Psychological Test and Assessment Modeling, Volume 59, 2017 (4), 471-492

The impact of design decisions on measurement accuracy demonstrated using the Hierarchical Rater Model

Jodi M. Casabianca[1,2] *& Edward W. Wolfe*[2]

Abstract

When humans assign ratings in testing contexts, concern arises about whether rater effects impact the accuracy of the resulting measures. Those who lead scoring efforts implement several activities and utilize various designs to minimize the impact of these rater errors. This article uses the Hierarchical Rater Model (HRM) to demonstrate how the magnitude of rater errors and numbers of ratings associated with various measurement facets (e.g., raters & items) impact the accuracy of measures. Additionally, we demonstrate how the level at which decisions are made about the measures (e.g., test taker item scores, test taker total scores, test taker classifications) impact measurement accuracy.

Keywords: rater effects, measurement accuracy, hierarchical rater model, rating designs

[1] *Correspondence concerning this article should be addressed to:* Jodi M. Casabianca, PhD, Research Scientist, Educational Testing Service, 660 Rosedale Road, MS T-03, Princeton, NJ 08541, USA; email: jcasabianca@ets.org

[2] Educational Testing Service, Princeton, USA

Judgments of the quality of an object are collected in numerous contexts, and the raters, frequently humans with a relevant expertise, utilize numerical or ordinal rating scales to depict the relative quality of a collection of artifacts being judged. In assessment contexts, raters employ rating scales to depict the quality of test taker responses to constructed-response items that appear on educational and certification tests; in these scenarios, the artifacts are the test taker responses which may take the form of essays, mathematical proofs, or performances to name but a few potential response formats. The assigned ratings, sometimes referred to as subjective ratings, are then accumulated across multiple test items, perhaps even including scores assigned to objectively scored items, and the resulting total score is used to make educational and certification assessments regarding the test taker. Inherent in these contexts are decisions about the assessment design and the levels at which assessments might be used. The purpose of this article is to demonstrate how various design decisions such as the number of ratings per response, may impact the quality of measures at different levels. We make this demonstration using the hierarchical rater model (HRM; Casabianca, Junker, & Patz, 2016; Patz et al., 2002), a multilevel item response theory (IRT) model used to scale individuals while also accounting for rater severity and variability.

We selected the HRM because it is a rater model that addresses the problem that comes about from complex ratings designs that include multiple items, multiple raters, and multiple ratings per item × test taker combination. That is, the hierarchical structure of the HRM explicitly models the natural hierarchy that exists in ratings data when there are multiple raters assigning multiple scores to the same responses. This is what Wilson and Hoskens (2001) called the "repeated rating" problem. As Mariano (2002) showed, the problem is that ignoring the hierarchical structure of the ratings data results in an information accumulation problem. In models that ignore this hierarchy, there is a downward bias of standard errors with added raters' ratings per item and a corresponding overestimated reliability. Indeed, one of the most popular rater models, the Facets model (Linacre, 1989), is one that ignores the nesting of multiple ratings within a test taker's response and considers the information from each rating as if it were an item contributing information. Most likely, it is widely used because it is straightforward to understand and well documented. However, researchers in the early 2000s introduced models for ratings that address this information accumulation issue – these models include the HRM, the model for multiple ratings (MMR) by Verhelst and Verstralen (2001) and the rater bundle model (RBM) by Wilson and Hoskens (2001). More recently, DeCarlo, Johnson and Kim (2011) introduced a version of the HRM (HRM-SDT, or HRM-signal detection theory) that uses an expanded signal detection model for rater effects, resulting in richer information about raters compared to the Patz et al. (2002) version of the HRM.

We use the Patz et al. (2002) HRM to discuss measurement at different levels because it is relatively simpler to use for demonstration purposes. The simplest level at which an assessment can be made about a test taker is at the item response level to which a rating is assigned. The rating constitutes a measure of the test taker's performance on the item in question, as interpreted by the rater in question. We can improve upon the quality of that measure by collecting ratings from more than one rater and creating a composite measure of the test taker's performance on that item. Furthermore, we can require the test taker to respond to multiple items (e.g. multiple essay prompts), thus increasing the quality of the measure of the construct in question. That is, by making multiple observations of the test taker's behaviors, we have begun to expand the scope of our consideration beyond the test taker's performance on an individual item and to the latent traits that the items jointly elicit. Often, the multiple ratings across the multiple items are scaled to create a total score, and those scaled scores define a continuum of measures that reveal whether two test takers differ in terms of their performance and also provide depictions of how much they differ. Those measures can be further collated to allow for consideration of the relative performances of multiple groups of test takers (e.g., classrooms or schools in educational settings). Further, bands of scores can be combined to define levels of performance as is often done in educational (e.g., proficient/non-proficient) or certification (e.g., pass/fail) testing.

As different levels of interpretation have implications for the degree to which rater errors will impact the accuracy of the interpretation of measures and because design choices that are geared toward reducing those errors have implications for the cost and feasibility of implementation, it is very important to take into account these various levels at which we may want to make assessments. Thus, the purpose of this article is to demonstrate how design choices impact the accuracy of measures that are estimated by the HRM. We specifically investigate how rater selection (and the associated magnitude of rater errors in the pool of raters), the number of ratings, and the number of items, relate to differences in measurement accuracy at different levels of the score (item vs. total vs. performance category) and the test taker (individuals vs. groups).

We address the following research questions:

1. What is the impact of *rater pool quality* on measurement accuracy?
2. What is the impact of *multiple ratings per item response* on measurement accuracy?
3. What is the impact of *test length* on measurement accuracy?
4. When considering the measurement of individuals, how robust are scores at different levels (item, total, pass/fail) to the impact of these aforementioned design decisions?
5. How does the impact of these design decision differ when measuring individuals versus measuring groups of individuals?

In the remainder of this article, we discuss theoretically how different aspects of the assessment design can impact measurement accuracy and how the HRM can be used to score test takers and estimate rater effects. We then provide a demonstrative example to answer our research questions and close the article with discussion and conclusions.

Impact of design decisions on measurement accuracy

In addition to actions designed to reduce rater errors such as training, calibration, and backreading (see Wolfe, 2014), there are design considerations that may improve the measurement process in a more predictable fashion. Figure 1 provides a schematic that depicts how different design decisions can impact accuracy at different levels. In this study, as we will describe, we follow the framework of this figure and focus on how design decisions impact item-level, total-level, and pass/fail level scores, as well as scores at the individual and at the group level. The design decisions include the number of ratings per response, the number of items on the assessment, and the selection criteria for raters and items, which relate to different levels of rating and item quality. Figure 1 shows these design decisions in a funnel that yields varying levels of measurement accuracy depending on the decisions made. The generalizability theory framework allows us to predict the expected improvement in the "dependability" or the reliability of relative and/or absolute decisions about test takers (Brennan, 2001) when we increase or decrease the number of elements associated with a design facet in the measurement system (e.g., raters, items). In addition, one may wish to weigh the benefits of certain design decisions on scoring at different levels with the added implementation costs, if applicable. We show at the bottom of the funnel in Figure 1 the resultant measurement accuracy, which we know will vary by the score level and the test taker level.

