Evaluating Rater Accuracy in Rater-Mediated Assessments Using an Unfolding Model

Educational and Psychological
Measurement

2016, Vol. 76(6) 1005–1025

© The Author(s) 2015

Reprints and permissions:

sagepub.com/journalsPermissions.nav

DOI: 10.1177/0013164415621606

epm.sagepub.com

Jue Wang¹, George Engelhard Jr.¹, and Edward W. Wolfe²

Abstract

The number of performance assessments continues to increase around the world, and it is important to explore new methods for evaluating the quality of ratings obtained from raters. This study describes an unfolding model for examining rater accuracy. Accuracy is defined as the difference between observed and expert ratings. Dichotomous accuracy ratings (0 = inaccurate, 1 = accurate) are unfolded into three latent categories: inaccurate below expert ratings, accurate ratings, and inaccurate above expert ratings. The hyperbolic cosine model (HCM) is used to examine dichotomous accuracy ratings from a statewide writing assessment. This study suggests that HCM is a promising approach for examining rater accuracy, and that the HCM can provide a useful interpretive framework for evaluating the quality of ratings obtained within the context of rater-mediated assessments.

Keywords

rater accuracy, rater-mediated assessments, unfolding models, hyperbolic cosine model

Assessment systems that go beyond selected-response items and incorporate constructed-response items that require scoring by raters can be defined as rater-mediated assessment systems. Some examples of rater-mediated assessment systems are performance assessments, essays, and portfolios (Johnson, Penny, & Gordon,

¹The University of Georgia, Athens, GA, USA

²Pearson, Iowa City, IA, USA

Corresponding Author:

Jue Wang, The University of Georgia, 126C Aderhold Hall, 110 Carlton Street, Athens, GA 30602, USA.
Email: cherish@uga.edu

2009). The most commonly used indicator of rating quality in rater-mediated assessments is rater agreement. There are other models for evaluating rater behaviors that include the rater bundle model (Wilson & Hoskens, 2001), hierarchical rater model (Patz, Junker, Johnson, & Mariano, 2002), a signal detection rater model (DeCarlo, Kim, & Johnson, 2011), and a latent trait model (Wolfe & McVay, 2012). Rasch measurement models have also been used to evaluate the psychometric quality of ratings related to rater errors and biases (Myford & Wolfe, 2003, 2004), as well as rater accuracy (Engelhard, 1996, 2013; Engelhard, Davis, & Hansche, 1999; Razynski, Engelhard, Cohen, & Lu, 2015). Models based on generalizability theory can also be used to examine sources of variations that attribute to persons, judges, and tasks (Brennan, 1992).

One of the key questions underlying rater-mediated assessment is: How do we know that raters are providing good ratings? Rater agreement indices (von Eye & Mun, 2005) as well as other rater error and bias indices can be considered *indirect measures* of rating quality. On the other hand, rater accuracy indices offer *direct measures* for exploring how closely a set of observed ratings matches a set of known *true* ratings obtained from expert raters. True ratings can be defined by a panel of experts who assign ratings to a set of performances that used to evaluate the quality of the ratings. Wolfe and McVay (2012) have also defined the true ratings based on average ratings across a group of raters. Engelhard (1996, 2013) defined the rater accuracy as a latent variable that can be objectively evaluated by the distance between observed rating and true ratings that are obtained from a panel of experts. Wolfe et al. (2014) summarized severity/leniency, centrality/extremity, and accuracy/inaccuracy as major types of rater effects. Marcoulides and Drezner (1993, 1997, 2000) have developed another approach to evaluate performance assessments based on an extension of generalizability theory. This model provides an opportunity to include reliability and diagnostic information at both the group and individual levels (Marcoulides & Drezner, 1997). Each of these different measurement theories provides different perspectives for thinking about rater-mediated assessments and offer different statistical models for evaluating the quality of ratings obtained from raters.

Unfolding models were originally developed in the context of attitude measurement by Thurstone (1927, 1928). They were used by Thurstone and Chave (1929) in the development of a scale for measuring attitude toward the church. The study of unfolding models has been approached from a deterministic perspective by Coombs (1964), and probabilistic unfolding models have been proposed by several researchers (Andrich, 1988, 1995; Luo, 1998, 2001; Roberts & Laughlin, 1996; Roberts, Donoghue, & Laughlin, 2002). Bennett and Hays (1960) also described an extension of Coombs's unfolding methods (1950, 1952) for multidimensional unfolding for ranked preference data. Unfolding models have been used in several substantive areas including preference studies (Coombs & Avrunin, 1977), studies of human development (Davison, 1977), and analyses of voting patterns among political parties (Poole, 2005). Currently, it is lack of research on unfolding models for rater evaluation within the context of performance assessments in educational and psychological research.

