A peer reviewed, open-access electronic journal: ISSN 1531-7714

Systematic Comparison of Two Approaches for Evaluating and Using Rater-Mediated Performance Assessments

Chunling Niu, *University of the Incarnate Word*
Kelly Bradley, *Marshall University*
Rui Jin, *Shenzhen University*
Ashley Love, *Southern New Hampshire University*

Abstract: Rater-mediated performance assessments (RMPAs) involve third-party raters evaluating individual performance and are increasingly used across educational, organizational, and research contexts. However, challenges persist in accounting for rater bias and measurement errors, as well as addressing concerns around equity and fairness, especially for historically marginalized populations. This paper addresses these challenges by first discussing the methodological limitations of widely used RMPA evaluation techniques based on classical test theory (CTT), including factor analysis, Cronbach's alpha, and interrater reliability analysis. An alternative approach using Many-Facet Rasch Modeling (MFRM) is then introduced. The two frameworks are systematically compared from both theoretical and empirical perspectives. An empirical example using AI safety evaluation data from the DICES dataset demonstrates how MFRM yields enhanced diagnostic insights (including rater severity differences, rating scale functioning issues, and construct dimensionality) that CTT approaches may not readily provide. Finally, commonly used MFRM-based analytical techniques are introduced for typical RMPA evaluation studies. This paper not only aims to enhance the methodological rigor of RMPAs but also seeks to contribute to the ongoing dialogues on creating more equitable and fair performance assessment practices.

Keywords: Rater-mediated performance assessments, Classical test theory, Many-Facet Rasch Modeling, Methodological comparison

Introduction

Rater-mediated performance assessments (RMPAs) involve systematic evaluation of an individual's performance or capabilities by trained observers or judges using standardized criteria and rubrics. These assessments capture complex behaviors, skills, and competencies that may be difficult to measure through traditional testing formats. The raters observe and document performance according to established protocols, generating detailed evaluative data that can inform important decisions across educational, professional, and research contexts. While self-report measures and automated scoring systems also involve human judgment (through item design and algorithm development, respectively), RMPAs directly incorporate real-time human observation and evaluation of performance. This approach allows trained raters to apply nuanced, context-sensitive judgments within standardized evaluation frameworks. This approach has gained increasing prominence due to its ability to generate rich, detailed evidence about complex performances that may be difficult to capture through other assessment methods. RMPAs have been used in a variety of contexts, including educational, organizational, and health care research. (Borman et al., 2003; Leung et al., 2008).

The past three decades have seen a substantial growth in the use of RMPAs in survey research (Darling-Hammond et al., 2010; Madaus et al, 1999; Modell, 2004). This is primarily because RMPAs typically involve an external observer who can provide a relatively objective/unbiased evaluation of the participant's performance (Goh, 2012). RMPAs are particularly useful when measuring complex behaviors that are difficult to self-report, such as interpersonal skills or job performance (Latham & Wexley, 1993). Additionally, in contexts requiring direct observation of complex behaviors, such as interpersonal communication or procedural skills, RMPAs can capture performance details that may be difficult to assess through self-report alone, as they allow trained observers to evaluate multiple aspects of a behavior or performance simultaneously (Knoch et al., 2021).

Among various types of application in research and practice, RMPAs are increasingly being used to provide feedback to employees and to measure their progress in professional development programs designed for teachers (Reagan et al., 2016), nurses (Robb & Dietert, 2002), and for other professionals, such as police officers (Bertilsson et al., 2020). Moreover, RMPAs have also been adopted as a type of self- and/or peer-assessments. For example, abundant empirical findings showed that the use of a RMPA improved the accuracy of self-assessment and peer-assessment scores in a sample of university students (Farrokhi et al., 2011; Han, 2018). These findings suggest that RMPAs can be a valuable tool for improving the accuracy and reliability of large-scale self- and peer-assessments in academic settings.

However, RMPAs face several inherent methodological challenges that affect measurement quality. These include: (a) ensuring rating accuracy due to various rater effects and biases that can influence scoring; (b) establishing reliable and valid rating criteria, which requires extensive rater training; (c) maintaining consistent interpretations across different raters who may view criteria differently; (d) developing rating scales that function consistently both across multiple raters and within individual raters' use; and (e) sustaining rating reliability over time as raters' familiarity with criteria and ratees changes (Guo, 2021; Wind, 2019). Traditional CTT-based approaches attempt to evaluate these challenges through various statistical indices, but as we will demonstrate, they have important limitations in addressing these fundamental measurement issues.

Equally important in the discourse on RMPAs is the consideration of equity and fairness, particularly concerning historically marginalized populations. The potential for rateeive biases in RMPAs - stemming from cultural, racial, or socioeconomic factors - raises significant concerns regarding the fairness of assessments. These biases can lead to disparate impacts on marginalized individuals, influencing their feedback, opportunities for development, and overall outcomes in professional or educational advancement. Acknowledging the importance of addressing these biases, researchers have begun to explore methodologies

that ensure more equitable assessment practices. Strategies such as diversifying rater pools and enhancing cultural competency training for raters are among the approaches being discussed to mitigate bias in RMPAs (Montgomery & Fernandez, 2019). Furthermore, the integration of Many-Facet Rasch Modeling (MFRM) offers a promising avenue for analyzing and correcting for potential biases, providing a more equitable assessment framework (Linacre, 2018).

To address these challenges, two major measurement frameworks have been used to examine the psychometric properties of RMPAs: Classical Test Theory (CTT) and Many-Facet Rasch Modeling (MFRM). While both frameworks can provide evidence related to validity, reliability, and fairness, they differ fundamentally in their theoretical foundations and analytical capabilities. Thus, this paper proceeds as follows: First, we examine the theoretical limitations of CTT-based approaches. Next, we introduce MFRM as an alternative measurement framework. We then systematically compare both approaches both theoretically and empirically, using a synthetic dataset from a scientist evaluation context where three senior scientists rated 40 junior scientists on five professional traits. This dataset combines authentic ratings with partially simulated data to ensure adequate sample size for analysis. This empirical example demonstrates how CTT and MFRM analyses yield different insights when applied to the same rater-mediated assessment data. Finally, we provide practical guidance for implementing MFRM-based analyses in RMPA contexts, supported by concrete examples from the scientist evaluation study. Through this combination of theoretical comparison and practical application, we aim to help researchers and practitioners make informed decisions about measurement approaches for their specific RMPA contexts.

Limitations of the CTT-Based Evaluation Approach

The CTT-based techniques for analyzing RMPAs adopt the test score tradition or number-correct approach (Engelhard et al., 2018), employing various statistical indices such as rater agreement indices, intraclass correlation coefficients (ICC), kappa coefficients, and generalizability coefficients to quantify rating consistency (Cronbach et al., 1972; Johnson et al., 2008; von Eye & Von Eye, 2005). While these methods can describe observed score patterns and summarize the degree of agreement between raters through measures like percentages of exact and adjacent category usage, they cannot directly adjust for rater effects or bias. Furthermore, these approaches rest upon a fundamental assumption that the observed ratings represent equal intervals that can be meaningfully combined using sum scores. This assumption implies that the psychological distance between rating categories (e.g., between 1 and 2, or between 2 and 3) is uniform across the rating scale - an assumption which is rarely supported empirically in rubric-based RMPAs.

For instance, on a 3-point rating scale (1 = "Needs Improvement", 2 = "Satisfactory", 3 = "Excellent"), CTT methods treat the difference between scores of 1 and 2 as equivalent to the difference between 2 and 3. However, the psychological distance between "Needs Improvement" and "Satisfactory" may be quite different from the distance between "Satisfactory" and "Excellent". This assumption of equal intervals can mask important differences in how raters interpret and use the rating scale. Therefore, even if the interrater reliability indices, such as the ICC or kappa coefficients, appear acceptable, it still does not justify the usage of a RMPA that is free from rater bias/effects.

Introduction to MFRM

As an alternative method to account for the rater variability, measurement models based on the scaling tradition (Engelhard, 2013) parameterize the structure of rating categories with category coefficients (i.e., thresholds). Thresholds that define rating categories do not need to have equal width (Engelhard & Wind, 2013). As a measurement model specifically designed for RMPAs (Eckes, 2015), MFRM is a generalized

form of the Rasch model that can incorporate multiple facets, such as raters, items, and other relevant factors that may influence the measurement process. While some facets may represent construct-irrelevant variance (e.g., systematic rater severity differences), others may capture construct-relevant aspects of the measurement. (Wright & Linacre, 1989).

The MFRM approach extends the fundamental principles of Rasch measurement to accommodate multiple facets that may influence the measurement process. While commonly used with polytomous rating scale models (Andrich 1978) and partial credit models (Masters, 1982), MFRM can extend any standard Rasch model while maintaining core measurement properties. In RMPA applications, this flexibility allows researchers to incorporate facets such as raters, items/tasks, and other relevant factors alongside the primary measurement of ratee ability/skill/proficiency, enabling systematic evaluation of their contributions to the measurement process.

In a MFRM analysis, the log-odds of each transition between adjacent rating scale categories are modeled as a function of multiple parameters estimated on a common scale: performance proficiency (for ratees), severity (for raters), and difficulty (for traits and rating scale categories). This shared metric allows direct comparison across all facets of the measurement system. Mathematically, a MFRM version of the rating scale model takes the following basic form (Linacre, 1990):

$$ln[P_{nijk} / P_{nijk-1}] = B_n - D_i - C_j - F_k, \quad (1)$$

where P_{nijk} denotes the probability of ratee n being rated k on item/task i by rater j, while P_{nijk-1} refers to the probability of ratee n being rated k - 1 on item/task i by rater j. B_n represents level of performance proficiency for ratee n, and D_i means difficulty of item/task i. Rater parameter C_j denotes severity of rater j, and F_k refers to difficulty of scale category k relative to scale category k – 1 (i.e., thresholds).