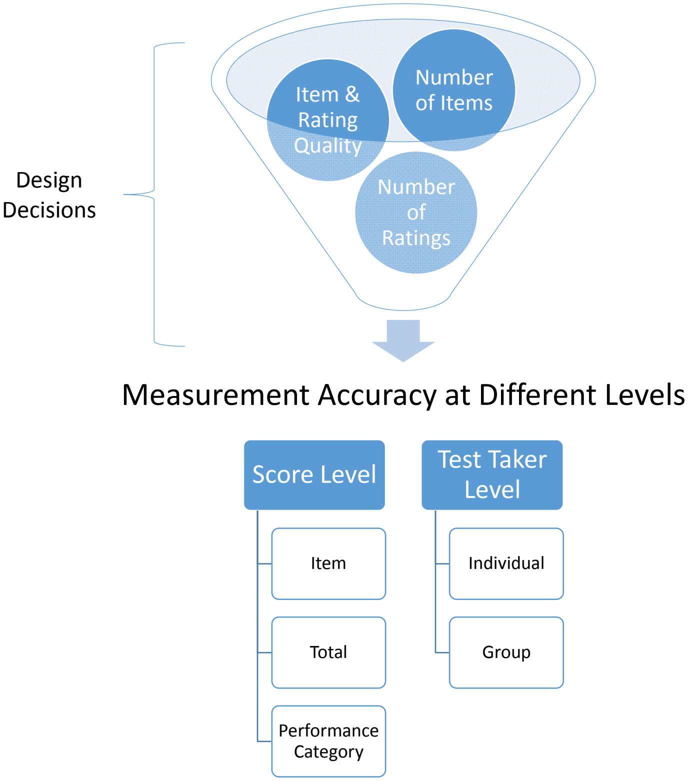

Figure 1:
Considerations in developing rating designs and their impact on measurement accuracy at different levels

Constructed response (CR) items, such as essays, typically require more effort and response time from the test taker, and therefore, we generally do not see tests with many CR items. With this being said, collecting responses from multiple items is by far the most effective way to improve measurement accuracy, assuming the items have adequate discriminating power[3]. The Spearman-Brown prophecy predicts the improved reliability of a test after adding parallel items (Brown, 1910; Spearman, 1910). The relationship between the number of items and reliability is not linear such that when the test reliability is high, it will take many additional parallel items to approach the maximum value of 1.0. However, in general, the addition of items, or in other cases, weighted parallel components to make up a composite test, increases the reliability. The same is true under the IRT framework where item information quantifies the item-level reliability. With more items, the test will have greater test information. Practically, the administration and scoring of multiple items will cost more than doing so for just one item. However, the potential psychometric benefits of additional items may exceed the cost in some cases.

Typically, rating designs incorporate some percentage of double-scored responses in order to perform a rater reliability check. In some instances, all of the responses will be scored by multiple raters, not to estimate rater reliability, but because it is believed that multiple ratings will yield a more accurate representation of the test taker's quality of response. These multiple ratings may be summed or averaged with the hopes that any rater effects are "averaged away" by taking multiple measurements. Naturally, multiple ratings require more rater effort. This can be very costly, so it is important to determine whether or not these additional ratings contribute to measurement quality in a useful way and also understand how this impacts scores at different levels. For example, do improvements in test taker classification accuracy exist due to multiple ratings, or are the positive effects of this design decision lost at the item- or total-score level? Under the generalizability framework, we can study the effect of multiple ratings on scores, and while multiple ratings may improve reliability slightly, research has shown that increasing the number of ratings does not significantly improve measurement accuracy (e.g. see: Brennan, Gao, & Colton, 1995; Kim & Wilson, 2009). When considering both raters and items as facets, adding more raters does not substantially reduce the residual variance, and certainly less so than adding items, as they make different contributions to residual variance. Therefore, while we may observe small improvements in observed scores due to additional ratings, they will not likely amount to the improvements due to additional items. Under the IRT framework, there are variations in the way multiple ratings of the same responses are treated. We discuss this later in this section.

Design decisions about the test development process and the rating process may also lead to an improvement in measurement quality. Specifically, this relates to selection of items and raters. Informed selection of highly discriminating items that are also not overly

[3] Note that here we are referring to the situation in which there are actually multiple items or prompts eliciting multiple responses from a test taker. The situation may also be that there is one response to a single item / prompt but a rater is applying a rubric with multiple dimensions and thus assigns multiple ratings reflecting different evaluations of the test taker. For simplicity, we restrict our discussion to the former case.

difficult or easy and are not overly-subjective to score will yield better quality of measurement by improving inter-item correlations and thus the overall quality of final measures. Similarly, data-driven selection of "high performing" raters, based on their accuracy rates and indices of their rater effects, will yield better ratings with reduced rater errors thereby improving the quality of response level observations. It is also important to consider raters who are responsive to training and feedback when they do make errors. The cost associated with these selection decisions are not as quantifiable as the other design decisions. The expense of writing and testing CR items is typically built into the assessment process. However, selecting high performing raters does require effort related to establishing performance indices, as well as the possible additional monetary compensation these raters may require due to their expertise.

As we alluded to earlier, when considering the impact of these decisions, we must consider the level at which of decision-making will occur with respect to score interpretation. For example, how impactful are these decisions at the level of the item score versus the total score (average of CR item scores) versus classification in performance categories? How much of a difference do these decisions make when considering an individual test taker's score versus a summary of a group's overall score? Understanding the impacts on these different levels and how they interact with measurement goals will motivate design decisions, especially if cost is a deciding factor.

Scoring models treat ratings of constructed responses differently. In an observed score framework, scoring may include an aggregation of ratings. There are various scoring possibilities under a latent variable framework, including: (i) general IRT models assuming no rater effects (e.g. 2-parameter logistic model, generalized partial credit model), (ii) IRT models with rater parameters (e.g. facets), and (iii) IRT rater models that incorporate the hierarchical structure of the rating design, or hierarchical rater models. The last type of scoring possibility includes the class of models that do not treat multiple ratings of the same response as additional information; the hierarchy is that there are multiple ratings which are a function of the "true" quality of the response along with rater error by way of rater parameters reflecting different rater effects. This article demonstrates the impact of different design decisions on the accuracy of scores at different levels (as depicted by Figure 1) using the HRM.