According to Andrich (1997), the responses of persons to items can be viewed as the result of a cumulative or noncumulative (unfolding) process. In cumulative response processes, the probability of a positive response increases monotonically as a function of the latent variable, while unfolding response processes do not increase monotonically—they reflect single-peaked response functions. Applying unfolding models can yield new indices to identify raters and essays on a continuum of accuracy. Current indices of rater accuracy do not provide information regarding the direction of inaccuracy (Wolfe, 2014). For example, the accuracy index suggested by Engelhard (1996) cannot differentiate raters who tend to give ratings that are lower or higher than deserved by the performances. In other words, it will be useful to have rater quality indices that include information regarding the directionality of inaccurate ratings. It will also be useful to identify the types of essay performances that lead to observed ratings where a rater tends to be more or less accurate.

Purpose of Study

The major purpose of this study is to describe an unfolding model for examining rater accuracy. Specifically, we use the hyperbolic cosine model (HCM; Andrich & Luo, 1993) as a new interpretive framework for examining rater accuracy within the context of rater-mediated assessments. We will briefly describe the idea of unfolding processes and the HCM. This is followed by illustrative and empirical data analyses to demonstrate the use of the HCM for evaluating rater accuracy. Suggestions are made for future research on unfolding models, and the implications for evaluating the quality of ratings obtained in rater-mediated assessments.

Conceptualizing Rater Accuracy as an Unfolding Process

In this study, we define accuracy as a latent variable, and indices of rater accuracy are developed to draw inferences about rating quality (Engelhard, 2013). We define the accuracy ratings as the absolute differences between observed ratings of operational raters and expert ratings of a carefully selected panel of expert raters. If expert ratings are not available, then other approaches can be used to define criterion ratings (e.g., average ratings across operational raters). For example, if we have six essays rated dichotomously (0 = Fail, 1 = Pass) by one rater:

Ratings	Essays					
	1	2	3	4	5	6
Observed	1	0	1	1	0	1
Expert	0	1	1	1	0	0
Difference	1	-1	0	0	0	1
Accuracy	0	0	1	1	1	0

This rater rated Essays 1, 3, 4, and 6 as passing, while the expert ratings provided by an expert panel indicated that Essays 2, 3, and 4 were passing. Then this rater has an accuracy rate of 50% (3 accurate ratings out of 6 ratings). The difference between observed and expert ratings can be statistically modeled in several different ways. One approach suggested by Engelhard (1996) has defined accuracy ratings, A_{ni} , as

$$A_{ni} = \max\{|R_{ni} - B_i|\} - |R_{ni} - B_i|,$$

where R_{ni} is the observed rating from operational Rater n on Essay i , and B_i is the expert rating on Essay i . In this way, all the possible values of accuracy ratings are in the positive direction. This is the approach for defining accuracy that was used by Razynski et al. (2015). Sulsky and Balzer (1988) should be consulted for other indices of rater accuracy.

When an unfolding model is used to analyze the distances between observed and expert ratings, the essay locations as a latent variable can be used to evaluate rater accuracy. In essence, we are creating an approach for monitoring raters that is conceptually similar to indices that are used in quality control situations with an explicit definition of known values to monitor rating quality (Shewhart, 1939).

The distinction between cumulative and unfolding data structures within the context of rater accuracy is illustrated in Table 1. The accuracy rate for an essay is the percentage of accurate ratings for each essay. Similarly, the accuracy rate for a rater is the percentage of accurate ratings for each rater. Panel A in the left column of Table 1 shows the cumulative accuracy ratings (0 = inaccurate, 1 = accurate) for seven raters. The essays vary from Essay 1 that the fewest raters score accurately to Essay 6 that most of the raters score accurately. This pattern of accuracy ratings mimics the structure of a Guttman scale with the iconic triangular pattern of ratings. These accuracy ratings can be modeled with a cumulative response model (e.g., the Rasch model), and an example of a cumulative response function is shown in Panel B. The probabilities of accurate responses are monotonically increasing as a function of the latent variable of rater accuracy on a logit scale.

Panel C in the right column of Table 1 shows a set of unfolding accuracy ratings. This set of seven raters exhibits a deterministic unfolding pattern that reflects a distinctive parallelogram pattern. The underlying assumption of a cumulative data pattern (i.e., Guttman pattern) is that essays are ordered from easy to difficult to score. However, raters may be more accurate on certain types of essays and less accurate on other essays with characteristics that make them more difficult to score accurately. Importantly, the raters may also vary in accuracy across the essays. Wolfe and his colleagues indicated that raters with different proficiencies focus on different characteristics of essays, and raters may use different strategies in scoring essays based on these characteristics (Wolfe & Feltovich, 1994; Wolfe & Kao, 1996; Wolfe, Song, & Jiao, 2016). The underlying assumption of the unfolding data pattern (i.e., parallelogram) is that raters can score accurately on the essays that located close on the line, while essays located below or above may be scored less accurately. The accuracy ratings in Panel C suggest an unfolding response process with a parallelogram pattern in

Table 1. Response Patterns for Cumulative and Unfolding Accuracy Ratings (0 = Inaccurate, 1 = Accurate).