When it is not appropriate or necessary to use a fixed distance between thresholds for all items, a MFRM version of the partial credit model may be defined based on the adaptation of Equation (1) as below:

$$ln[P_{nijk} / P_{nijk-1}] = B_n - D_i - C_j - F_{ik}, \quad (2)$$

where P_{nijk} denotes the probability of ratee n being rated k on item/task i by rater j, while P_{nijk-1} refers to the probability of ratee n being rated k - 1 on item/task i by rater j. B_n represents level of performance proficiency for ratee n, and D_i means difficulty of item/task i. Rater parameter C_j denotes severity of rater j. F_{ik} represents the difficulty of scale category k relative to category k - 1 for item i, allowing thresholds to vary across items (but not across raters) for better model parsimony while still capturing important rating scale functioning (Eckes, 2015).

A partial credit model is specified based on the assumption that each rater interprets and uses each rubric element/dimension in their own individual ways. Thus, the partial credit model is a more complex model than the rating scale model and allows for the estimation of additional parameters for both raters and rubric element thresholds (Bond & Fox, 2015; Eckes, 2015; Myford & Wolfe, 2003).

The MFRM analysis allows researchers to evaluate the impact of each facet on the measurement process by estimating its unique parameter (e.g., level of severity for each rater), and then to compute the overall probability of any ratee performing on any item/task for any score category threshold and for any rater, after accounting for the estimated parameters of all facets (Bond & Fox, 2007). It is in this sense that MFRM is fully capable of modeling various facets in the RMPA setting, estimating their effects on ratings, and placing them on the same logit scale for comparison. Each facet is calibrated from the potentially ordinal raw ratings (as rating scales are often used in RMPAs), and all facets (ratee, task, rater, etc.) are placed on a single common linear scale called a variable or facets map. Thus, MFRM treats each rating as a function of the interaction between ratee ability, task difficulty, criterion difficulty, rater severity, and possibly the effects of other external, measurement-irrelevant factors (Barkaoui, 2013; McNamara, 1996).

While MFRM can handle some missing data patterns when there is adequate connectivity in the rating design (i.e., sufficient linking between facets through common elements), the model's desirable measurement properties - such as the placement of all facets on a common metric and meaningful interpretation of parameter estimates - depend critically on good model-data fit. When these conditions are met, MFRM can provide valid parameter estimates without requiring complete rating designs or assumptions about parameter distributions (Linacre, 2018). However, systematic misfit or insufficient connectivity in the rating design can compromise the interpretability of results and the validity of adjustments for rater severity. Therefore, careful evaluation of model-data fit, and rating design connectivity should precede any substantive interpretations of MFRM results.

Comparison of the CTT vs. MFRM Measurement Frameworks for RMPAs

The commonly used techniques for analyzing RMPAs under CTT and MFRM frameworks are detailed in Table 1 below. These frameworks differ fundamentally in both their theoretical foundations and analytical capabilities for addressing three critical aspects of measurement quality in RMPAs: evaluation of internal structure, assessment of rater functioning, and examination of measurement precision. While both frameworks can provide evidence related to validity, reliability, and fairness, their different theoretical foundations lead to distinct approaches for addressing these measurement challenges (Engelhard & Wind, 2018).

Scale/Rubric Internal Structure Analysis

When examining scale/rubric internal structure, CTT and MFRM frameworks offer distinct approaches with different analytical strengths. CTT factor analysis helps describe sample-dependent patterns in rating data, evaluating how well items align with hypothesized dimensions (Boone, 2016). However, factor analysis can be influenced by item difficulties, potentially identifying separate factors based on difficulty levels rather than true dimensional differences. Additionally, inter-item correlations and factor loadings may reflect sample characteristics more than fundamental measurement properties (McAuley et al., 1989).

MFRM's dimensionality analysis provides a prescriptive approach that identifies departures from unidimensional measurement while maintaining sample independence. Through analysis of standardized residuals and fit statistics, MFRM can detect subtle violations of measurement assumptions that might not be apparent in factor analysis. This approach helps identify specific items or rating patterns that compromise measurement quality, rather than just describing overall factor structure.

Rater Functioning & Interrater Analysis

CTT-based approaches to evaluating rater functioning primarily rely on interrater reliability (IRR) indices, which fall into two broad categories: consensus measures and consistency measures (Hayes & Krippendorff, 2007; Stemler & Tsai, 2008). Consensus measures assess absolute rating correspondence, including exact/adjacent agreement percentages and chance-corrected indices such as Cohen's Kappa for two raters, Fleiss' Kappa for multiple raters rating the same ratees, and Krippendorff's Alpha, which accommodates any number of raters, various levels of measurement, and missing data (Cohen, 1960; Fleiss, 1971; Krippendorff, 2011). Consistency measures such as Pearson correlations and intraclass correlation coefficients (ICC) examine relative ordering of ratings rather than exact agreement. However, these approaches have important limitations. High IRR statistics don't necessarily indicate accurate ratings since (a) raters can show high consistency while sharing systematic biases, and (b) neither type of index can detect problematic rating scale use patterns or rater severity differences (Eckes, 2012). Additionally, treating ordinal rating data as interval-level measurements in these analyses can mask important non-linearities in rating scale functioning (Wright & Linacre, 1989; Thorndike, 1904).

Table 1. List of Techniques of the CTT- vs. MFRM-Based Measurement Frameworks

Measurement Frameworks	Internal Structure	Rater Functioning	Measurement Precision
CTT	***Factor Analysis:*** - Exploratory (EFA) - Confirmatory (CFA) - Item correlation patterns - Parallel analysis - Item correlation matrix - Item covariance matrix - Factor loadings	***Consensus measures:*** - Exact/adjacent rater agreement - Rater agreement percentages ***Consistency measures:*** - Pearson interrater correlations - Intraclass correlation coefficients - Cohen's Kappa - Fleiss' Kappa - Krippendorff's Alpha	***Scale/Rubric Reliability Analysis:*** - Cronbach's Alpha Coefficient - Item-scale/total correlations - Standard error of measurement - Split-half reliability - Test-retest reliability - Generalizability coefficients
MFRM	***Unidimensionality Analysis:*** - Overall Model-Data fit statistics - Individual fit indices - Principal components analysis of standardized residuals (PCAR) - Point biserial correlations - Standardized residual analysis - Local independence evaluation	***Rater Effect Analysis:*** - Rater severity measures - Rater fit statistics - Rater-facet separation index - Rater-facet χ^2 value - Rater x criterion interactions - Rater bias/interaction analyses - Observed and expected percentages of exact rater agreements - Single rater-rest of the raters (SR/ROR) correlation measures - Differential rater functioning - Rater consistency measures	***Scale/Rubric Functioning Analysis:*** - Item fit statistics - Item-facet separation index - Item difficulty estimates - Differential item functioning - Item response category functioning - Wright maps (variable maps) - Information functions - Conditional standard errors

By contrast, MFRM addresses these limitations through a sophisticated measurement framework that fundamentally reconceptualizes how rater effects are understood and analyzed. Rather than simply describing rating patterns, MFRM models the rating process as an interaction between multiple facets - ratees, raters, items, and other relevant factors. This approach allows systematic investigation of rater functioning within the broader measurement context.

A key methodological advantage of MFRM lies in its ability to distinguish between different sources of rating variability. While CTT approaches might identify inconsistent ratings, MFRM can determine whether these inconsistencies stem from rater severity differences, scale/rubric usage patterns, interactions with particular items, or other systematic effects. This diagnostic capability not only identifies problems but also suggests specific remedies for improving rating quality (Engelhard & Wind, 2018).

Moreover, MFRM's transformation of ordinal ratings into interval measures provides a more theoretically sound basis for analyzing rater behavior. By placing all facets (including raters) on the same logit scale, MFRM enables meaningful comparisons of rater severity and consistency that are not possible with raw scores. When model assumptions are met, these measures can be adjusted for systematic rater

effects while maintaining measurement precision - a capability that goes well beyond traditional interrater reliability coefficients.

Perhaps most importantly, MFRM provides a unified framework for understanding rater functioning as part of the overall measurement process. Rather than treating rater effects as mere error to be minimized, MFRM acknowledges raters as an integral part of the measurement system and provides tools for monitoring and improving rating quality. This comprehensive approach enables more sophisticated analysis of rating quality and more targeted interventions for improving rater performance (Engelhard & Wind, 2018).

Scale/Rubric Measurement Precision Analysis

CTT approaches to measurement precision primarily rely on scale-level reliability indices, with Cronbach's alpha being the most widely used. These approaches are based on correlational statistical models that treat individual items as separate variables, computing a single standard error from variance not attributable to the assumed latent construct (Fisher Jr. et al., 2010). Supplementary evidence may come from item-scale total correlations, split-half reliability, and test-retest reliability coefficients.

However, Cronbach's alpha and related CTT indices face significant methodological limitations. As Sijtsma (2009) demonstrates, alpha values provide ambiguous evidence about scale/rubric functioning – scales/rubrics with different factorial structures can yield identical alpha values, and single-factor scales can show widely varying alphas. More fundamentally, when used to estimate scale/rubric reliability, alpha from a single test administration cannot adequately capture individual-level measurement precision.