Using the Hierarchical Rater Model to estimate rater severity and unreliability

Measurement error is introduced into test taker measures in several ways, but the introduction of error due to raters and the rating process is of particular concern when responses are scored through a subjective decision-making process. When rater errors influence ratings in a consistent manner, recognizable patterns can be observed in the ratings, and those patterns are indicative of rater effects. Numerous rater effects exist, so we focus our attention on just two commonly observed effects that are captured by the HRM, severity and individual rater unreliability/inconsistency.

When raters assign ratings that are consistently lower or higher than a known-to-be-valid rating, we say that the rater exhibits *severity* or *leniency*. A severe rater assigns ratings that are too low, given the test taker's true performance, and a lenient rater assigns ratings that are too high. This results in underestimation of the test taker's performance on the item by severe raters and overestimation of the test taker's performance on the item by lenient raters.

When raters assign ratings that consistently exhibit less or more random variability around known-to-be-valid ratings, we say that the rater is exhibiting an *accurate* or *inaccurate* rating pattern. An accurate rater assigns ratings that are very similar to the true performances of the test takers, which is desirable. Hence, accurate ratings tend to be consistent with known-to-be-valid ratings. On the other hand, an inaccurate rater assigns ratings that exhibit a high level of random deviation from the true performances of test takers. This results in an accurate estimation of test taker performance by accurate raters and generally poor estimation of the performance of test takers regardless of their level of performance by an inaccurate rater's ratings. As we will discuss, in the case of the HRM, a rater's variability is somewhat related to accuracy, however the variability is captured around the known-to-be-valid rating offset by an individual rater's bias.

The HRM posits that a test taker's response to an item may be (hypothetically) judged to have some true rating or quality, we call this a test taker's "ideal rating" on an item *j* (*j* = 1, …, *J*). Then, a series of *R* raters evaluate the responses, giving observed ratings based on their observations and their understanding of the scoring rubric. The HRM hierarchy connects this two-stage rating process with an IRT model and a signal detection model. Specifically, in the first stage, an IRT model defines the relationship between the ideal ratings and the latent trait. In the second stage, a "signal-detection-like" model defines the relationship between the ideal rating of an indicator and multiple raters' observed ratings.

In the HRM, the first level of the hierarchy models the distribution of ratings given the quality of response (or ideal rating/response), the second level models the distribution of a test taker 's response (ideal ratings) given their latent trait, and the third level models the distribution of the latent trait θ. The hierarchical representation of the HRM is given by

$$\begin{aligned}
X_{ijr} \mid \xi_{ij}, \tau_r^2, \phi_r &\sim \text{polytomous signal detection model}, r = 1,\ldots,R, \text{for each } i, j. \\
\xi_{ij} \mid \theta_i, \beta_j, \gamma_{jk} &\sim \text{polytomous IRT model}, j = 1,\ldots,J, \text{ for each } i \\
\theta_i &\sim N(\mu_\theta, \sigma_\theta^2), i = 1,\ldots,N \text{ where } \sigma_\theta^2 = 1/\omega \\
\omega &\sim Gamma\left(a_\omega, b_\omega\right) \\
\beta_j &\sim N(\mu_\beta, \sigma_\beta^2) \\
\gamma_{jk} &\sim N(\mu_\gamma, \sigma_\gamma^2) \\
1/\tau_r^2 &\sim Gamma\left(a_{1/\tau^2}, b_{1/\tau^2}\right) \\
\phi_r &\sim N(\mu_\phi, \sigma_\phi^2).
\end{aligned} \tag{1}$$

Here, θ_i, the latent trait for test taker i $(i = 1,\ldots,N)$ is normally distributed with mean μ_θ and σ_θ^2, ξ_{ij} is the ideal rating for test taker *i* on item *j,* and X_{ijr} is the observed rating given by rater *r* for test taker *i*'s response to item *j*. The model is estimated as a Bayesian model with Markov chain Monte Carlo (MCMC) estimation. Therefore, the model depiction in (1) assumes prior distributions for unknown parameters in the model including the precision of the latent traits ω, the difficulty parameters β_{jk} and step parameters γ_{jk} of the IRT model, and the rater parameters, $1/\tau^2$ and ϕ, which are rater precision and bias (severity/leniency), respectively. We will discuss this in more detail later in the next section which focuses on a simulated example.

The ideal ratings, ξ_{ij}, represent the quality of person *i*'s response to item *j* and are latent variables modeled using a polytomous IRT model, such as the *K*-category partial credit model (PCM; Masters, 1982). With ideal rating ξ_{ij} and *K* possible scores $(k = 1,\ldots,K)$, the PCM is given by:

$$P\left(\xi_{ij} = \xi \middle| \theta_i, \beta_j, \gamma_{j\xi}\right) = \frac{\exp\left\{\sum_{k=1}^{\xi}\left(\theta_i - \beta_j\right) - \gamma_{jk}\right\}}{\sum_{h=0}^{K-1}\exp\left\{\sum_{k=1}^{h}\left(\theta_i - \beta_j\right) - \gamma_{jk}\right\}}. \tag{2}$$

From the PCM component of the HRM we estimate β_j, the item difficulty for the j^{th} item, γ_{jk}, the k^{th} item step parameter for item *j*, and the latent traits, θ_i.

Table 1:
Matrix of Rating Probabilities in the SDM Component of the HRM

Ideal Rating (ξ)	Observed Rating (*k*)				
	0	1	2	3	4
0	p_{00r}	p_{01r}	p_{02r}	p_{03r}	p_{04r}
1	p_{10r}	p_{11r}	p_{12r}	p_{13r}	p_{14r}
2	p_{20r}	p_{21r}	p_{22r}	p_{23r}	p_{24r}
3	p_{30r}	p_{31r}	p_{32r}	p_{33r}	p_{34r}
4	p_{40r}	p_{41r}	p_{42r}	p_{43r}	p_{44r}

The signal detection model (SDM) in the HRM follows a matrix of rating probabilities (see Table 1). In this matrix, the probabilities are conditional on the ideal rating. For example, p_{00r} is the probability that rater *r* assigns a score of 0 when the ideal rating is 0. Thus, a separate table describes each rater's rating probabilities conditional on ideal ratings. To model patterns of rating behavior per rater, the SDM in the HRM considers the probabilities in each row of the matrix to be proportional to a Normal density with mean $\xi + \phi_r$ and standard deviation τ_r :

$$p_{\xi kr} = P\left(X_{ijr} = k \middle| \xi_{ij} = \xi\right) \propto \exp\left\{-\frac{1}{2\tau_r^2}\left[k - (\xi + \phi_r)\right]^2\right\}. \qquad (3)$$