Panel A: Cumulative accuracy ratings							Panel C: Unfolding accuracy ratings						
Raters	Essays						Raters	Essays					
	1	2	3	4	5	6		1	2	3	4	5	6
A	1	1	1	1	1	1	A	1	1	0	0	0	0
B	0	1	1	1	1	1	B	1	1	1	0	0	0
C	0	0	1	1	1	1	C	0	1	1	0	0	0
D	0	0	0	1	1	1	D	0	0	1	1	0	0
E	0	0	0	0	1	1	E	0	0	1	1	1	0
F	0	0	0	0	0	1	F	0	0	0	1	1	0
G	0	0	0	0	0	0	G	0	0	0	1	1	1
Accuracy rate (%)	14.3	28.6	42.9	57.1	71.4	85.7	Accuracy rate (%)	28.6	42.9	57.1	57.1	42.9	14.3
Panel B: Response function for cumulative data							Panel D: Response function for unfolding data

the ratings. Essays vary from Essays 3 and 4 with accuracy rates of 57.1% to Essay 6 with an accuracy rate of 14.3%. In this case, Essays 2 and 5 have comparable accuracy rates (42.9%), but different sets of raters score different essays accurately: Raters A, B, and C score Essay 2 accurately, while Raters E, F, and G score Essay 5 accurately. These accuracy ratings can be modeled with an unfolding model that offers the potential of separating out these differences between essays and raters. The probabilities of accurate responses that reflect an unfolding model with a single-peaked response function is shown in Panel D.

Andrich (1988) proposed a probabilistic item response theory (IRT) model called Squared Simple Logistic Response model (SSLM) for analyzing unfolding preference data. He compared SSLM with Bradley–Terry–Luce (BTL) model for analyzing cumulative responses. In his work, a parameter representing the distance between two stimuli is found in both models, but an additional term referring to the distance between person location and the midpoint of two stimuli is included only in the SSLM. Andrich (1988) also indicated that the nature of the task and the data could be used to decide whether cumulative or unfolding models are more appropriate. Andrich (1988) proposed three ways to discover the underlying response function. First, we can check if the person response patterns with ordered statements are consistent with the theoretical data structure. If the nature of the data is cumulative, then the response patterns should be similar to Guttman patterns with the items ordered with a cumulative model. When the response patterns form a parallelogram, then an unfolding response process is suggested with the item responses matching an unfolding model. Second, we can evaluate the empirical order of the items to check if it matches a theoretical order based on an unfolding model. Finally, we can examine model-data fit with a chi-square test of the correspondence between observed and expected responses based on an unfolding model.

The Hyperbolic Cosine Model

Within the context of rater-mediated assessments, if a rater's unfolding location is close to the essay's location on the underlying accuracy continuum, then this rater tends to be accurate on this essay. Raters who assign higher ratings than experts are in the inaccurate above latent category, while those who assign lower ratings are in the inaccurate below latent category. This response process leads to a single-peaked response function that can be modeled with an unfolding model, such as the HCM used in this study.

There have been several different unfolding models proposed in the literature (Davison, 1977; Kyngdon, 2005; Poole, 1984; Post, 1992; Roberts & Laughlin, 1996; van Schuur, 1989). Luo and Andrich (2005) provided an overview and a discussion of several unidimensional unfolding models in terms of their information functions. HCM has different properties since the unit parameter is a property of the data and independent of the scale compared to PARELLA model (Hojtink, 1990). The HCM uses the math function $\cosh(x)$ to unfold the responses instead of squaring the

parameters, which is used in SSLM. In this study, we focus on the use of HCM to evaluate rater accuracy. The HCM is derived from the Rasch model with three ordered response categories (Andrich & Luo, 1993). The HCM can be written in the following form:

$$P\{x=0\} = \frac{\cosh(\rho_i)}{\cosh(\rho_i) + \cosh(\beta_n - \delta_i)}, \quad (1)$$

$$P\{x=1\} = \frac{\cosh(\beta_n - \delta_i)}{\cosh(\rho_i) + \cosh(\beta_n - \delta_i)}, \quad (2)$$

where x denotes the observed responses with 0 representing an inaccurate rating and 1 as an accurate rating, so that the corresponding probability function is the probability of being inaccurate and accurate. β_n represents the location of Rater n , and δ_i refers to the location of Essay i . The parameter ρ_i is a unit parameter for an essay that is the distance between essay location δ_i and unfolded thresholds τ_1 and τ_2 for three latent categories. The distance between these two thresholds reflects a zone of accuracy within the context of rater-mediated assessments.