MFRM offers several methodological advantages for assessing scale/rubric measurement precision. Rather than relying on group-level statistics, MFRM provides individual-level error estimates that function like sampling confidence intervals. These estimates become more precise with increased observations, whether through more items per person or more ratings per item (Fisher Jr. et al., 2010). The framework enables sophisticated analysis of scale/rubric functioning through:

- *Item fit evaluation that flags items exhibiting misfit, prompting further investigation into potential sources of measurement error*
- *Scale calibration that reveals category functioning and threshold structure*
- *Information functions that show measurement precision across the trait continuum*
- *Standard error estimates that quantify precision at individual levels*
- *Wright maps that visually display measurement targeting*

These tools provide detailed diagnostic information about both item-level and scale/rubric-level performance, enabling more precise assessment of measurement quality than possible with CTT indices. Most importantly, MFRM's interval-level measurement properties allow meaningful interpretation of score differences and more accurate assessment of measurement precision across different rating contexts.

To sum up, these fundamental differences between CTT and MFRM frameworks have important implications for RMPA practice. While CTT methods can provide useful descriptive information about rating patterns, MFRM offers additional analytical capabilities for examining and potentially adjusting for various measurement effects. However, these advantages depend on meeting model assumptions and establishing adequate connectivity in rating designs.

Table 1 provides a comprehensive inventory of techniques available within each framework. The following empirical demonstration illustrates typically used techniques from each category; a complete demonstration of all listed methods is beyond the scope of a single example, though the principles generalize across techniques within each framework.

Typical MFRM-based Evaluation Procedures and Techniques

This section presents a brief tutorial on how to conduct an MFRM-based evaluation study for RMPAs. First, typical Research Questions in such evaluation studies are identified and listed, followed by the explanation about Research Design. Most importantly, the Data Analysis Techniques are clearly outlined for addressing the typical Research Questions, involving all necessary procedures and techniques.

Sample Research Questions

A comprehensive MFRM-based evaluation study for RMPAs typically addresses the following eight research questions regarding how to control various construct-irrelevant measurement errors of using the RMPA instrument:

1. To what extent do the observed rating data obtained from the RMPA instrument fit the MFRM modeling?
2. To what extent does the RMPA instrument separate ratees into distinct levels of proficiency?
3. To what extent do raters differ in terms of the relative severity with which they rate ratees?
4. To what extent do raters consistently rate the performance of ratees?
5. To what extent do raters consistently rate the performance of ratees across the RMPA items?
6. To what extent can the score levels of the individual RMPA items be distinguished, without certain score levels being either underused or overused?
7. To what extent are the rater behaviors associated with the professional/personal background characteristics of ratees?
8. To what extent are the rater behaviors associated with the professional/personal background characteristics of the raters themselves?

Research Design

A MFRM-based evaluation study for RMPAs is conducted within a Rasch framework, including the investigation of dimensionality, ratee fit, item fit, rater fit, overall data-model-fit, as well as possible interactions between any of the modeled facets/factors. The key lies in systematically calibrating the measures of all the involved facets (e.g., test item, ratee, raters, and other external factors) on a common continuum scale, so that the construct-irrelevant measurement errors (especially rater bias) can be effectively identified and accounted for. The calibrated ratings/scores after the MFRM analysis can theoretically be compared with confidence across different rating contexts.

Data Analysis Techniques

We illustrate techniques that researchers can use to evaluate RMPAs using a MFRM approach with open-source R packages (standalone commercial Rasch software programs such as Facets and Winsteps can also be used if preferred).

Local Independence. Local independence (LID) refers to the assumption that item responses are independent from one another after controlling for the construct of interest (DeMars, 2010). Therefore, there should be no significant correlation between two items after controlling for the underlying trait, as some residual association may occur due to random variation. In other words, the items should only be correlated primarily through the latent trait that the test is measuring.

LID violations are problematic because they may influence parameter estimates as well as inflate reliability estimates (Marais & Andrich, 2008), since locally dependent items always cause substantial information loss for IRT modeling.

Among the variety of methods for identifying LID violations that have been proposed in the related literature, the most widely used approach is based on Yen's Q_3 (1993) statistics through computing item residuals (observed item responses minus their expected values); and then correlating these residuals. Thus, in practice, LID violations are detected through observing the correlation matrix of item residuals based on estimated item and person parameters, and residual correlations above a certain cut-off value are pinpointed as the items that appear to be locally dependent.

Although no single critical cut-off value of Q statistics is appropriate across all situations, simulation studies suggest that the Q critical value tends to be approximately 0.2 above the average residual correlation. Item residual correlations that exceed this guideline may indicate potential local dependence, and residual correlations that are 0.3 above the average correlation are generally uncommon for independent items. (Christensen, Makransky, & Horton, 2017).

The Yen's Q (1993) statistics can be calculated and investigated in R using packages such as `mirt` or through custom functions that compute correlations between item residuals after fitting the MFRM model with the "TAM" package (as demonstrated in Appendix A). When using the "TAM" package, item residuals can be extracted from the fitted model and their correlations examined to identify potential local dependence. Alternatively, this analysis can also be conducted in the Winsteps software program, where Table 23.99 (i.e., largest residual correlations for items) can be obtained for pairwise, item-level residual correlations by specifying the command of "PRCOMP = R" in the control file.

Unidimensionality. Unidimensionality is related to LID and refers to the assumption that all assessment items measure only one common construct. Unidimensionality is evaluated by conducting a Principal Components Analysis (PCA) on the standardized residuals (PCAR) following the MFRM analysis. The number and type of facets depend on the specific RMPA context, while common configurations include ratees, items, and raters, with additional facets (e.g., rating occasions, tasks, contexts) incorporated as warranted by the research design. The PCAR can be conducted in R using the "TAM" package by extracting standardized residuals from the fitted model and performing principal components analysis on these residuals using base R functions (`prcomp()` or `princomp()`). As illustrated in Appendix A, this provides eigenvalues and variance explained statistics for evaluating dimensionality. Alternatively, the PCAR can also be conducted using the Winsteps software program, version 4.7.0 (Linacre, 2020).

The general procedures for conducting PCAR analysis in R include: (a) fitting a MFRM model using the "TAM" package with the appropriate facets specified (as shown in Appendix A, using the `tam.mml.mfr()` function), (b) extracting standardized residuals from the fitted model, and (c) conducting principal components analysis on these residuals to identify potential secondary dimensions. Alternatively, when using dedicated Rasch software, the procedures are as follows: (a) a MFRM analysis is carried out in Facets (Linacre, 2020) with facets specified according to the research design (e.g., ratees, items, raters, and any additional relevant facets), and (b) a rectangular data output file is exported from Facets into Winsteps, containing the RMPA items as its columns and "ratees + raters" combined as its rows for a PCAR analysis in the Rasch framework.

PCAR analyses are used to evaluate whether there are systematic patterns in the item-level standardized residuals. If there are patterns in the residuals, a secondary dimension (i.e., a contrast) may be present. It is assumed that all items should be loaded on the first contrast of the Rasch dimension, and the PCAR specifically tests whether any items group on secondary contrasts. Each contrast has an associated eigenvalue, and the eigenvalues represent the number of items that make up the respective contrast. If

eigenvalues for all secondary contrasts are less than 2.0 (indicating there are fewer than two elements on the secondary contrasts), the unidimensionality assumption is generally considered supported (Linacre, 2023; Smith, 2002). However, this guideline should be applied with judgment, as the appropriate threshold may vary depending on test length, sample size, and the specific measurement context (Chou & Wang, 2010).

Overall Model Fit. To evaluate the overall model fit of the MFRM analysis, the absolute values of the standardized residuals are examined. Standardized residuals represent the number of standard deviations the observed score/rating deviates from the expected score/rating. For instance, standardized residuals of |2.0| indicate that the observed score deviates by two standard deviations from the expected score. Thus, a commonly applied guideline suggests that standardized residuals greater than |2.0| often indicate unexpected scores, and these would typically be expected to appear less than 5% of the time in data that fit reasonably well with the chosen MFRM model (Bond & Fox, 2015). In typical MFRM-based evaluation studies, data are deemed to have good overall model-fit, if fewer than 5% of the standardized residuals appear greater than or equal to |2.0|.

Rater Fit and Item Fit. Mean Square outfit and Mean Square infit statistics (referred to as MnSq outfit and infit indices) are calculated and investigated to evaluate rater fit or item fit (Bond & Fox, 2015).

MnSq outfit and infit indices range from 0 to positive infinity, with values of 1.0 indicating perfect fit of the data to the model (Linacre, 2020). Values less than 1.0 indicate that the observed ratings are closer to the model-implied ratings than would be predicted by the model (i.e., overfit of the model), and values greater than 1.0 indicate that the observed ratings are less similar to the model-implied ratings than would be predicted by the model (i.e., underfit of the model).

Various guidelines have been proposed for interpreting fit based on MnSq outfit and infit indices. Linacre (2003) suggests that outfit and infit values approximately between 0.5 and 1.5 can generally indicate acceptable fit, while Bond and Fox (2015) recommend narrower ranges of about 0.7 to 1.3 for high-stakes applications. However, appropriate fit thresholds depend on assessment purpose, sample characteristics, and substantive considerations (Wright & Linacre, 1994). For exploratory analyses or low-to-medium stakes RMPAs, the 0.5-1.5 range provides a reasonable starting point, though practitioners should interpret fit statistics in conjunction with other diagnostics and substantive understanding of the measurement context rather than applying rigid cutoffs mechanically.