The rater bias parameter, ϕ_r , indicates a rater's deviation from the ideal rating and reflects a consistent bias in the rater's ratings. When ϕ_r approaches 0, the rater has only a small deviation from the ideal rating. When ϕ_r is negative, the rater exhibits a severity effect (or negative bias). Conversely, when ϕ_r is positive the rater exhibits a leniency effect (or positive bias). Typically, values smaller than 0.5 in absolute value are not considered substantial because they lie within 1 score point from the ideal rating. Values beyond 0.5 in absolute value, however, indicate a tendency to score a full score point or more away from the ideal (Casabianca, Junker, & Patz, 2016; Patz et al., 2002). The spread parameter, τ_r , indicates a rater's variability around $\xi + \phi_r$; values near 0 indicate high consistency or reliability in rating and high values indicate poorer consistency in rating. It is important to note that this parameter is interpreted in relation to a rater's ϕ_r . If ϕ_r is 0 and τ_r is small (< 0.5) then the rater consistently scores with no bias; their probability of scoring in categories above or below the ideal rating category is low. If ϕ_r is 0 and τ_r is larger (> 0.5) then the rater scores inconsistently around 0 bias. If ϕ_r is 1.25, for example, and τ_r is small (< 0.5), then the rater has consistent positive bias. Finally, if ϕ_r is 1.25 and τ_r is large (for example, $\tau_r = 1$), then the rater has positive bias but with a lot of variation, or inconsistency. The larger τ_r , the more errors a rater will make relative to their own central tendency. In other words, the HRM captures rater inconsistency around the rater's typical scoring behavior. As we mentioned earlier, this is different from the traditional definition of accuracy/inaccuracy (and the notion of random errors or variation around known-to-be-valid ratings), however, if the rater was fairly unbiased, then τ_r could be a measure of inaccuracy/accuracy.

Values of τ_r greater than 0.5 indicate that a rater is scoring roughly consistently around $\xi + \phi_r$ and values greater than this indicate raters are scoring with more variability and will perhaps assign ratings at the next score level (above or below). Based on this reasoning, for this study we decided to consider values greater than approximately 0.75 to be larger than desired, and certainly values larger than 1.0 to be large and a sign of rater unreliability. This criterion was selected in relation to the length of the score scale. If the scale were longer, for example, if $K = 9$, then perhaps we would use a less stringent criterion for classifying a rater as unreliable.

Demonstrative example

Using simulated datasets we investigated our five research questions: (i) What is the impact of *rater pool quality* on measurement accuracy? (ii) What is the impact of *multiple ratings per item response* on measurement accuracy? (iii) What is the impact of *test length* on measurement accuracy? (iv) When considering the measurement of individuals, how robust are scores at different levels (item, total, pass/fail) to the impact of these aforementioned design decisions? and (v) How does the impact of these design decisions differ when measuring individuals versus measuring groups of individuals? To answer these questions, we generated ratings data from the HRM for 1,000 test takers (N = 1,000), for constructed response items each on a 5-point scale ($K = 5$) scored by $R = 100$ raters. We selected this number of examinees and raters because this would be roughly the scenario in an administration of a large-scale assessment and it has also been used in other related simulation studies.

To address research question (i), we varied the quality of the rater pools in the simulated datasets. Specifically, we varied the type of rater pool quality (100% Normal, 20% Unreliable, 20% Severe, which were defined by manipulating HRM rater parameters as explained below). Unreliability and severity are two rater effects that impact ratings, and they are the two modeled by the HRM. We examined a sample with severity and another one with unreliability, but not one with a combination of the two; we decided to manipulate rater effects in isolation to be able to explain the results with clarity. The 20% prevalence of raters in the pool exhibiting the effect is similar to other studies. To address (ii) we varied the number of ratings per item (S = 2, 4, 8). Two ratings per item is a common rating design ("100% double-scored"). Having 4 or 8 ratings per item is not common, but we included those levels in order to demonstrate what occurs as we add ratings. To address (iii), we varied the number of items (J = 2, 4, 8) and selected 2, 4, and 8 item tests because typically a test made up of CR items is not very long due to the efforts required to respond to it and score it. Two-item CR tests may consist of two essays, for example, such as in TOEFL Writing where there are two separate writing tasks. These factors were completely crossed to yield 27 (3*3*3) datasets to which the HRM was fitted and resulting parameter estimates evaluated for measurement accuracy. Note that this is an analysis of simulated datasets and that we analyzed only one simulated dataset per condition.

Research questions (iv) and (v) were answered by analyzing the observed scores and estimated traits at different score levels (ratings level, composite level, pass/fail) and for individuals versus groups.

Ratings data generation

To generate observed ratings from the HRM we used three sets of rater parameters, and Figure 2 graphically illustrates those three sets of parameters. The first set of rater parameters contained 100 raters with bias and variability within normal ranges (i.e., "normal" raters). We define normal raters, or raters who do not exhibit aberrant behavior, to have bias between $-0.5 \leq \phi_r \leq 0.5$ and variability $\tau_r \leq 0.75$. Ratings data generated with these raters will have minimal noise and thus will not greatly impact measure estimates for test takers. The second set of rater parameters contained 80 raters with "normal" parameters and 20 raters with normal bias parameters but relatively large rater SDs, which simulates a situation in which 20% of raters exhibit an unreliablity rater effect. Note that in these cases, the raters all had absolute rater bias values less than 0.5 (i.e., none exhibited a severity or leniency rater effect). The third set of rater parameters contained 80 raters with "normal" parameters and 20 raters with larger negative bias parameter values which simulates a situation in which 20% of raters exhibit a severity rater effect but no unreliability effect.

We randomly drew rater parameter values from the distributions as listed in Table 2 and kept those true rater parameter values fixed across the conditions varying the number of items and ratings within the rater quality condition. The 1,000 true latent trait values

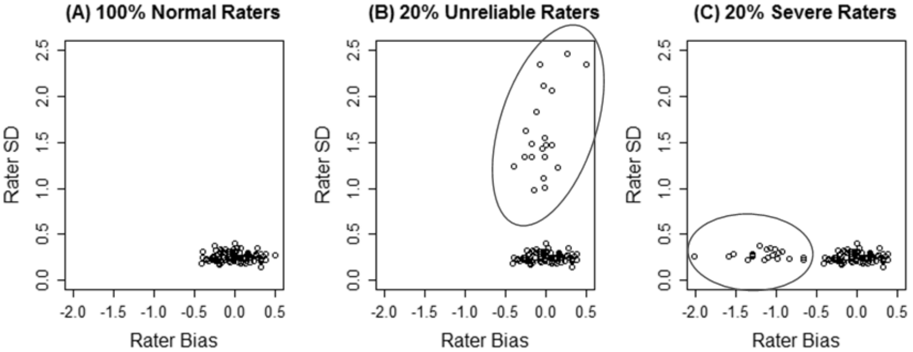

Figure 2:
True rater parameter values for each Rater Quality condition. The average rater bias was -0.007 for the Normal and Unreliable conditions and -0.232 for the Severe condition. The average rater SD was 0.255 for the Normal and Severe conditions and 0.518 for the Unreliable condition. The circled points are aberrant raters. (Note: SD = standard deviation).

