2905	0.72	0.21	0.08	0.22	0.65	2.13*	0.03
2744	1.15	0.22	0.63	0.23	0.52	1.62	0.11
2423	0.68	0.21	0.37	0.22	0.31	1.01	0.31
2702	1.10	0.22	0.91	0.25	0.18	0.56	0.58
3051	0.91	0.22	0.74	0.24	0.17	0.53	0.59
3083	0.82	0.21	0.68	0.24	0.14	0.42	0.67
730	0.41	0.21	0.32	0.22	0.09	0.30	0.76
2696	0.27	0.21	0.22	0.22	0.05	0.18	0.86
2569	0.36	0.21	0.32	0.22	0.05	0.15	0.88
2481	0.36	0.21	0.37	0.22	0.00	-0.01	0.99
2505	0.72	0.21	0.74	0.24	-0.01	-0.04	0.97
3076	0.72	0.21	0.74	0.24	-0.01	-0.04	0.97
698	0.63	0.21	0.79	0.24	-0.16	-0.50	0.62
1714	0.72	0.21	0.91	0.25	-0.19	-0.58	0.56
3372	0.63	0.21	0.85	0.24	-0.22	-0.68	0.50
2911	0.45	0.21	0.85	0.24	-0.40	-1.23	0.22
3153	0.50	0.21	0.98	0.25	-0.48	-1.45	0.15
2020	0.18	0.21	0.74	0.24	-0.56	-1.73	0.08
3084	0.09	0.22	0.98	0.25	-0.89	-2.68*	0.01

Note. IDOC = ideas, development, organization, and cohesion, LUC = language usage and conventions, and SE = standard errors, * $p < .05$.

```
+---------------------------------------------------------+
|Logit|+Rater                     |-Essay      |-Domains|
|-----+---------------------------+------------+--------|
              High Accuracy           Hard         Hard
|   2 +                           +            +        |
|     |                           |            |        |
|     |                           |            |        |
|     |                           |            |        |
|     |                           |            |        |
|     |                           | *          |        |
|     |                           |            |        |
|     |                           | *          |        |
|     |                           | *****      |        |
|     |                           |            |        |
|   1 + 2702                      +            +        |
|     | 2744                      | ****       |        |
|     | 1714 3051 3083 3271       | *****      |        |
|     | 698  2505 3076 3153 3372  | ******     |        |
|     | 2911                      | **         |        |
|     | 2423 3084                 | *****      |        |
|     | 730  2020 2481 2905       | ****       | IDOC   |
|     | 2569                      | ***        |        |
|     | 2696                      | *********  |        |
|     |                           | ********   |        |
*   0 *                           * ******     *        *
|     |                           | **         |        |
|     |                           |            |        |
|     |                           | ********   |        |
|     |                           | ******     | LUC    |
|     |                           | ***        |        |
|     |                           |            |        |
|     |                           | *******    |        |
|     |                           | ******     |        |
|     |                           |            |        |
|  -1 +                           + ***        +        |
|     |                           |            |        |
|     |                           |            |        |
|     |                           |            |        |
|     |                           | **         |        |
|     |                           |            |        |
|     |                           |            |        |
|     |                           | *          |        |
|     |                           |            |        |
|     |                           |            |        |
|  -2 +                           + ***        +        |
             Low Accuracy             Easy         Easy
|-----+---------------------------+------------+--------|
```

Note. IDOC = ideas, development, organization, and cohesion, LUC = language usage and conventions.

Figure 5:

Wright Map for Rater Accuracy Model

Discussion

In this study, we briefly discussed two perspectives on evaluating the quality of ratings in rater-mediated assessments: a psychometric perspective and a cognitive perspective. As shown in Figure 1, rater-mediated assessments rely on both perspectives to have reliable, valid, and fair ratings in a rater-mediated assessment system of performances. Much of the current research on rating quality has been dominated by a psychometric perspective with relatively little research on the cognitive processes of human raters. In order to meaningfully evaluate and interpret the quality of ratings, it is important to explicitly consider both theory of measurement and theory of rater cognition. Ideally, these two perspectives should be complementary and congruent. The psychometric perspective used in this study is based on Rasch measurement theory, and the cognitive perspective is based on Brunswik's lens model. In particular, we emphasized the use of a rater accuracy model (RAM) to illustrate our major points.