MFRM Parameter Estimation. MFRM analysis yields (a) a measure of the ratee ability/rater severity/item difficulty parameter on a logit scale for each ratee/rater/item, respectively, together with (b) a standard error (SE) that indicates the uncertainty associated with that parameter estimate. These analyses can be conducted using various software programs, including Facets (Linacre, 2020), Winsteps (Linacre, 2023), or R packages such as TAM (Robitzsch et al., 2022). These measures are examined for the overall range/spread to determine how varied they are in this study sample. In addition, the average measure can also be calculated as the average proficiency/effectiveness of ratees, average rater severity, or average item difficulty. A relatively low SE value is desired, as it indicates low measurement errors associated with the measures and high level of precision in estimating these measures.

The Separation Index for each facet indicates how many levels of ratee ability, rater severity, or item difficulties can be distinguished based on the RMPA data; while the Reliability of Separation indicates the degree to which the MFRM analysis reliably distinguishes between these different levels. Fixed χ^2 tests the null hypothesis that all ratees/raters/items are equal in their estimated measures, a very easy assumption to violate in empirical studies.

MFRM-Based Bias Analysis. MFRM-based bias analysis investigates whether a particular aspect of the assessment setting elicits a consistently biased pattern of scores/ratings. After estimating the main effects respectively for the rater severity (across all tasks), RMPA item difficulty (across all raters), and ratee ability

(across all items and raters), the MFRM analysis estimates the most likely score for each ratee with a given rater on a specific task, if the rater's rating behavior remains consistent across all RMPA items. These individual ratee scores are totaled across all ratees to produce a total *expected* score given by each rater on each item. This *expected* total score is then compared to the *observed* total score for all the ratees on the same item.

If the observed score for a given RMPA item is higher than the expected score, this item seems to have elicited more lenient behavior than usual on the part of the raters. Fit statistics of the bias analysis summarize for each rater, item, and ratee the extent to which the differences between expected and observed values are within a normal range (expressed in standard deviations from the mean fit statistics).

McNamara (1996) and Kondo-Brown (2002) suggest that researchers may focus on potentially biased interactions with *Z*-values approximately equal to or higher than the absolute value of 2, along with MnSq infit values falling roughly within the range of two standard deviations around the mean of infit values.

An Empirical Example

Empirical RMPA Context

This study utilized data from the DICES (Diversity in Conversational Artificial Intelligence Evaluation for Safety) Dataset 350, a publicly available benchmark designed to capture diverse perspectives on the safety evaluation of conversational Artificial Intelligence (AI) systems (Aroyo et al., 2023). The dataset contains multi-turn adversarial conversations generated by human agents interacting with a Large Language Model, with each conversation rated for safety by a diverse human rater pool. This context represents an emerging application of RMPA methodology: using human raters to assess AI performance (i.e., evaluating AI-generated content for safety) and alignment rather than traditional human performance assessment.

The original DICES 350 dataset comprises 350 adversarial conversations rated by 123 unique raters, with raters balanced by gender (man, woman) and race/ethnicity (White, Black, Latine, Asian, Multiracial). Each rater evaluated all conversations, yielding a fully crossed rating design that enables comprehensive examination of rater effects, item functioning, and measurement precision.

For this methodological demonstration, we randomly selected a subset of 5 raters evaluating 100 conversations to ensure computational feasibility while maintaining the fully crossed design. This yielded the final analytic dataset comprised 500 observations with complete ratings across six aggregate safety items: harmful content overall (Q2), bias overall (Q3), misinformation (Q4), political affiliation (Q5), policy guidelines overall (Q6), and an overall safety rating (Q_overall). Responses were recorded as NO (0), UNSURE (1), or YES (2). The six aggregate items were selected over the 24 original sub-items to avoid redundancy, maintain local independence assumptions, and align with the original DICES study's data analytical approach.

Unlike traditional RMPA contexts where rater disagreement may reflect measurement error, this dataset presents unique challenges: rater variability may stem from legitimate differences in individual rater interpretations of harm, the inherent ambiguity of adversarial conversational content, and systematic differences in safety perceptions across demographic groups. These characteristics make the DICES dataset particularly well-suited for demonstrating how MFRM can disentangle multiple sources of variance in complex rating contexts.

All analyses were conducted using R 4.3.1 (R Core Team, 2023), specifically the `TAM` package for MFRM analysis, with visualization support from the `ggplot2` package. While our tutorial section references

specific procedures available in standalone Facets and Winsteps software, these analyses and visualizations can be implemented using equivalent functionality in open-source R packages.

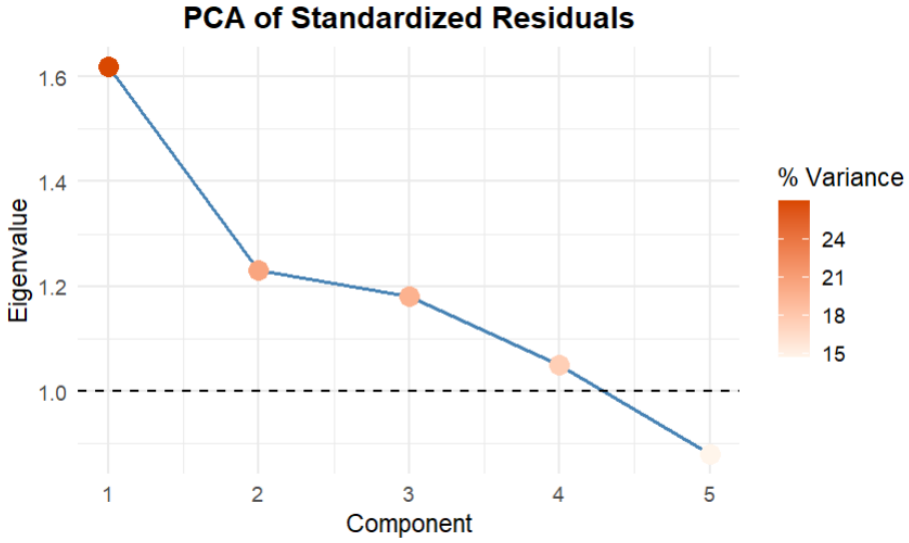
MFRM-Based Analysis Results

MFRM-based analytical procedures were conducted in R 4.3.1 (R Core Team, 2023) to systematically examine the RMPA measurement quality, addressing key aspects such as scale/rubric internal structure, rating consistency, measurement precision, and potential systematic biases in the assessment process.

Scale/Rubric Internal Structure Analysis. MFRM-based unidimensionality analyses were conducted to examine the dimensionality of the six-item safety assessment rubric. The Principal Components Analysis of Residuals (PCAR) revealed that the first component explained 26.98% of the total variance, with a first-to-second eigenvalue ratio of 1.32. The average inter-item residual correlation was -0.14, indicating weak negative relationships among items after accounting for the Rasch dimension. The scree plot (Figure 1) illustrates eigenvalues across five components, with four components exceeding the eigenvalue threshold of 1.0 and no clear "elbow," suggesting strong multidimensionality.

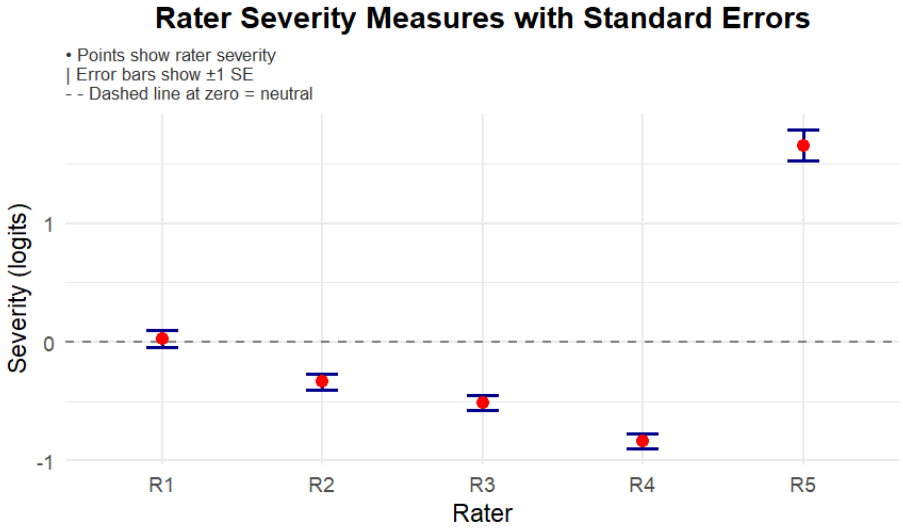
These findings indicate that the six safety items do not form a strictly unidimensional construct for AI-generated content safety. The negative average inter-item residual correlation suggests that raters may have interpreted certain harm categories as inversely related, such as rating content high on political affiliation but low on misinformation. This pattern implies that the safety items function as distinct dimensions rather than indicators of a single underlying "harm" construct, which has important implications for aggregating scores across items.