Table 2:
Generating Rater Parameter Distributions for Simulated Data

Rater Quality Condition	Normal Raters		Aberrant Raters	
	Rater Bias	Rater SD	Rater Bias	Rater SD
100% Normal	$\phi \sim$ N(0,0.25)	Log(τ) ~ N(0.50, 0.13)	N/A	N/A
20% Unreliable	$\phi \sim$ N(0,0.25)	Log(τ) ~ N(0.50, 0.13)	$\phi \sim$ N(0,0.25)	Log(τ) ~ N(1.2, 0.14)
20% Severe	$\phi \sim$ N(0,0.25)	Log(τ) ~ N(0.50, 0.13)	$\phi \sim$ N(-1, 0.30)	Log(τ) ~ N(0.50, 0.13)

Note. SD = standard deviation; ϕ = rater bias; τ = rater SD.

were drawn from a $N(0,1)$ distribution and were kept fixed across all datasets. Each test taker’s true pass/fail status was determined by their true θ value – if θ > 0 then they passed, otherwise they failed. We selected PCM parameters for eight items from the literature (Donoghue, 1994; Li & Baser, 2012) to use in our simulations and nested the 2-item test within the 4-item test which were nested within the full 8-item test.

Using the true latent trait values, true PCM item parameters, and true rater parameters, we generated observed ratings for a fully-crossed rating design in which all 100 raters scored all items. To manipulate the number of ratings per item, we trimmed the fully-crossed data by randomly selecting either 2, 4, or 8 ratings per item. This generated incompletely-crossed datasets in which each response has only 2, 4 or 8 ratings (instead of 100 ratings, 1 rating from each of 100 raters). Since the ratings were removed randomly from the fully-crossed dataset, there is no relationship between the response (or test taker who gave the response) and the rater who assigned the rating to the response. Furthermore, the number of ratings per response is uniform within a dataset. What did vary is the number of ratings each rater assigned within a dataset.

We evaluated the utility of scores at different levels to reflect a test taker’s true ability. Thus we evaluated correlations between true θ values and item scores, HRM-based θ estimates, and pass/fail classifications derived from HRM-based θ estimates. A test taker was designated to have passed if that test taker’s estimated θ was greater than 0 and designated to have failed otherwise. We examined results at the individual test taker and the group level.

Parameter estimation

The HRM is a hierarchical Bayesian model with parameters estimated with MCMC methods which means that we must place priors on all estimated parameters (for more information on MCMC estimation in the IRT context and more specifically in the HRM context, see references such as: Patz & Junker, 1999ab; Patz. et al., 2002; Junker, Patz, &

Van Houdnos, 2016). On the rater bias parameters ϕ_r we placed a *N*(0, σ_ϕ^2 =10) and on the rater precision parameters $1/\tau_r^2$ we placed a *Gamma*(1,1). We placed a *N*(0,1) prior on the latent traits, item difficulties, and item step parameters.[4] Using weakly informative priors such as these (versus uninformative priors, e.g., *N*(0,10), which provide less certainty about the likely values), contributes to efficient estimation but still provide enough flexibility for precise estimates. In general, using a truly informative prior may be too restrictive, and using an uninformative prior would allow the data to dominate the estimation procedure. However, it is unnecessary to fully rely on the data since we know the likely range of values for all parameters.

We used JAGS (Plummer, 2003) via the R2Jags package from R (Su & Yajima, 2012) to fit the HRM to the data. For each dataset we ran 3 chains, each with 20,000 iterations, and a burn-in of 10,000. To reduce autocorrelation we thinned the chains by keeping every 10th iteration resulting in 3,000 iterations in the final posterior sample (3 chains x 1,000 iterations). We evaluated the convergence of chains according to the Gelman-Rubin convergence diagnostic ($\hat{R}$; Gelman & Rubin, 1996); all $\hat{R}$ values were below 1.1 which is the criterion indicative of convergence.

Results: Measurement accuracy at the individual level

Presentation of our results focuses on how the decisions that we make about test takers based on the HRM estimates are impacted by four design decisions: (a) rater pool quality (i.e., rater selection), (b) the number of ratings per response, (c) the number of items on the assessment, and (d) the level at which we interpret test taker measures. We also discuss the impact of utilizing the HRM. In this section, we focus exclusively on interpretation at the levels of the test taker. We focus on group-level decisions in the following section.

Table 3 presents correlations that illustrate the impact of design decisions regarding the number of items (2, 4, or 8), the quality of the rating pool (Normal, Unreliable, and Severe), the number of ratings for each response (2, 4, or 8), and the level at which test taker measures are interpreted (Item, Total, and Pass/Fail classifications), on concordance with true latent ability. Specifically, this table contains the correlations between generating latent trait values and observed/estimated parameter values. Item level correlations are between the true generating θs and the observed item scores. Item scores were comput-

[4] Note that, typically, identification of the PCM in the Bayesian context would entail constraining the location of the scale either by setting $\mu = 0$ in the Normal prior on the latent traits *or* constraining the item difficulty parameters using either a hard or soft constraint (using priors). In addition, there is another location indeterminacy in the item step parameters. Thus, for purposes of identification, we did use a *N*(0,1) prior on the latent traits and we applied a sum-to-zero constraint on the item step parameters. We also used the same prior on both the difficulties and item step parameters since in some of these datasets, especially with the two-item test, there is not a lot of information/data to estimate all of the HRM parameters.