Our study was guided by the following three questions:

- What psychometric perspectives can be used to evaluate ratings in rater-mediated assessments?
- What cognitive perspectives can provide guidance on how to model judgments obtained in rater-mediated assessments?
- How can we connect these two theoretical perspectives to improve rater-mediated assessments?

In answer to the first question, we believe that a scaling perspective based on item response theory in general and Rasch measurement theory in particular provides the best match to the models of judgment in rater-mediated assessments. Rasch measurement theory specifies the requirements necessary for developing and maintaining a psychometrically sound performance assessment system. There are two versions of the Rasch model that can be used to evaluate rater accuracy. A Rasch model with observed ratings and a Rasch model with accuracy ratings which is called Rater Accuracy Model. The first model focuses on two assessment systems (one based on expert raters and the second on operational raters) with the latent variable defining the object of measurement for both groups of raters. The second model (i.e., RAM) focuses on rater accuracy directly as the latent variable with the raters defined as the objects of measurement. RAM offers a direct evaluation of rater accuracy measures with accuracy ratings which are defined as the differences between observed and criterion ratings.

Turning now to the second question, we selected cognitive perspectives based on Brunswik's Lens Model as the basis for examining human judgments in rater-mediated assessments. Lens models connect the criterion system and the judgmental system which can best represent operational raters' cognition processes while making judgments. We have described two lens models. *Lens Model I* is for measuring student proficiency (e.g., writing competency) as the distal variable (Figure 3). *Lens Model II* is for measuring rater accuracy directly as the distal variable (Figure 4), which emphasizes the evaluation of the raters or judges by modeling the distances between operational ratings and criterion ratings.

The final question raises an important issue about the congruence between a statistical theory of measurement and a substantive theory regarding human cognition and judgment. Lens models can be conceptually linked to both the Many-Facet Rasch Model and the RAM with the major distinctions between the objects of measurement in two models. For both models, it is substantively useful to visualize the locations of the object of measurement on a Wright Map, to define the latent variable in terms of the specific cues used by the raters as *lens*, and to conceptualize two systems -- criterion system and judgmental system. The Many-Facet Rasch Model analyzes the two systems separately and then compares the results. The measurement focuses on student proficiency as a latent continuum in each system, and the consistency between two systems reflects the rater accuracy. On the other hand, the RAM is used to model accuracy ratings defined as the distances between the two systems. This approach directly reflects rater accuracy by modeling it as the underlying latent trait.

Using illustrative data from a rater-mediated writing performance assessment, we demonstrated the statistical procedures for modeling rater accuracy. Specifically, we calculated accuracy ratings by matching operational ratings and the criterion ratings for individual raters. Then we used the RAM to analyze accuracy ratings to obtain the accuracy measures for individual raters, the difficulty associated with scoring accuracy for student performances (i.e., essays), and the difficulty associated with scoring accuracy for the domains that were specified in the analytic scoring rubric. To evaluate differential rater functioning, we examined the interaction between individual raters and domains. Lastly, we interpreted the statistical results of RAM based on the five potential sources of inaccuracy. These sources of inaccuracy also provide a frame of reference for interpreting the statistical results in terms of specific rater issues in operational performance assessments.

We want to stress that the statistical theories of measurement and substantive theories of human cognition and judgment for evaluating rating quality should be complementary and congruent. Ideally, research on rater-mediated assessments should balance concerns with both cognitive and psychometric perspectives. In practice, the development and evaluation of how well our theories match one another remains a challenging puzzle. As progress is made in both areas, the nexus between psychometrics and cognition for rater-mediated assessments promises to be an exciting area of research.