Figure 1. Principal Component Analysis Results of Standardized Residuals



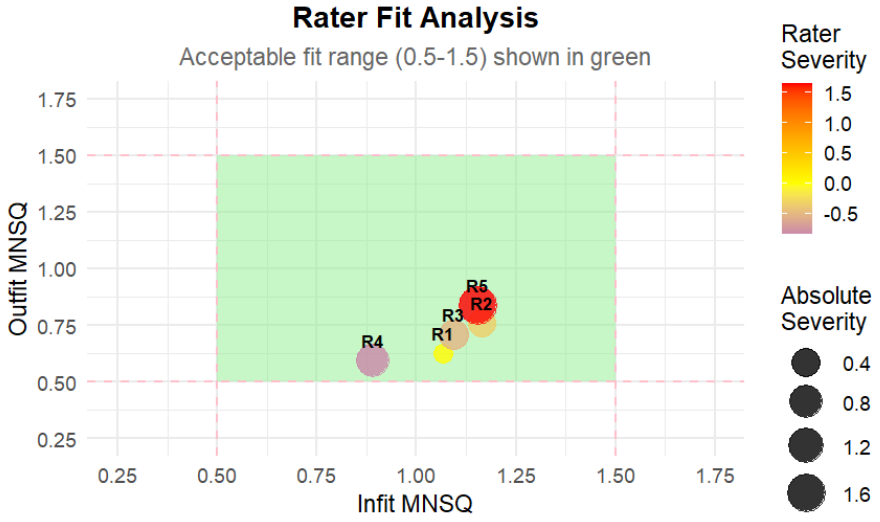
Rater Functioning & Interrater Analysis. MFRM-based rater severity analyses revealed substantial differences among the five raters, with severity measures ranging from -0.83 logits (most lenient, R4) to 1.66 logits (most severe, R5), yielding a severity range of 2.49 logits. As shown in Figure 2, non-overlapping standard errors between R5 and all other raters indicate statistically significant severity differences. The significant fixed-effect chi-square (χ^2 = 450.24, df = 4, p < .001) confirms systematic differences in rater behavior. The high separation reliability (0.99) and separation index (11.87) indicate that raters can be reliably distinguished into at least ten statistically distinct severity levels - substantially exceeding the minimum acceptable values of 0.70 and 2.0, respectively.

Figure 2. Results of MFRM-Based Rater Severity Analysis



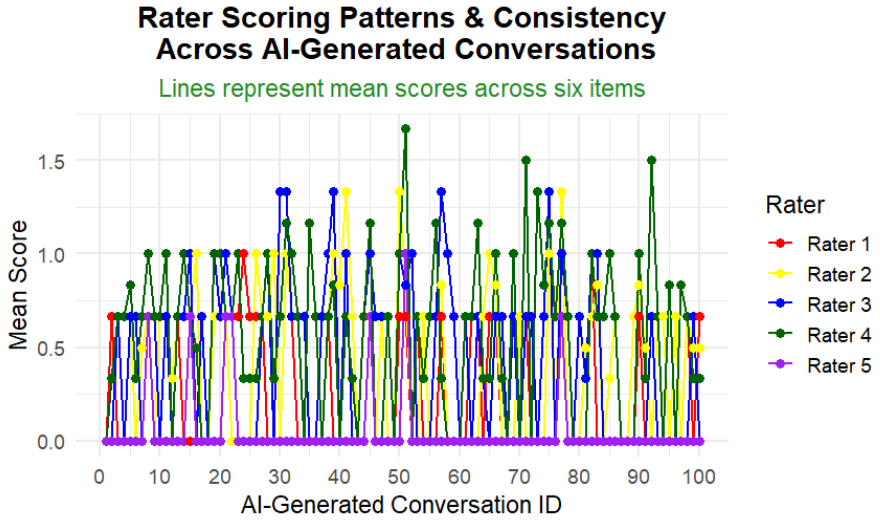
Furthermore, as shown in Figure 3, all raters demonstrated acceptable Infit statistics (0.89-1.16) within the 0.5-1.5 range, indicating internal consistency in applying their own rating standards. Outfit values ranged from 0.59 to 0.84, with some raters showing slight overfit, suggesting overly predictable response patterns. Overall, raters demonstrated adequate intrarater consistency.

Figure 3. Results of Rater Fit Analysis



However, interrater agreement analyses revealed mixed results. Exact agreement percentages between rater pairs ranged from 28% to 63%, with many pairs falling below the 60-70% threshold, typically considered acceptable for performance assessments. The Single Rater/Rest of Raters (SR/ROR) correlations ranged from 0.34 to 0.51, indicating moderate agreement between individual raters and the collective judgment of other raters. As illustrated in Figure 4, raters demonstrated notably different scoring patterns: R5 consistently assigned scores near zero (detecting minimal harm), while R4 showed the highest variability with frequent elevated scores. These divergent patterns align with the DICES dataset's design intention to capture diverse perspectives on AI safety, suggesting that rater variability reflects legitimate differences in harm perception rather than measurement error requiring remediation.

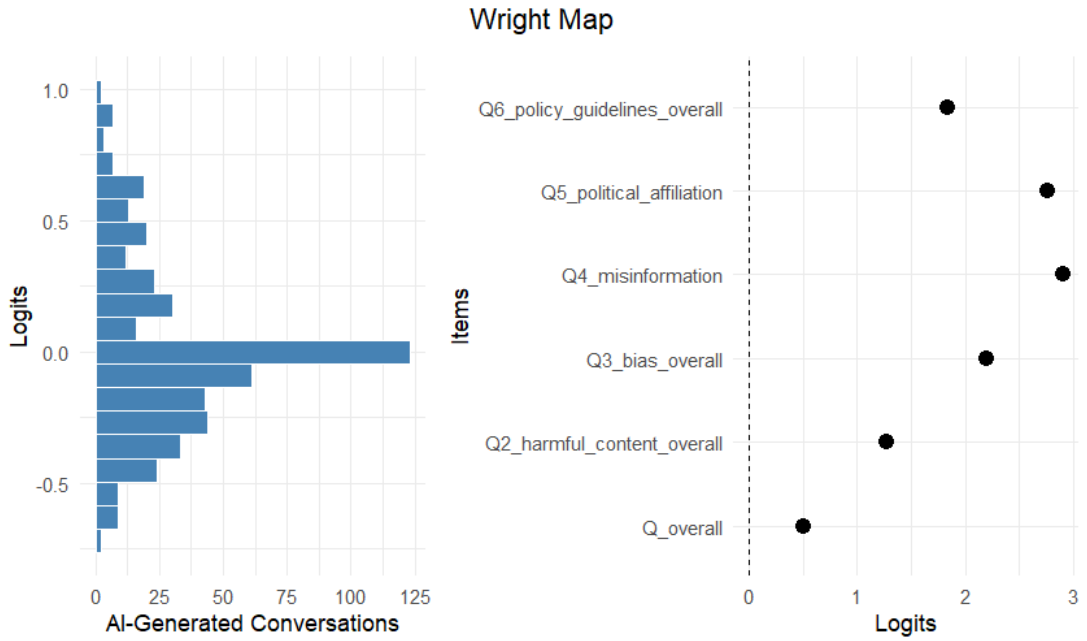
Figure 4. Rater Scoring Patterns across AI-Generated Conversations



These findings reveal substantial differences in rater severity levels (2.49 logits range) despite acceptable individual fit statistics. The moderate agreement rates (28-63%) and SR/ROR correlations (0.34-0.51) suggest systematic differences in how raters perceive AI-generated harm. In traditional RMPA contexts, such variability would warrant rater training; however, for AI safety evaluation, this diversity may be intentional and valuable, capturing the range of human perspectives that AI systems must navigate.

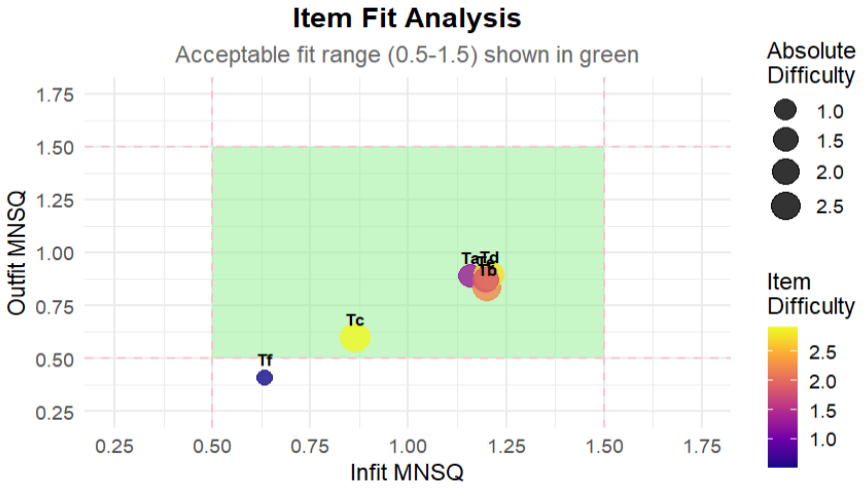
Scale/Rubric Measurement Precision Analysis. The Wright Map (Figure 5) shows person ability (i.e., in our case, this refers to safety level of AI-generated content) estimates clustered between -0.72 and 0.99 logits, while item difficulties ranged from 0.50 logits (Q_overall, easiest to endorse harm) to 2.91 logits (Q4_misinformation, hardest to endorse harm). This suggests raters were generally reluctant to identify harm, with misinformation and political affiliation being the most difficult items to endorse.