Table 3:
Correlations between True and Observed / Estimated Measures at the Individual-level

Number of Items	Rating Quality	Number of Ratings	Score Levels			
			Observed Scores		HRM-based Scores	
			Item	Total	Total	P / F
2	Normal	2	.685	.797	.797	.835
		4	.686	.795	.796	.836
		8	.686	.795	.796	.836
	Unreliable	2	.666	.783	.802	.827
		4	.676	.791	.807	.833
		8	.684	.799	.806	.834
	Severe	2	.672	.788	.799	.834
		4	.681	.794	.800	.833
		8	.686	.799	.800	.833
4	Normal	2	.672	.876	.877	.921
		4	.672	.875	.878	.922
		8	.676	.878	.878	.922
	Unreliable	2	.634	.856	.875	.879
		4	.654	.868	.880	.882
		8	.666	.875	.881	.881
	Severe	2	.640	.859	.875	.892
		4	.653	.864	.876	.883
		8	.661	.869	.876	.884
8	Normal	2	.665	.924	.927	.939
		4	.666	.924	.927	.948
		8	.666	.924	.927	.942
	Unreliable	2	.634	.916	.928	.946
		4	.656	.923	.930	.947
		8	.664	.925	.930	.947
	Severe	2	.639	.920	.928	.943
		4	.648	.921	.928	.941
		8	.654	.924	.928	.943

Note. The item score correlations are based on the observed item scores (average of multiple ratings for an item) and the true latent trait values. The first total score column contains correlations between true latent traits and observed total test scores. The second total score column contains correlations between true and estimated latent traits. The pass/fail correlations are based on the classifications based on the estimated and true latent trait values. The pass/fail correlations are biserial correlations. HRM = hierarchical rater model; P/F = pass/fail.

ed as the average of the observed ratings for an item for each test taker, or $\sum_{r=1}^{D} X_{ijr} \Big/ D$ where D is the number of ratings of test taker i's response to item j (in this study, $D = 2$, 4, or 8). The observed item scores are values that reflect a test taker's item-level performance, taking into account all raters' ratings. Because each test taker responds to at least two items, we computed correlations (one per item) separately and then averaged the correlations across items within a condition to prevent dependencies from artificially inflating the correlations. For example, in the conditions with two items and two ratings per item, we computed the correlation between the true θ and the item score for item 1 (average of the two ratings) and the correlation between the true θ and the item score for item 2. We applied a Fisher transformation (Fisher, 1915) to each correlation and computed the mean correlation on the *z* scale. We then transformed the mean back to the correlation scale. Those values appear in the column labeled "Item" in Table 3. There are two "Total" columns in this table that describe the relationship between the true θ value and scores summarizing total test performance. The "Total" column under "Observed Scores" contains Pearson correlations between the generating θ values and the observed total test score, computed as the sum of the item scores. The "Total" column under "HRM-based Scores" contains Pearson correlations between the generating θ values and the estimated θ values. We provided both to demonstrate the difference between explicitly modeling rater severity and unreliability in the latent trait context versus the observed score framework. The "P / F" column (where "P / F" is for Pass/Fail) contains biserial correlations describing the relationship between the pass or fail classification based on the estimated θs and the true θ values. Thus, all correlations in Table 3 describe the relationship between scores observed at different levels and the true underlying ability.

Impact of rating quality & effect of utilizing HRM

We manipulated rating quality by generating ratings based on a rating pool with or without aberrant raters. The HRM is a model that explicitly estimates rater severity and unreliability and thus controls for these effects in the resulting latent trait estimates. For this reason, we would not expect to see a large impact due to rating pool quality on the latent trait estimates (i.e., HRM-based total scores). In other words, we expect to see similar correlations across the rating quality conditions because the model corrects for those effects. Indeed, this is the case. That is, for the HRM-based total scores, when comparing the correlations across the rating quality conditions (with the same number of items and the same number of ratings), the differences in the correlations were generally very small (< .011).

The similarities in correlations observed across rating quality conditions under the HRM-based total score correlations are not as apparent in the observed score correlations, indicating the usefulness of the HRM. Comparing the two total score correlations, we see essentially no differences in correlations in the Normal conditions – that is, the correlations based on observed total test scores and estimated latent traits are the same when the rater pool contained only Normal raters. However, in the Unreliable and Severe conditions, the correlations between the observed total test score and the generating latent trait values were weaker. The Unreliable conditions, matching the number of items and raters, had correlations approximately .005 to .019 lower when correlating the observed score with the true latent trait values. The Severe conditions, also matching the other factors in the study, had correlations approximately .001 to .016 lower. These differences across score types (observed vs. estimated) relate to differences in the scoring mechanisms' capacities to reflect true abilities under less than optimal rating quality conditions; though they are relatively small (maximum difference of .02) they are potentially impactful in high-stakes situations.

Overall, across the four score levels, comparisons between Normal, Unreliable, and Severe rating quality conditions with the same number of items and ratings reveal a largest absolute difference of only .042 (comparing the P/F correlations for Normal rating quality for 4 items and 2 ratings, .921, to the correlation for Unreliable rating quality for the same number of items and ratings, .879). This suggests that at least in this study, with these data and under these modeling strategies, the existence of rater effects only has a small impact on the accuracy of the parameter estimates. It is worth noting that the larger differences tend to appear when comparing Normal rating quality to Unreliable and Severe rating quality and that larger differences tend to appear when making comparisons at the P/F score level and, to some degree, at the item score level. In addition, most of the largest differences occur when there are 4 items on the test.

Impact of number of ratings

The results indicate no significant improvements in agreement with the true latent trait due to the inclusion of more ratings, a result that is consistent with prior research conducted within a generalizability theory framework (for example, see Brennan, Gao, & Colton, 1995). That is, if you compare the values of the correlations when the rating quality and number of items is the same within each score level column, the differences are generally very close to 0. A few exceptions to this trend exist at the Item score level, but the absolute differences are about .03. Most notably, in the Unreliable conditions at the Item score level, the greatest improvements in agreement with the true latent traits occur for 4 and 8 items when increasing the number of ratings from 2 to 8 (e.g., for 4 items, 2 ratings produces a correlation of .634 while 8 ratings produces a correlation of .666). There are also some differences in the correlations between the total observed score and the true latent traits, mainly in the Unreliable and Severe rating quality conditions. For example, the same condition that exhibited sensitivity to the number of ratings based on item scores also revealed the same sensitivity based on observed total test scores, but slightly less so. The difference between the correlation in the 4-item, Unreliable condition with 2 ratings and 8 ratings was .019. The corresponding difference in correlations based on the true and estimated θs was .006. Importantly, these findings are valid only within the context of the conditions in this study, namely, two, four and eight ratings per response.

Impact of the number of items

The results associated with comparisons between rows that contain the same rater effect and the same number of ratings but different numbers of items reveals what one would expect – more items results in better recovery of the test taker’s true ability, but only at the Total and P/F score levels. Generally, the correlations for the Item score level were around the average value of .67. However, at the Total and P/F score levels, the typical correlations were roughly .80 to .83 for 2 items to about .88 to .90 for 4 items to about .93 to .94 for 8 items. That is, at these two levels, correlations increased by a value of more than .10, on average, when comparing a 2 item test to an 8 item test.