The joint maximum likelihood estimation method with Newton–Raphson iteration algorithm is used for parameter estimation. One constraint is that the summation of estimated essay locations is 0. Three parameters are estimated: β_n , δ_i , and ρ_i . The information function is obtained in a similar way as Rasch models by incorporating Fisher information, which is based on Cramér–Rao inequality (Fisher, 1922). Following Samejima (1969, 1977, 1993), Luo and Andrich (2005) proposed an item information function for HCM as follows:

$$I = I_{ni} - P_{ni}(1 - P_{ni})\tanh^2(\beta_n - \delta_i). \quad (3)$$

The information function is moderated by two components that are the distance between β_n and δ_i and the parameter ρ_i that is used to estimate P_{ni} . P_{ni} refers to the probability of accurate responses of Rater n for Essay i as shown in Equation 1. The information is 0 when β_n equals to δ_i . It reaches a maximum when P_{ni} equals $1 - P_{ni}$ (i.e., P_{ni} equals .50), that is when the distance between β_n and δ_i is equal to ρ_i . The range for $\tanh(x)$ is between -1 and 1 , so the maximum value for the information function of HCM is approaching .25. The Hyperbolic Cosine ($\cosh$) and Hyperbolic Tangent ($\tanh$) mathematical functions used in Equations 1, 2, and 3 are defined as

$$\cosh x = \frac{\exp(x) + \exp(-x)}{2}, \quad (4)$$

$$\tanh x = \frac{\exp(x) - \exp(-x)}{\exp(x) + \exp(-x)}. \quad (5)$$

Andrich (1995) proposed two statistical tests to evaluate model-data fit. The first one is an overall test of fit. The hypothesis for the overall test of fit is that the responses correspond to a single-peaked form based on an unfolding model. It uses a Pearson χ^2 statistic:

$$\chi^2 = \sum_{i=1}^I \sum_{g=1}^G \frac{\left(\sum_{n \in N_g} X_{ni} - \sum_{n \in N_g} P_{ni} \right)^2}{\sum_{n \in N_g} P_{ni}(1 - P_{ni})}, \quad (6)$$

where $g = 1, \dots, G$ are the number of clusters that raters are divided into and N_g refers to a set of raters in each cluster g . This statistic approximates the χ^2 distribution when the number of clusters and essays increase. If the Pearson χ^2 statistic is not significant with degrees of freedom of $(G - 1)(I - 1)$, it indicates an acceptable overall fit. Second, a likelihood ratio test is used to examine whether the unit parameter is equal across all the essays. It uses the model comparison idea to evaluate if the model has a significant improvement by comparing the likelihood obtained with variant units ($L_{\hat{\rho}_i}$) to the likelihood of common unit ($L_{\hat{\rho}}$). The null hypothesis is that ρ_i is equal across items. A significant χ^2 statistic with degrees of freedom of $I - 2$ suggests that variant units should be used. This statistic is given

$$\chi^2 = -2 \log(L_{\hat{\rho}}/L_{\hat{\rho}_i}). \quad (7)$$

Using HCM With Illustrative Data

Illustrative unfolding data including seven raters (A to G) and six essays are shown in Table 2. Rater and essay locations are estimated based on HCM (Table 2). The *RateFOLD* computer program (Luo & Andrich, 2003) is used for data modeling. It is informative to see the relationship between HCM locations and the accuracy rates of essays (Figure 1). A simple polynomial model is fit to these data, and this model fits very well with $R^2 = .98$. It highlights very clearly the distinction between Essays 2 and 5 that shared the same accuracy rates, but that are estimated to have different locations based on HCM. Although this study does not focus on a comparison between measurement theories, the essay locations were also estimated by a dichotomous Rasch model using the Facets computer program (Linacre, 2015). The Rasch model locations for these items are 0.94, 0.29, -0.30 , -0.30 , 0.29, and 1.83 logits, respectively. Essays that share the same accuracy rates have the same Rasch location estimates; for example, Essays 2 and 5 have the same location of .29 logits, and Essays 3 and 4 share the same location of -0.30 logits. Therefore, if we replace the accuracy rates by the Rasch location estimates for essays in Figure 1, we will also obtain a polynomial curve. This supports the earlier observation that essays with same accuracy rates may be scored accurately by different groups of raters, and HCM can capture this information.

Table 2. Illustrative Data and HCM Location Estimates for Essays and Raters.

	Essays						Rater accuracy rate (%)	HCM locations for raters (SE)
	1	2	3	4	5	6		
Raters								
A	1	1	0	0	0	0	33.3	-5.79 (1.93)
B	1	1	1	0	0	0	50.0	-3.93 (1.38)
C	0	1	1	0	0	0	33.3	-2.54 (1.24)
D	0	0	1	1	0	0	33.3	-0.02 (1.14)
E	0	0	1	1	1	0	50.0	1.19 (1.21)
F	0	0	0	1	1	0	33.3	2.61 (1.51)
G	0	0	0	1	1	1	50.0	4.16 (1.37)
Essay accuracy rate (%)	28.6	42.9	57.1	57.1	42.9	14.3		
HCM locations for essays (SE)	-5.98 (0.72)	-4.49 (0.68)	-1.44 (0.48)	2.20 (0.61)	3.38 (0.65)	6.32 (0.70)		

Note. HCM = hyperbolic cosine model.