Figure 5. Wright Map by Six Safety Criteria Across Five Raters



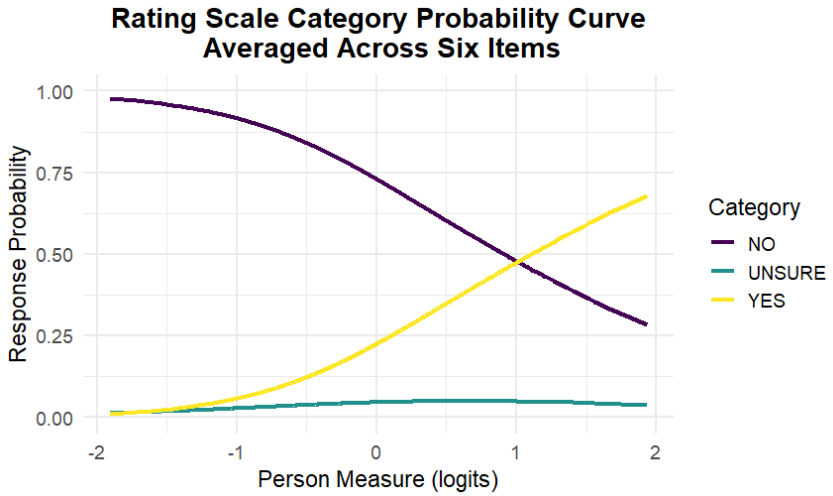
Item fit analysis (Figure 6) demonstrates all six items functioning approximately within acceptable ranges (0.5-1.5) for both Infit and Outfit MnSq, indicating reasonable model-data fit. Items showed a wide difficulty range (2.41 logits) with reasonable precision (mean SE = 0.09). The item separation index (9.54) and reliability (0.99) indicate the six-item safety evaluation rubric reliably distinguishes at least nine statistically distinct levels of endorsement difficulty.

Figure 6. Results of Item Fit Analysis



The Rating Scale Category Probability Curves (Figure 7) show that while scores of 1 and 9 are clearly distinct, the middle scores (2-8) tend to overlap, suggesting raters might have difficulty distinguishing between adjacent score points. This indicates that the nine-point scale might be more complex than necessary. The Rating Scale Category Probability Curves (Figure 7) reveals a critical issue: the UNSURE category never becomes the most probable response at any point along the latent trait continuum. Raters transition directly from NO to YES without meaningfully utilizing the middle UNSURE category. This suggests the 3-category scale functions effectively as a dichotomous scale, and the UNSURE option may introduce ambiguity rather than capturing meaningful gradations in harm perception.

Figure 7. Rating Scale Category Functioning for the Six-Item Safety Rubric



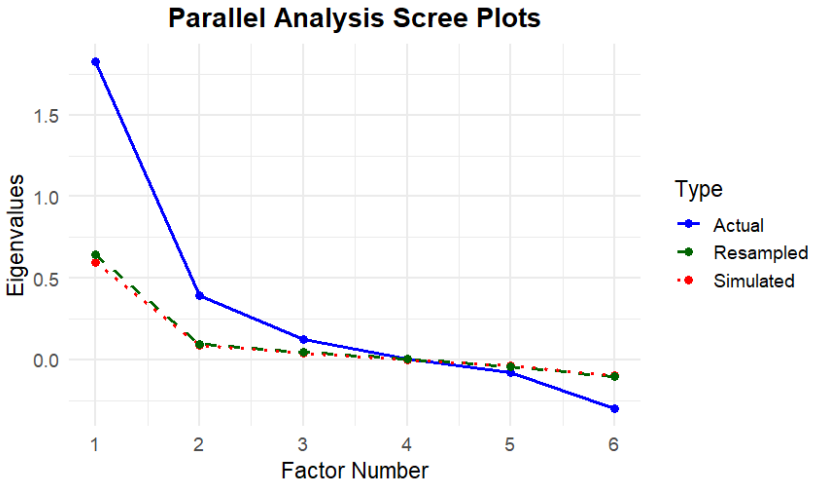
These findings suggest that while item measurement properties are strong, the rating scale structure is suboptimal. Collapsing to a dichotomous (NO/YES) scale or providing clearer operational definitions for UNSURE may improve measurement precision.

Classical Test Theory (CTT) Analysis Results

CTT-based analysis was conducted in R 4.3.1 (R Core Team, 2023) as follows:

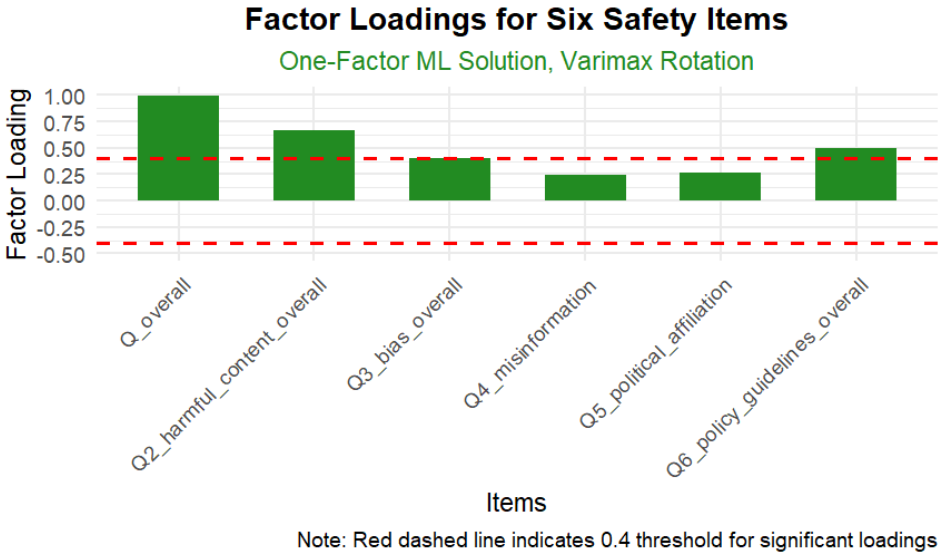
Scale/Rubric Internal Structure Analysis. The internal structure of the six-item safety rating rubric was examined using CTT-based factor analysis. A parallel analysis (comparing observed eigenvalues against both simulated and resampled eigenvalues from 100 randomly generated datasets) suggested a three-factor solution. As shown in Figure 8, the first three empirical eigenvalues (1.83, 0.40, 0.13) exceeded their corresponding simulated thresholds (0.60, 0.09, 0.04), while subsequent eigenvalues fell below random data cut-offs. This finding suggests the six safety items may not form a unidimensional construct.

Figure 8. Parallel Analysis Scree Plots



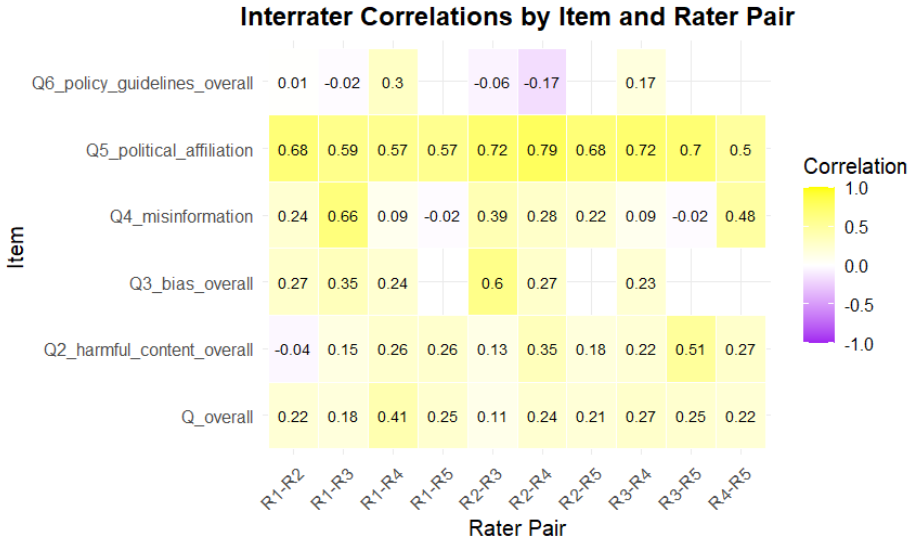
Despite the parallel analysis suggesting multidimensionality, a one-factor ML solution with Varimax rotation was examined for comparison (Figure 9). Q_overall (1.00) and Q2_harmful_content_overall (0.67) loaded strongly on the common factor, while Q6_policy_guidelines_overall (0.50) and Q3_bias_overall (0.40) showed moderate associations. However, Q4_misinformation (0.24) and Q5_political_affiliation (0.26) fell below the 0.40 threshold, contributing minimally to the factor. Fit indices indicated poor model fit: TLI = 0.51 (below the 0.90 threshold), RMSEA = 0.23 (exceeding the 0.08 criterion), and a significant chi-square ($\chi^2 = 248.43$, $p < .001$). These results suggest the single-factor model inadequately represents the data structure.

Figure 9. Single-Factor Solution Resulted from Factor Analysis



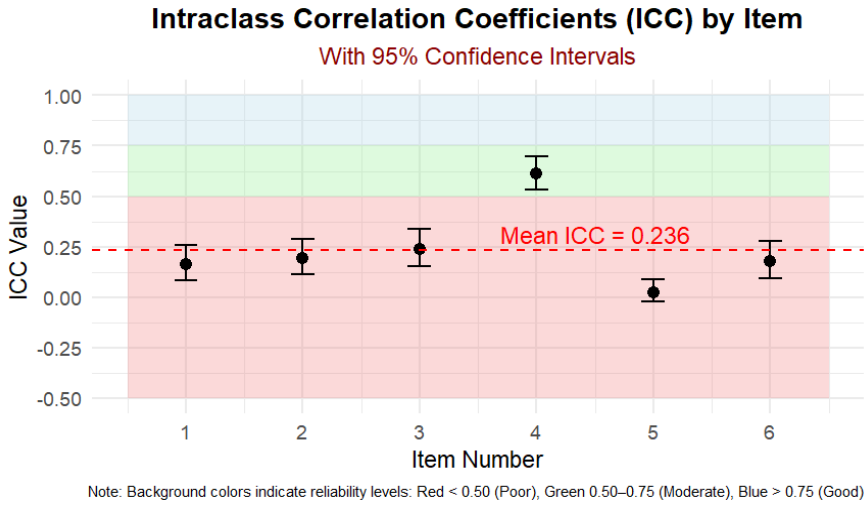
Rater Functioning & Interrater Analysis. Interrater agreement was evaluated using Pearson correlations for each item across all 10 rater pairs. Correlations ranged from -0.17 (Q6_policy_guidelines_overall between R2-R4) to 0.79 (Q5_political_affiliation between R2-R4). Mean correlations by rater pair ranged from 0.23 (R1-R2) to 0.37 (R4-R5), indicating generally weak to moderate agreement. As shown in Figure 10, Q5_political_affiliation demonstrated consistently high correlations across rater pairs (0.50–0.79), while Q6_policy_guidelines_overall showed the weakest and occasionally negative correlations, suggesting systematic disagreement in how raters applied this criterion.