Finally, the title of this study reflects an indirect reference to the opening lines in *A Tale of Two Cities* (Charles Dickens, 1859):

> *It was the best of times, it was the worst of times, it was the age of wisdom, it was the age of foolishness, it was the epoch of belief, it was the epoch of incredulity, it was the season of Light, it was the season of Darkness, it was the spring of hope, it was the winter of despair…*

Some researchers who evaluate rater-mediated assessments have numerous justifiable concerns about human biases and errors (e.g., intentional and random), and their perspectives may reflect despair over the current state of the art. From our perspective, we have hope that many of the concerns about human scoring can be minimized and the promise of performance assessments become a reality in education and other contexts. In particular, we believe that explicit considerations of both psychometric and cognitive perspectives have important implications for improving the training and maintaining the quality of ratings obtained from human raters in performance assessments.

References

Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika, 43*, 561-73.

Athanasou, J.A., & Kaufmann, E. (2015). Probability of responding: A return to the original Brunswik. *Psychological Thought, 8*(1), 7–16.

Barsalou, L. W. (1992). *Cognitive psychology: An overview for cognitive scientists*. Psychology Press.

Bergin, C., Wind, S. A., Grajeda, S., & Tsai, C.-L. (2017). Teacher evaluation: Are principals' classroom observations accurate at the conclusion of training? *Studies in Educational Evaluation*, *55*, 19–26.

Brunswik, E. (1952). *The conceptual framework of psychology*. Chicago, IL: University of Chicago Press.

Brunswik, E. (1955a). Representative design and probabilistic theory in a functional psychology. *Psychological Review, 62*(3), 193-217.

Brunswik, E. (1955b). In defense of probabilistic functionalism: A reply. *Psychological Review, 62*(3), 236-242.

Brunswik, E. (1956). *Perception and the representative design of psychological experiments* (2nd ed.). Berkeley, CA: University of California Press.

Casabianca, J. M., Junker, B. W., & Patz, R. J. (2016). Hierarchical rater models. In W. J. van der Linden (Ed.), *Handbook of item response theory* (Vol. 1, pp. 449–465). Boca Raton, FL: Chapman & Hall/CRC.

Cooksey, R. W. (1996). *Judgment analysis: Theory, methods, and applications*. United Kingdom: Emerald Group Publishing Limited.

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.

Cooksey, R. W., Freebody, P., & Wyatt-Smith, C. (2007). Assessment as judgment-in-context: Analysing how teachers evaluate students' writing. *Educational Research and Evaluation, 13(5),* 401-434.

Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. (1972). *The dependability of behavioral measurements: Theory of generalizability for scores and profiles.* New York: Wiley.

Dickens, C. J. H. (1859). *A tale of two cities (Vol. 1).* Chapman and Hall.

Eckes, T. (2005). Examining rater effects in TestDaF writing and speaking performance assessments: A many-facet Rasch analysis. *Language Assessment Quarterly*, *2*(3), 197–221.

Eckes, T. (2015). *Introduction to many-facet Rasch measurement: Analyzing and evaluating rater mediated assessments* (2nd ed.). Frankfurt am Main: Peter Lang.

Engelhard Jr, G. (1992). The measurement of writing ability with a many-faceted Rasch model. *Applied Measurement in Education, 5(3),* 171-191.

Engelhard, G. (1994). Examining rater errors in the assessment of written composition with a many-faceted Rasch model. *Journal of Educational Measurement, 31(2),* 93-112.

Engelhard, G. (1996). Evaluating rater accuracy in performance assessments. *Journal of Educational Measurement*, *33*(1), 56-70.

Engelhard, G. (2002). Monitoring raters in performance assessments. In G. Tindal and T. Haladyna (Eds.), *Large-scale assessment programs for ALL students: Development, implementation, and analysis,* (pp. 261-287). Mahwah, NJ: Erlbaum.

Engelhard, G. (2013). *Invariant measurement: Using Rasch models in the social, behavioral, and health sciences*. New York: Routledge.

Engelhard, G., & Myford, C. (2010). Comparison of single and double assessor scoring designs for the assessment of accomplished teaching. In Garner, M., Engelhard, G., Wilson, M., & Fisher, W. (Eds.). *Advances in Rasch measurement* (Vol. 1, pp. 342-368). Maple Grove, MN: JAM Press.