Figure 1. Plot of accuracy rates and HCM essay locations for illustrative data.

On the other hand, Raters A, C, D, and F all have the same accuracy rates, but these raters are accurate on different sets of essays (Figure 2). Under HCM, their locations on the accuracy continuum are -5.79 , -2.54 , -0.02 , and 2.61 logits respectively. The location estimates of the raters based on the dichotomous Rasch model also depend on the raw scores (or comparable to accuracy rates). Raters A, C, D, and F have a same location estimate of -0.33 logits with a same raw score of two. Similarly, Raters B, E, and G have a single estimate 0.44 logits with a raw score of three. Unfolding model can differentiate the raters who have the same accuracy ratings obtained on different essays.

HCM also includes a unit parameter ρ_i for each essay, and the estimate of this unit parameter can be used to identify a zone of accuracy by defining thresholds (τ_1 and τ_2) as plus or minus one unit about an essay's location. This zone of accuracy is an additional feature compared to the Rasch model that can help the researcher to identify the raters who have a probability greater than .50 of scoring a specific essay accurately. It also provides information about the raters that tend to score inaccurate below and inaccurate above (Figure 3). As shown in Figure 3, raters within thresholds τ_1 and τ_2 have an estimated probability of accurate response that is higher than .50. The distance between the location δ_3 and each threshold is the unit parameter ρ_3 . The statistical test for equal unit parameters supports the inference that the units are equal based on the likelihood ratio test, $\chi^2(4) = 2.74, p = .59$.

Unlike the single-peaked rater information functions of cumulative responses, the rater information functions for unfolding responses are bimodal (Figure 4). The information function curve has two peaks. It reaches 0 when the rater location is the same

Figure 2. Variable map for unfolding rater accuracy for illustrative data.

as the essay location. It has the maximum when the distance between rater and essay locations is equal to the unit parameter.

The data in this section illustrated the use of HCM for modeling accuracy data. As expected, the HCM unfolding model fit the illustrative data with a parallelogram data structure with a good overall test of fit, $\chi^2(35) = 9.26, p > .999$.

Using HCM With Empirical Data

The use of HCM for analyzing empirical data within the context of rater-mediated assessments is illustrated in this section. Writing data from Gyagenda and Engelhard (2010) are used with essays from 8th grade students ($I = 50$) rated by randomly selected raters ($N = 20$) from a large-scale statewide writing assessment. The essays were also rated by a validity panel that defined the expert ratings. The original data had four domains, and the dichotomized accuracy ratings for the domain of Style is

Figure 3. Probability function for three latent ordered categories for illustrative data.

Note. δ_3 represents the location of Essay 3. The parameter ρ_3 is a unit parameter for Essay 3, which is the distance between the essay location δ_3 and each unfolded thresholds τ_1 and τ_2 for three latent categories. The distance between these two thresholds is the zone of accuracy for Essay 3.

Figure 4. Information function of illustrative data for Essay 3.

used in this section (0 = inaccurate, 1 = accurate). The accuracy rates for the essays range from 48.0% to 78.0%. The raters are moderately accurate with almost all the raters having an accuracy rate between 50.0% and 80.0%. As in the previous section, the *RateFOLD* computer program (Luo & Andrich, 2003) is used to conduct the data analyses.

The overall test of fit for common units is acceptable, $\chi^2(949) = 927.78$, $p = .68$. The likelihood ratio test indicated that there is no significant improvement by using variant units as compared to common units for all essays, $\chi^2(48) = 20.06$, $p = .999$. For illustrative purposes, the results with both common and variant units are reported.

Figure 5. Variable maps for unfolding rater accuracy.
Note. Common units have a fixed zone of accuracy, variant units have different zones of accuracy for each essay.

The variable map shows the locations for both essays and raters on the same scale (Figure 5). Overall, raters are located closer to the essays they tend to score accurately. Even though the test for equal units is not statistically significant, the location estimates for essays are different on the variable maps in the two panels. As expected, raters are more spread out on the variable map when variant units are used.

A simple polynomial model between essay locations of HCM and accuracy rates of essays fits the empirical data quite well with $R^2 = .98$ (Figure 6). It differentiates Essays 46 and 33 clearly. These two essays share the same accuracy rate (50%) with equivalent location estimate obtained with the dichotomous Rasch model as -0.79 logits. However, HCM provides different locations for Essays 46 (-4.27) and 33 (2.65). Essays 29, 45, 12, and 44 all have the same accuracy rate (95%) and Rasch location estimate (2.25 logits), but they have different HCM locations. Essays 29 and 45 have equal location of -1.78 logits, and Essays 12 and 44 also have equivalent location of 0.082 logits. Similarly, raters who have the same location estimate in the Rasch model have been estimated differently under HCM. Therefore, HCM is capable of differentiating essays and raters with different response patterns.