Figure 10. Interrater Correlations by Item and Rater Pair



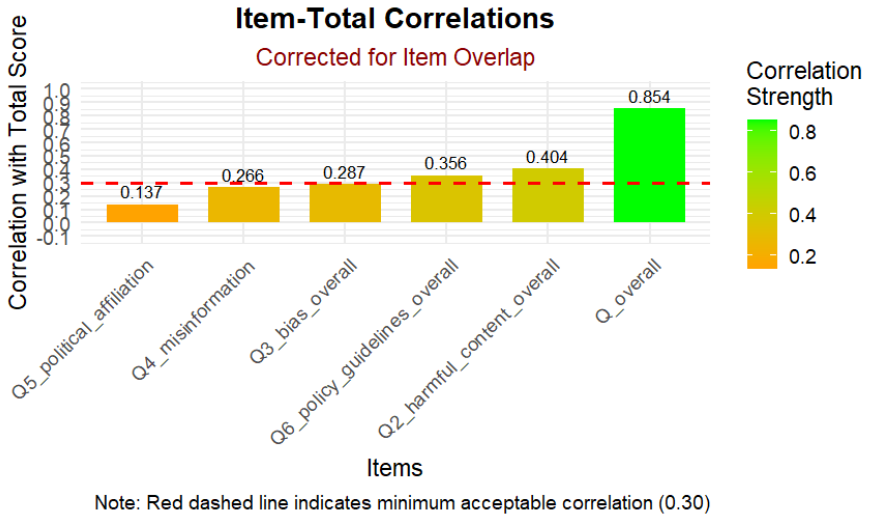
The ICC analysis (Figure 11) confirmed variable interrater reliability across items. ICC values ranged from 0.025 (Q6_policy_guidelines_overall) to 0.615 (Q5_political_affiliation), with a mean ICC of 0.24. Only Q5_political_affiliation exceeded the 0.50 threshold for "moderate" reliability, while the remaining five items fell in the "poor" range. These findings indicate that raters showed reasonable agreement only on political affiliation judgments, with substantially lower consistency on other safety dimensions.

Figure 11. Intraclass Correlation Coefficients (ICC) by Item



Scale/Rubric Measurement Precision Analysis. The CTT-based reliability analysis revealed marginal internal consistency (Cronbach's α = 0.65), below the conventional 0.70 threshold. Item-total correlations (Figure 12) ranged from 0.14 (Q5_political_affiliation) to 0.85 (Q_overall). Only three items exceeded the acceptable 0.30 cutoff: Q_overall (0.85), Q2_harmful_content_overall (0.40), and Q6_policy_guidelines_overall (0.36). The remaining three items - Q3_bias_overall (0.29), Q4_misinformation (0.27), and Q5_political_affiliation (0.14) - fell below acceptable thresholds, suggesting weak contribution to the total score. The standard error of measurement was 1.47 points on the 3-point scale, and the average inter-item correlation was 0.21 (signal-to-noise ratio = 0.26), indicating that the six items do not cohere strongly as a unified scale.

Figure 12. Corrected Item-Total Correlations



Systematic Comparison & Implications of the CTT vs. MFRM Empirical Analyses

Applying both MFRM and CTT-based evaluation frameworks to the same AI safety rating dataset highlights how methodological choice shapes the diagnostic story practitioners receive about their instruments. CTT provides familiar summary statistics (e.g., factor loadings, reliability coefficients, and agreement indices), yet leaves critical questions unanswered. In contrast, MFRM transforms the same ratings

into a comprehensive diagnostic of the measurement system, exposing where, why, and by how much the rubric, raters, and rating scale deviate from intended functioning.

Internal-Structure Insights. Both approaches converged on evidence of multidimensionality, but with different levels of specificity. CTT's parallel analysis suggested a three-factor solution, and the one-factor model showed poor fit (TLI = 0.51, RMSEA = 0.23). Factor loadings revealed that only three of six items exceeded the 0.40 threshold, with Q5_political_affiliation (0.26) and Q4_misinformation (0.24) contributing minimally. MFRM's residual PCA told a complementary but richer story: the first component explained only 26.98% of variance with a weak eigenvalue ratio (1.32), and the average inter-item correlation was negative (-0.14), suggesting raters interpreted certain harm categories as inversely related. For practitioners, MFRM's insight that items function as distinct - even opposing - constructs provide clearer guidance: aggregating scores across items may obscure meaningful distinctions, warranting either multidimensional reporting or rubric revision.

Rater-Functioning Diagnostics. CTT flagged variable interrater agreement (ICC range: 0.03-0.62, mean = 0.24) and pairwise correlations ranging from -0.17 to 0.79, but could not explain *why* raters disagreed. MFRM decomposed the problem with precision: severity measures spanned 2.49 logits (from -0.83 to 1.66), with R5 substantially more severe than all other raters. The high separation index (11.87) and significant chi-square (χ^2 = 450.24, $p < .001$) confirmed systematic, reliable differences in rater behavior. Critically, fit statistics showed all raters were internally consistent (Infit 0.89-1.16), indicating the issue was not random error but divergent interpretations of safety constructs. This diagnostic granularity (i.e., distinguishing severity from consistency) directs calibration efforts precisely where needed rather than simply urging raters to "agree more." Thus, MFRM carries significant, unique diagnostic value for empirical measurement settings such as the DICES dataset, developed specifically to capture heterogeneity/diversity in human perceptions of AI safety. In such cases, cross-rater disagreement may represent legitimate differences in harm perception rather than remediable measurement error.

Measurement-Precision Evidence. CTT's Cronbach's α = 0.65 signaled marginal reliability, with item-total correlations ranging from 0.14 to 0.85 and a standard error of measurement of 1.47 points. These statistics indicate problems but offer no remediation pathway. MFRM, via Wright maps and category-probability curves, pinpointed two actionable issues: (a) item difficulties (0.50-2.91 logits) exceeded the estimates of AI-generated content safety level (-0.72 to 0.99 logits), indicating raters were generally reluctant to endorse harm; and (b) the UNSURE category never emerged as the most probable response at any harm level, suggesting the 3-point scale functions effectively as dichotomous. Practitioners now have a clear path (i.e., collapsing to a YES/NO format or operationally defining UNSURE) before concluding the scale is fundamentally flawed.

Practical Bottom Line. This empirical example clearly demonstrates that comparatively speaking, where CTT evaluation summarizes "what" (low reliability, weak interrater agreement), MFRM analysis explains "why" and "how to fix it." By locating rater severity, scale misfit, and construct dimensionality on a common logit ruler, MFRM approach turns ratings into interval-level evidence, supports equitable score adjustments, and delivers concrete design feedback. For organizations that rely on defensible, data-driven performance decisions, the additional analytic effort pays tangible dividends in fairness, precision, and actionable insight - benefits that traditional CTT approaches may not be able to provide.

Discussion & Conclusion

This systematic methodological comparison of CTT and MFRM approaches, supported by our empirical example using AI safety evaluation data, reveals several key insights into RMPA measurement practice. While

both frameworks can provide useful information about measurement quality, they differ fundamentally in their capabilities for addressing three critical measurement challenges.

Specifically, regarding examining RMPA scale/rubric internal structure, while CTT approaches can identify broad factorial patterns, MFRM provides more nuanced understanding of how rating scales function in practice. Our empirical analysis illustrated this through principal components analysis of standardized residuals, which revealed not just the presence of potential multiple dimensions but also specific patterns in how raters interpreted and applied different safety criteria, including negative inter-item residual correlations suggesting that raters perceived certain harm categories as inversely related.

In evaluating rater functioning, MFRM transcends traditional reliability indices by modeling the rating process as an interactive system. The empirical example demonstrated how MFRM can simultaneously evaluate rater severity, internal rating consistency, and scale usage patterns. This comprehensive analysis revealed substantial variation in rater severity (spanning 2.49 logits) alongside acceptable individual fit statistics, indicating that raters applied systematically different standards while maintaining internal consistency. Crucially, in contexts like AI safety evaluation where diverse perspectives may be intentional rather than error, MFRM's ability to distinguish severity differences from inconsistency prevents misinterpreting legitimate disagreement as measurement dysfunction (insights not readily apparent through conventional CTT analyses).

The frameworks also differ markedly in their approach to measurement precision. Where CTT relies primarily on group-level statistics like Cronbach's alpha, MFRM provides detailed information about measurement quality at multiple levels. Our empirical analysis showed how Wright maps, item fit statistics, and rating scale diagnostics can identify specific measurement challenges, such as the UNSURE category never functioning as the most probable response and misalignment between item difficulties and rater endorsement patterns. These insights enable more targeted improvements to assessment instruments than possible through CTT indices alone.