Impact of score level decision

The results for the score level reveal another expected result – as we decrease the granularity of the score we interpret (Item --> Total --> P/F), the score's representation of underlying ability is improved. Specifically, the average correlation at the Item score level average was about .67, while the average correlations at the Total score and P/F levels equal .86 and .89, respectively. That is, when you make decisions based on a broader scope, your depictions of the test taker’s ability are more accurate.

Results: Measurement accuracy at the group level

Yet another level of analysis is at the group level. Table 4 summarizes statistics that demonstrate the impact of focusing on a group of test takers rather than individual test takers. Specifically, Table 4 provides the mean of the ability estimates $M\left(\hat{\theta}\right)$, the mean deviation of ability estimates from true abilities $M\left(\theta-\hat{\theta}\right)$, the mean of the standard errors of the ability estimates $M\left[SE\left(\hat{\theta}\right)\right]$, and the standard error of the mean of the ability estimates $SE\left[M\left(\hat{\theta}\right)\right]$. Generally, the mean of the ability estimates and the biases are all close to the expected value of 0.00, and they do not vary much across any of the three factors that we varied, although the values of the mean of the estimates do tend to approach zero more consistently when there are 4 or 8 items rather than only 2.

Two comparisons that are revealing have to do with the mean of the standard errors of the estimates and the standard error of the mean of the estimates. First, note that the mean of the standard errors of the estimates $M\left[SE\left(\hat{\theta}\right)\right]$ decrease as the number of items increase. That is, the mean for 2 items equals 0.61, the mean for 4 items equals 0.49, and the mean for 8 items equals 0.37. This reinforces the observation made in the previous section that the only design decision that has a significant impact on our results is the number of items

Table 4:
Summary Statistics for Estimated Traits Describing Measurement Accuracy at the Group Level

Number of Items	Rating Quality	Number of Ratings	$M(\hat{\theta})$	$M(\theta - \hat{\theta})$	$M[SE(\hat{\theta})]$	$SE[(M(\hat{\theta})]$
2	Normal	2	-0.004	-0.002	0.610	0.025
		4	-0.003	-0.003	0.606	0.025
		8	-0.002	-0.004	0.607	0.025
	Unreliable	2	-0.002	-0.004	0.618	0.024
		4	-0.003	-0.003	0.605	0.025
		8	-0.002	-0.004	0.605	0.025
	Severe	2	-0.004	-0.002	0.612	0.025
		4	-0.003	-0.003	0.606	0.025
		8	-0.003	-0.003	0.606	0.025
4	Normal	2	-0.001	-0.005	0.487	0.027
		4	-0.001	-0.005	0.485	0.027
		8	-0.002	-0.004	0.485	0.027
	Unreliable	2	0.001	-0.007	0.497	0.027
		4	0.001	-0.007	0.487	0.027
		8	0.001	-0.007	0.487	0.027
	Severe	2	0.000	-0.006	0.493	0.027
		4	0.000	-0.006	0.489	0.027
		8	-0.001	-0.005	0.489	0.027
8	Normal	2	0.000	-0.006	0.374	0.029
		4	-0.001	-0.005	0.372	0.029
		8	-0.001	-0.005	0.373	0.029
	Unreliable	2	-0.001	-0.005	0.379	0.029
		4	0.000	-0.006	0.372	0.030
		8	-0.001	-0.005	0.372	0.030
	Severe	2	-0.001	-0.005	0.371	0.028
		4	-0.001	-0.005	0.369	0.028
		8	0.001	-0.007	0.369	0.028

Note. Means and mean deviations of estimated θs (from true θs) reveal no differences across conditions. An effect due to the level of decision making is shown when comparing the mean of the SEs to the SE of the mean, indicating higher precision at the group level. θ = true latent trait; $\hat{\theta}$ = estimated latent trait; *M* = mean; *SE* = standard error.

administered to test takers. As one might expect, not only does an increase in the number of items improve the correlation between estimates and true abilities (Table 3), but the estimates are more precise with the increase in the number of items (Table 4). Note that rating quality and the number of ratings assigned by raters have no impact on the precision of the estimates under the HRM. Second, comparing the two rightmost columns reveals that, when decisions are made at the group level rather than the individual level, the precision of the statistic in question is higher. On average, the standard error of the mean $SE\left[M\left(\hat{\theta}\right)\right]$ equals 0.03, and its value does not vary significantly across any of the conditions that we explored. Note that $SE\left[M\left(\hat{\theta}\right)\right]$ does get marginally smaller when decreasing the number of items, however this relates to an underestimation of the true variance of the latent trait distribution and should not be confused with improved precision.

Discussion & Conclusions

The design of constructed response scoring systems requires many choices that inherently impact the quality of the collected ratings, resulting scores, and decisions made about test takers. We examined how rater selection (rating quality), the number of items, and the number of ratings per item per test taker impact scores at different levels under IRT scoring with the HRM. There are several levels of evaluation possibly resulting from an assessment. Diagnostic information about an individual may be gleaned from a single rating on a single response to an item. Summative information about an individual's overall ability on some domain may be gathered from a total score based on multiple items. Information on whether or not a candidate achieves a sufficient score to show mastery on a domain may be gathered by applying an established cutpoint to a total score. Furthermore, these three score levels may be considered at the individual test taker level, or at the group level when performing item or test analyses. Our goal in this article was to study how different design decisions impact the information about individuals and groups at these various score levels.

The HRM is a multilevel IRT model for ratings which explicitly models and therefore accounts for rater bias (severity and leniency) and rater variability. Scores computed from the HRM are latent trait estimates that have been refined to account for the rater bias and unreliability detected by the model. In addition, the multilevel component of the HRM includes a nesting of observed ratings assigned by human raters within ideal ratings. By including this hierarchy, the model acknowledges that multiple ratings of the same work should not add information to the measurement (only additional items should contribute to the test information).