With common units, the zones of accuracy for all of the essays are the same. In order to better understand the parameters, the probability curves for the essays with

Figure 6. Plot of accuracy rates and HCM essay locations with empirical data.

variant units are shown in Figure 7. First, the zone of accuracy indicates the range of raters who tend to score accurately on this essay. Second, the raters who are located outside the zone of accuracy tend to score this essay inaccurately. Raters who tend to score inaccurate below or above may due to different types or features of the essays. Essay 35 has a smaller zone of accuracy than Essay 29. With a smaller zone of accuracy, it indicates that this essay has fewer raters scoring accurately. If it is an essay used in rater training, it may be merit examination by content specialists. The information regarding scoring this essay provided by two groups of inaccurate raters as well as the few accurate raters could also be useful for examining studies of rater perceptions and judgments. With an additional feature of a zone of accuracy for each essay of HCM, it is possible and convenient to find subsets of raters who are accurate, inaccurate below, and inaccurate above for a set of essays.

The information functions for Essays 35 and 29 are shown in the Figure 8. Similar to the illustrative data analysis, the information function has two peaks. It reaches 0 for the raters who have the same location with the essay, and it reaches the maximum when the probability of accurate response is .5 (i.e., the distances between rater and essay locations are equal to the unit parameter). The theoretical maximum value of information for each essay is .25, and it relates to its unit parameter. The information function reflects the precision of the estimates for the rater accuracy.

The expected curves provide information about model-data fit (Figure 9). The observed ratings of 5 groups that 20 raters are divided ($N = 4$ in each group) and the model-based expected curves are displayed. A nonsignificant p value indicates model and data fit. Essay 10 is clearly a misfitting essay. The observed responses do not fall on the expected curve. And it is the only misfitting essay with an alpha value of .01. Essays 15 and 35 both fit the model. The observed responses are relatively close to the expected curve. A larger p value or smaller χ^2 value given the same degree of freedom indicates a better fit between model and data. Therefore, Essay 35 has a better fit than Essay 15.

Figure 7. Probability functions for selected essays.

Note. δ_i represents the location of Essay i . The parameter ρ_i is a unit parameter for essays, which is the distance between the essay location δ_i and each unfolded thresholds τ_1 and τ_2 for three latent categories. The distance between these two thresholds reflects a zone of accuracy. Raters within the zone of accuracy are more likely to score accurately on the corresponding essay.

Discussion and Summary

Most of the reservations [about rating scales], regardless of how elegantly phrased, reflect fears that rating scale data are subjective (emphasizing, of course, the undesirable connotations of subjectivity), biased, and at worst, purposefully distorted. (Saal, Downey, & Lahey, 1980, p. 413)

Accuracy ratings can be defined as the distance between observed ratings from operational raters and expert ratings defined by a panel. Engelhard (1996, 2013) proposed

Figure 8. Information functions for selected essays.

the use of Rasch measurement theory based on the Many Faceted Rasch Model for measuring rater accuracy by using accuracy ratings. One of the limitations of previous approaches for quantifying rater accuracy is that they do not differentiate the direction of inaccuracy (inaccurate below expert ratings and inaccurate above expert ratings). Another limitation of previous research is that no information is provided regarding the range or the zone of accuracy exhibited by raters. The unfolding models proposed by Andrich and Luo (Andrich, 1988, 1997; Andrich & Luo, 1993; Luo & Andrich, 2005) offer a promising approach for evaluating rater accuracy. Unfolding models have typically been used to measure attitudes (Andrich, 1988) with the latent continuum defined by a set of items that are rated using several categories (e.g., agree, neutral, disagree). In this study, accuracy ratings are treated as unfolding data.

Figure 9. Expected curves for selected essays.

Typically, Rasch measurement and IRT models view rating scores as reflecting a cumulative response process. In contrast, unfolding models view the ratings as non-monotonic functions of the underlying latent continuum.

This study illustrates the use of the HCM for unfolding rater accuracy. Our goal was to present both conceptual and empirical evidence to support the potential usefulness of the HCM for evaluating rating accuracy. Research is still needed to investigate why the essays are ordered on a continuum (Wolfe et al., 2016). Future research is needed to refine the selection of the essays that define the underlying continuum. It is important to explore whether or not it is possible to deliberately create or identify essays that meaningfully represent an underlying accuracy continuum: why are some essays rated lower than the expert ratings, and why are some essays rated higher than the expert ratings? Therefore, rater perception and cognition can be interactively examined with different characteristics of essays.

In addition to the HCM used in this study, future research should (1) analyze rater judgment toward characteristics of essays with mixed-methods design; (2) include detailed comparisons between HCM and other measurement models (e.g., Rasch measurement theory and generalizability theory); (3) apply HCM to polytomous data with an ordered categorical scales; and (4) examine other unfolding models developed for the measurement of attitudes that might be useful for modeling rater accuracy data. We look forward to more applications of the concept of rater accuracy within the context of other types of performance assessments that are rater-mediated.