Engelhard, G., & Wind, S.A. (2013). *Rating quality studies using Rasch measurement theory.* College Board Research Report 2013-3.

Guilford, J. P. (1936). *Psychometric methods.* New York: McGraw Hill.

Hammond, K. R. (1955). Probabilistic functioning and the clinical method. *Psychological review, 62(4),* 255.

Hammond, K. R. (1996). Upon reflection. *Thinking & Reasoning*, *2*(2-3), 239-248.

Hammond, K. R., Hursch, C. J., & Todd, F. J. (1964). Analyzing the components of clinical inference. *Psychological review, 71(6),* 438.

Hursch, C. J., Hammond, K. R., & Hursch, J. L. (1964). Some methodological considerations in multiple-cue probability studies. *Psychological review, 71(1),* 42.

Johnson, R. L., Penny, J. A., & Gordon, B. (2009). *Assessing performance: Designing, scoring, and validating performance tasks*. New York: Guilford Press.

Karelaia, N., & Hogarth, R.M. (2008). Determinants of linear judgment: A meta-analysis of lens model studies. *Psychological Bulletin, 134*(3), 404–426.

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.

Kendall, M. G., & Buckland, W. R. (1957). *Dictionary of statistical terms*. Edinburgh, Scotland: Oliver and Boyd.

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: Guilford.

Leighton, J. P. (2012). Editorial. *Educational Measurement: Issues & Practice*, *31*(3), 48-49.

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

Linacre, J. M. (2015) *Facets computer program for many-facet Rasch measurement, version 3.71.4.* Beaverton, Oregon: Winsteps.com

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika*, *47*(2), 149-174.

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

Myford, C. M. (2012). Rater cognition research: Some possible directions for the future. *Educational Measurement: Issues & Practice*, *31*(3), 48-49.

Nering, M.L., & Ostini, R. (2010). *Handbook of polytomous item response theory models.* New York: Routledge.

Patterson, B.F., Wind, S.A., & Engelhard, G. (2017). Incorporating criterion ratings into model-based rater monitoring procedures using latent class signal detection theory. *Applied Psychological Measurement*, 1-20.

Postman, L., & Tolman, E. C. (1959). Brunswik's probabilistic functionalism. *Psychology: A study of a science*, *1*, 502-564.

Rasch (1960/1980). *Probabilistic models for some intelligence and attainment tests.* Copenhagen: Danish Institute for Educational Research. (Expanded edition, Chicago: University of Chicago Press, 1980).

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

Tucker, L. R. (1964). A suggested alternative formulation in the developments by Hursch, Hammond, and Hursch, and by Hammond, Hursch, and Todd. *Psychological Review, 71*(6), 528-530.

von Eye, A., & Mun E. Y. (2005). *Analyzing rater agreement: Manifest variable methods*. Mahwah, NJ: Erlbaum.

Wang, J., Engelhard, G., & Wolfe, E. W. (2016). Evaluating rater accuracy in rater-mediated assessments with an unfolding model. *Educational and Psychological Measurement, 76*, 1005–1025.

Wang, J., Engelhard, G., 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.

Wesolowski, B. W., & Wind, S. A. (in press). Investigating rater accuracy in the context of secondary-level solo instrumental music. *Musicae Scientae*.

Wesolowski, B., Wind, S.A., & Engelhard, G. (2016). Examining rater precision in music performance assessment: An analysis of rating scale structure using the multifaceted Rasch partial credit model. *Music Perception, 33*(5), 662–678.

Wind, S.A., & Engelhard, G. (2013). How invariant and accurate are domain ratings in writing assessment? *Assessing Writing, 18,* 278–299.

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. (2014). 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.

Wang, J. & Engelhard, G. (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.

Wind, S. A., & Peterson, M. E. (2017). A systematic review of methods for evaluating rating quality in language assessment. *Language Testing*, doi: 10.1177/0265532216686999