Our results demonstrate six things about the impact of the measurement model and design decisions on the accuracy of score interpretation. First, we demonstrated that employing the HRM improves the accuracy of score interpretation when compared to interpretation of observed scores. The HRM accomplished this by removing the influence of rater severity and unreliability on test taker measures. In the simulation, we observed that the correlation between observed scores and true ability is lower than the correlation between HRM estimates and true ability, indicating that HRM estimates more closely approximate truth. Second, we demonstrated that rater severity and unreliability as modeled by the HRM have only a small impact on the accuracy of test taker measures. In our simulations, the maximum difference in correlation between true and estimated abilities was about .04 when comparing normal to unreliable raters, showing that the model did a good job of recovering truth, even when observed ratings contained more error. Third, we demonstrated that the number of items on the test has the biggest impact on the accuracy of test taker measures. In our example, the correlations between true and estimated abilities were different by more than .10 when comparing a 2 item test to an 8 item test. Fourth, we demonstrated that the number of ratings per response whether there were two, four, or eight ratings per response, had virtually no impact on the accuracy of test taker measures. Fifth, our simulations demonstrated how the granularity of decisions impacts the accuracy of score interpretation. For example, focusing on the test taker's scores on a single item resulted in correlations of about .67 between estimated and true ability while focusing on pass/fail decisions resulted in a correlation of about .87. Sixth, focusing on the group of test takers, rather than individual test takers, also produced more accurate decisions. For example, the average standard error of latent trait estimates equals 0.49, while the average standard error of the mean of the latent trait estimates equals 0.03.

Our conclusions from this study may be summarized into the following points:

- Rater selection is important for ensuring quality scores. Raters with experience, who are reactive to feedback, and who have shown minimal aberrant behavior are preferred for selection. However, given the potential limitations involved with hiring, training, and selecting raters, using the HRM for scoring provides a method by which to mitigate some of their errors. Even still, the HRM does not mitigate all rater effects. Generally, there are consequences to ignoring rater effects when using observed scores or scoring with IRT models that do not model these effects (Hombo, Donoghue, & Thayer, 2001).
- Collecting multiple ratings of the same work is a design component that may appear to provide better measurement, however, the payoff is not great, if any payoff exists at all. Under the HRM, there were no notable differences related to the number of ratings at the individual or group level, at least not under the conditions we studied.
- Test length is very important in measurement at the individual level. Our simulations supported what was already widely known about the advantages of longer tests. There is no benefit to a longer test when considering measures of this type at the group level.

Our results are based on a series of simulated datasets, and thus multiple replications would be the best way to derive more stable conclusions. However, combining our knowledge of psychometrics and our observations from this simulated example we are still able to make some high-level suggestions. For the design of CR assessments and scoring systems we suggest using multiple items whenever possible and applying a scoring model that accounts for the specific rater effects that have been found in ratings.

While the costs associated with multiple items are likely higher than costs associated with multiple ratings of fewer items because there is less training involved, the psychometric payoff is likely to be greater. It is important to note that not all IRT rater models will be appropriate in all situations. The ratings collected within the scoring system may reveal different rater effects and thus preliminary descriptive analyses may be helpful in determining which model is most appropriate to be applied to mitigate those errors.

References

Brennan, R. L. (2001). *Generalizability theory*. New York, NY: Springer.

Brennan, R. L., Gao, X., & Colton, D. A. (1995). Generalizability analyses of Work Keys listening and writing tests. *Educational and Psychological Measurement*, *55*(2), 157-176. doi: 10.1177/0013164495055002001

Brown, W. (1910). Some experimental results in the correlation of mental abilities. *British Journal of Psychology*, *3*(3), 296-322.

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

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*(3), 333-356. doi:10.1111/j.1745-3984.2011.00143.x

Donoghue, J. R. (1994). An empirical examination of the IRT information of polytomously scored reading items under the generalized partial credit model. *Journal of Educational Measurement*, *31*(4), 295-311.

Fisher, R. A. (1915). Frequency distribution of the values of the correlation coefficient in samples from an indefinitely large population. *Biometrika*, *10*(4), 507-521. doi:10.2307/2331838

Gelman, A., & Rubin, D. B. (1996). Markov chain Monte Carlo methods in biostatistics, *Statistical Methods in Medical Research*, *5*(4), 339-355. doi:10.1177/096228029600500402

Hombo, C. M., Donoghue, J. R., & Thayer, D. T. (2001). *A simulation study of the effect of rater designs on ability estimation* (Research Report No. RR-01-05). Princeton, NJ: Educational Testing Service.

Hoskens, M., & Wilson, M. (2001). Real-time feedback on rater drift in constructed-response items: An example from the Golden State Examination. *Journal of Educational Measurement*, *38*(2), 121-145. doi: 10.1111/j.1745-3984.2001.tb01119.x

Junker, B. W., Patz, R. J., & VanHoudnos, N. M. (2016). Markov chain Monte Carlo for item response models. In W. J. van der Linden (Ed.), *Handbook of modern item response theory* (pp. 271-312). Boca Raton, FL: Chapman & Hall/CRC.

Kim, S. C., & Wilson, M. (2009). A comparative analysis of the ratings in performance assessment using generalizability theory and the many-facet Rasch model. *Journal of Applied Measurement*, *10*(4), 403-423.

Li, Y., & Baser, R. (2012). Using R and WinBUGS to fit a generalized partial credit model for developing and evaluating patient-reported outcomes assessments. *Statistics in Medicine*, *31*(18), 2010-2026. doi: 10.1002/sim.4475

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

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

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika*, *47*(2), 149-174. doi:10.1007/BF02296272

Patz, R. J., & Junker, B. W. (1999a). A straightforward approach to Markov chain Monte Carlo methods for item response models. *Journal of Educational and Behavioral Statistics*, *24*(2), 146-178.

Patz, R. J., & Junker, B. W. (1999b). Applications and extensions of MCMC in IRT: Multiple item types, missing data, and rated responses. *Journal of Educational and Behavioral Statistics*, *24*(4), 342-366.

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. doi:10.3102/107699860 27004341

Plummer, M. (2003). JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling. In K. Hornik, F. Leisch, & A. Zeileis (Eds.), *Proceedings of the 3rd International Workshop on Distributed Statistical Computing (DSC 2003)* (pp. 1-10). Retrieved from https://www.R-project.org/conferences/DSC-2003/Proceedings/Plummer.pdf

Spearman, C. (1910). Correlation calculated from faulty data. *British Journal of Psychology*, *3*(3), 271-295. doi: 10.1111/j.2044-8295.1910.tb00206.x

Su, Y. S., & Yajima, M. (2012). R2jags: A Package for Running jags from R. *R Package Version 0.03-08.*

Verhelst, N. D., & Verstralen, H. H. (2001). An IRT model for multiple raters. In Boomsma, A., van Duijn, M., & T. Snijders (Eds.), *Essays on item response theory* (pp. 89-108). New York, NY: Springer.

Wilson, M., & Hoskens, M. (2001). The rater bundle model. *Journal of Educational and Behavioral Statistics*, *26*(3), 283-306. doi:10.3102/10769986026003283

Wolfe, E.W. (2014). *Methods for monitoring rating quality: Current practices and suggested changes*. Iowa City: Pearson.