No single indicator of the psychometric quality of ratings can identify all aspects of rater behaviors. Unfolding models offer a promising set of previously unexplored indices that can be added to the current array of indices for examining rater agreement, rater errors and biases, and rater accuracy. There are several implications of the use of unfolding models to examine rater accuracy. In this study, we estimated rater accuracy locations and zones of accuracy, and these indices hold promise for inclusion as a part of rater training, examining rater cognition, and the ongoing monitoring of rater performance in large-scale rater-mediated assessment systems. The idea of applying unfolding models to evaluate rater accuracy is new, and we believe that it offers a promising approach for evaluating the quality of rater-mediated assessments.

Authors' Note

Researchers supported by Pearson (the funding agency) are encouraged to freely express their professional judgment. Therefore, the points of view or opinions stated in Pearson-supported research do not necessarily represent official Pearson position or policy.

Acknowledgments

We would like to thank Professors David Andrich and James Roberts for helpful comments and discussions of unfolding models.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: Pearson provided support for this research.

References

- 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. (1995). Hyperbolic cosine latent trait models for unfolding direct responses and pairwise preferences. *Applied Psychological Measurement, 19*, 269-290.

- Andrich, D. (1997). A hyperbolic cosine IRT model for unfolding direct response of persons to items. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 399-414). New York, NY: Springer.
- Andrich, D., & Luo, G. (1993). A hyperbolic cosine latent trait model for unfolding dichotomous single-stimulus responses. *Applied Psychological Measurement*, 17, 253-276.
- Bennett, J. F., & Hays, W. L. (1960). Multidimensional unfolding: Determining the dimensionality of ranked preference data. *Psychometrika*, 25, 27-43.
- Brennan, R. L. (1992). Generalizability theory. *Educational Measurement: Issues and Practice*, 11(4), 27-34.
- Coombs, C. H. (1950). Psychological scaling without a unit of measurement. *Psychological Review*, 57, 145-158.
- Coombs, C. H. (1952). *A theory of psychological scaling* (Vol. 34). Ann Arbor: Engineering Research Institute, University of Michigan.
- Coombs, C. H. (1964). *A theory of data*. New York, NY: Wiley.
- Coombs, C. H., & Avrunin, C. S. (1977). Single-peaked functions and the theory of preference. *Psychological Review*, 84, 216-230.
- Davison, M. L. (1977). On a metric, unidimensional unfolding model for attitudinal and developmental data. *Psychometrika*, 42, 523-548.
- DeCarlo, L. T., Kim, Y., & Johnson, M. S. (2011). A hierarchical rater model for constructed responses, with a signal detection rater model. *Journal of Educational Measurement*, 48, 333-356.
- Engelhard, G. (1996). Evaluating rater accuracy in performance assessments. *Journal of Educational Measurement*, 33(1), 56-70.
- Engelhard, G. (2013). *Invariant measurement: Using Rasch models in the social, behavioral, and health sciences*. New York, NY: Routledge.
- Engelhard, G., Davis, M., & Hansche, L. (1999). Evaluating the accuracy of judgments obtained from item review committees. *Applied Measurement in Education*, 12, 199-210.
- Fisher, R. A. (1922). On the mathematical foundations of theoretical statistics. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, 309-368.
- Gyagenda, I. S., & Engelhard, G. (2010). Rater, domain, and gender influences on the assessed quality of student writing. In M. Garner, G. Engelhard, M. Wilson, & W. Fisher (Eds.), *Advances in Rasch measurement* (Vol. 1, pp. 398-429). Maple Grove, MN: JAM Press.
- Hooijtink, H. (1990). *PARELLA: Measurement of latent traits by proximity items*. Groningen: University of Groningen, Netherlands.
- Johnson, R. L., Penny, J. A., & Gordon, B. (2009). *Assessing performance: Designing, scoring, and validating performance tasks*. New York, NY: Guilford Press.
- Kyngdon, A. (2005). An introduction to the theory of unidimensional unfolding. *Journal of Applied Measurement*, 7, 260-277.
- Linacre, J. M. (2015). *Facets computer program for Many-facet Rasch measurement, version 3.71.4*. Retrieved from <http://www.winsteps.com/facets.htm>
- Luo, G. (1998). A general formulation for unidimensional unfolding and pairwise preference models: Making explicit the latitude of acceptance. *Journal of Mathematical Psychology*, 42, 400-417.
- Luo, G. (2001). A class of probabilistic unfolding models for polytomous responses. *Journal of Mathematical Psychology*, 45, 224-248.

- Luo, G., & Andrich, D. (2003). RateFOLD [Computer program]. Victoria, Western Australia, Australia: Social Measurement Laboratory, School of Education, Murdoch University.
- Luo, G., & Andrich, D. (2005). Information functions for the general dichotomous unfolding model. In S. Alagumalai, D. D. Curtis, & N. Hungi (Eds.), *Applied Rasch measurement: A book of exemplars* (pp. 309-328). Amsterdam, Netherlands: Springer.
- Marcoulides, G. A., & Drezner, Z. (1993). A procedure for transforming points in multidimensional space to a two-dimensional representation. *Educational and Psychological Measurement*, 53, 933-940.
- Marcoulides, G. A., & Drezner, Z. (1997). A method for analyzing performance assessments. In M. Wilson, G. Engelhard Jr., & K. Draney (Eds.), *Objective measurement: Theory into practice* (Vol. 4, pp. 261-277). Norwood, NJ: Ablex.
- Marcoulides, G. A., & Drezner, Z. (2000). A procedure for detecting pattern clustering in measurement designs. In M. Wilson & G. Engelhard Jr (Eds.), *Objective measurement: Theory into practice* (Vol. 5, pp. 287-302). Norwood, NJ: Ablex.
- Myford, C. M., & Wolfe, E. W. (2003). Detecting and measuring rater effects using many-facet Rasch measurement: Part I. *Journal of Applied Measurement*, 4, 386-422.
- Myford, C. M., & Wolfe, E. W. (2004). Detecting and measuring rater effects using many-facet Rasch measurement: Part II. *Journal of Applied Measurement*, 5, 189-227.
- 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.
- Poole, K. T. (1984). Least squares metric, unidimensional unfolding. *Psychometrika*, 49, 311-323.
- Poole, K. T. (2005). *Spatial models of parliamentary voting*. New York, NY: Cambridge University Press.
- Post, W. J. (1992). *Nonparametric unfolding models: A latent structure approach*. Leiden, Netherlands: DSWO Press.
- Razynski, K., Engelhard, G., Cohen, A., & 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.
- Roberts, J. S., Donoghue, J. R., & Laughlin, J. E. (2002). Characteristics of MML/EAP parameter estimates in the generalized graded unfolding model. *Applied Psychological Measurement*, 26, 192-207.
- Roberts, J. S., & Laughlin, J. E. (1996). A unidimensional item response model for unfolding responses from a graded disagree-agree response scale. *Applied Psychological Measurement*, 20, 231-255.
- Saal, F. E., Downey, R. G., & Lahey, M. A. (1980). Rating the ratings: Assessing the psychometric quality rating data. *Psychological Bulletin*, 88, 413-428.
- Samejima, F. (1969). *Estimation of latent ability using a response pattern of graded scores* (Psychometric Monograph No. 17). Richmond, VA: Psychometric Society. Retrieved from <http://www.psychometrika.org/journal/online/MN17.pdf>
- Samejima, F. (1977). A method of estimating item characteristic functions using the maximum likelihood estimate of ability. *Psychometrika*, 42, 163-191.
- Samejima, F. (1993). An approximation for the bias function of the maximum likelihood estimate of a latent variable for the general case where the item responses are discrete. *Psychometrika*, 58, 119-138.

- Shewhart, W. A. (1939). *Statistical method from the viewpoint of quality control*. Washington, DC: Graduate School of the Department of Agriculture.
- Sulsky, L. M., & Balzer, W. K. (1988). Meaning and measurement of performance rating accuracy: Some methodological and theoretical concerns. *Journal of Applied Psychology*, 73, 497-506.
- Thurstone, L. L. (1927). A law of comparative judgment. *Psychological Review*, 34, 278-286.
- Thurstone, L. L. (1928). Attitudes can be measured. *American Journal of Sociology*, 33, 529-554.
- Thurstone, L. L., & Chave, E. J. (1929). *The measurement of attitude: A psychophysical method and some experiments for measuring attitude toward the church*. Chicago, IL: University of Chicago Press.
- van Schuur, W. H. (1989). Unfolding German political parties: A description and application of multiple unidimensional unfolding. In G. de Soete, H. Ferger, & K. C. Klauer (Eds.), *New developments in psychological choice modeling* (pp. 259-277). Amsterdam, Netherlands: North Holland.
- von Eye, A., & Mun, E. Y. (2005). *Analyzing rater agreement: Manifest variable methods*. Mahwah, NJ: Erlbaum.
- Wilson, M., & Hoskens, M. (2001). The rater bundle model. *Journal of Educational and Behavioral Statistics*, 26, 283-306.
- Wolfe, E. W. (2014). *Methods for monitoring rating quality: Current practices and suggested changes*. Iowa City, IA: Pearson.
- Wolfe, E. W., & Feltovich, B. (1994, April). Learning to rate essays: A study of scorer cognition. Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA.
- Wolfe, E. W., & Kao, C. W. (1996, April). *Expert/novice differences in the focus and procedures used by essay scorers*. Paper presented at the annual meeting of the American Educational Research Association, New York, NY.
- 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., Jiao, H., & Song, T. (2014). A Family of Rater Accuracy Models. *Journal of applied measurement*, 16, 153-160.
- Wolfe, E.W., Song, T., & Jiao, H. (2016). Features of difficult-to-score essays. *Assessing Writing*, 27, 1-10.