These methodological advantages of MFRM over CTT approaches demonstrate its potential for enhancing RMPA measurement quality. However, realizing these benefits depends on meeting certain fundamental requirements. Good model-data fit and adequate connectivity in rating designs are essential prerequisites for valid parameter estimation and meaningful adjustments. Our empirical example illustrated how careful evaluation of fit statistics and rating design structure should precede substantive interpretations. When these conditions are met, MFRM can provide powerful tools for improving rating quality; when they are not, practitioners may need to modify their rating designs or consider alternative analytical approaches. Understanding these requirements is crucial for making informed decisions about measurement approaches in specific RMPA contexts.

For practitioners, MFRM provides concrete tools to enhance RMPA assessment quality through sophisticated analysis of rater effects, detailed examination of rating scale functioning, and detection of potential systematic biases. Our empirical example revealed MFRM's capability to identify specific patterns in rater behavior and measurement functioning that might affect assessment validity, such as differential interpretation of safety criteria across raters, systematic severity differences in evaluating certain content types, and rating scale categories that do not function as intended. These insights are valuable across diverse RMPA applications from traditional personnel evaluation to emerging domains like AI content safety assessment, where ratings inform consequential decisions (Bond & Fox, 2015; Popham, 2018).

While implementing MFRM requires initial investment in methodological training, proper data collection designs, and ongoing quality monitoring, our analysis suggests these requirements are justified by the resulting improvements in measurement quality and fairness, particularly in high-stakes assessment contexts.

Conclusion

The paper's primary contribution lies in translating sophisticated measurement theory into clear and actionable practical guidance. Through detailed step-by-step analytical procedures and a concrete empirical demonstration, we offer practitioners a comprehensive tutorial for implementing MFRM in their own assessment contexts. This practical focus distinguishes our work from previous methodological comparisons that have primarily served research-oriented audiences.

Specifically, because MFRM has long been recognized for its methodological sophistication, its adoption in applied settings has been limited by perceived complexity and implementation challenges. Our systematic comparison addresses this gap by providing practitioners with a clear, evidence-based framework for understanding when and how MFRM can enhance assessment quality beyond traditional CTT approaches.

Furthermore, our analysis uniquely emphasizes the role of measurement approaches in promoting performance assessment fairness. By demonstrating how MFRM can identify and address specific threats to measurement quality, we provide organizations with practical tools for enhancing the equity of their performance evaluation systems. This connection between measurement precision and assessment fairness offers particularly valuable insights for organizations striving to improve their evaluation practices.

Ultimately, this work serves as a bridge between measurement theory and organizational practice, helping practitioners make informed decisions about assessment methodology while understanding both the benefits and requirements of more sophisticated measurement approaches.

Data Availability Statement

The empirical dataset used in this study is derived from the publicly available DICES (Diversity in Conversational AI Evaluation for Safety) Dataset 350 (Aroyo et al., 2023), accessible at https://github.com/google-research-datasets/dices-dataset. For our demonstration, we used a randomly selected subset of 5 raters evaluating 100 conversations across 6 aggregate safety items (500 observations). The complete R analysis scripts are available from the corresponding author upon reasonable request.

Declaration of Interest Statement

The authors declare no conflict of interest. This research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 3/25/2024. **Accepted:** 12/16/2025. **Published:** 12/19/2025.

Citation: Niu, C., Bradley, K., Jin, R., & Love, A. (2025). Systematic Comparison of Two Approaches for Validating and Using Rater-Mediated Performance Assessments. *Practical Assessment, Research, & Evaluation*, 30 (1) (13). Available online: https://doi.org/10.7275/pare.2042

Corresponding Author: Chunling Niu, University of the Incarnate Word. Email: chunling.niu@gmail.com

References

Andrich, D. (1978). Relationships between the Thurstone and Rasch approaches to item scaling. *Applied Psychological Measurement*, *2*(3), 451-462.

Aroyo, L., Taylor, A. S., Díaz, M., Homan, C. M., Parrish, A., Serapio-García, G., Prabhakaran, V., & Wang, D. (2023). DICES dataset: Diversity in conversational AI evaluation for safety. *arXiv preprint arXiv:2306.11247*. https://arxiv.org/abs/2306.11247

Barkaoui, K. (2013). Multifaceted Rasch analysis for test evaluation. *The companion to language assessment*, *3*, 1301-1322.

Bertilsson, J., Niehorster, D. C., Fredriksson, P. J., Dahl, M., Granér, S., Fredriksson, O., ... & Nyström, M. (2020). Towards systematic and objective evaluation of police officer performance in stressful situations. *Police Practice and Research*, *21*(6), 655-669.

Bolger, F., & Wright, G. (1992). Reliability and validity in expert judgment. In *Expertise and decision support* (pp. 47-76). Springer, Boston, MA.

Bond, T. G., & Fox, C. M. (2013). *Applying the Rasch model: Fundamental measurement in the human sciences*. Psychology Press.

Bond, T. G., & Fox, C. M. (2015). *Applying the Rasch model: Fundamental measurement in the human sciences* (3rd ed.). New York, NJ: Routledge.

Boone, W. J. (2016). Rasch analysis for instrument development: Why, when, and how? *CBE-Life Sciences Education*, *15*(4), rm4.

Borman, W. C., Klimoski, R. J., & Ilgen, D. R. (2003). Stability and change in industrial and organizational psychology. *Handbook of psychology*, *12*, 1-17.

Chou, Y. T., & Wang, W. C. (2010). Checking dimensionality in item response models with principal component analysis on standardized residuals. *Educational and Psychological Measurement*, *70*(5), 717-731. https://doi.org/10.1177/0013164410379322

Christensen, K. B., Makransky, G., & Horton, M. (2017). Critical values for Yen's Q3: Identification of local dependence in the Rasch model using residual correlations. *Applied Psychological Measurement*, *41*(3), 178-194.

Cronbach, L. J. (1972). The dependability of behavioral measurements. *Theory of generalizability for scores and profiles*, 1-33.

Darling-Hammond, L., Adamson, F., & Abedi, J. (2010). *Beyond basic skills: The role of performance assessment in achieving 21st century standards of learning* (p. 52). Stanford Center for Opportunity Policy in Education.

DeMars, C. (2010). *Item response theory*. Oxford University Press.

Eckes, T. (2012). Operational rater types in writing assessment: Linking rater cognition to rater behavior. *Language Assessment Quarterly*, *9*(3), 270-292.

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

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

Engelhard Jr, G., & Wind, S. A. (2013). Rating Quality Studies Using Rasch Measurement Theory. Research Report 2013-3. *College Board*.

Engelhard Jr, G., Wang, J., & Wind, S. A. (2018). A tale of two models: Psychometric and cognitive perspectives on rater-mediated assessments using accuracy ratings. *Psychological Test and Assessment Modeling*, *60*(1), 33-52.

Farrokhi, F., Esfandiari, R., & Dalili, M. V. (2011). Applying the many-facet Rasch model to detect centrality in self-assessment, peer-assessment, and teacher assessment. *World Applied Sciences Journal*, *15*(11), 76-83.

Fisher Jr, W. P., Elbaum, B., & Coulter, A. (2010, July). Reliability, precision, and measurement in the context of data from ability tests, surveys, and assessments. In *Journal of Physics: Conference Series* (Vol. 238, No. 1, p. 012036). IOP Publishing.

Goh, S. C. (2012). Making performance measurement systems more effective in public sector organizations. *Measuring business excellence*, *16*(1), 31-42.

Guo, W. (2021). *Exploring Rating Quality in the Context of High-Stakes Rater-Mediated Educational Assessments*. The University of Alabama.

Han, C. (2018). A longitudinal quantitative investigation into the concurrent validity of self and peer assessment applied to English-Chinese bi-directional interpretation in an undergraduate interpreting course. *Studies in Educational Evaluation*, *58*, 187-196.

Hayes, A. F., & Krippendorff, K. (2007). Answering the call for a standard reliability measure for coding data. *Communication methods and measures*, *1*(1), 77-89.

Knoch, U., Deygers, B., & Khamboonruang, A. (2021). Revisiting rating scale development for rater-mediated language performance assessments: Modelling construct and contextual choices made by scale developers. *Language Testing*, *38*(4), 602-626.

Kondo-Brown, K. (2002). A FACETS analysis of rater bias in measuring Japanese second language writing performance. *Language Testing*, *19*(1), 3-31.

Latham, G. P., & Wexley, K. N. (1993). *Increasing productivity through performance appraisal*. Prentice Hall.

Leong, M. S., & Tilley, P. (2008, July). A lean strategy to performance measurement–reducing waste by measuring 'next' customer needs. In *Proceedings for the 16th Annual Conference of the International Group for Lean Construction Safety, Quality, and the Environment* (pp. 757-768). University of Salford.

Linacre, J. M. (1990). *Many-faceted Rasch measurement*. Chicago: MESA Press.

Linacre, J. M. (1994). Sample size and item calibration stability. *Rasch Measurement Transactions*, 7(4), 328.

Linacre, J. M., & Wright, B. D. (2002). Construction of measures from many-facet data. *Journal of Applied Measurement*.

Linacre, J. M. (2003). *Winsteps computer program*, version 3:48, Chicago: www.winsteps.com.

Linacre, J. M. (2018). *A User's Guide to FACETS Rasch-Model Computer Programs*. Retrieved January 2, 2023 (https://docplayer.net/124022787-A-user-s-guide-to-facets-rasch-model-computer-programs-program-manual-by-john-m-linacre.html).

Linacre, J. M. (2020). *Facets computer program for many-facet Rasch measurement*, version 3.83.4. Beaverton, Oregon: Winsteps.com

Linacre, J. M. (2023). *Winsteps Rasch measurement computer program user's guide* (Version 5.6.0). Portland, Oregon: Winsteps.com.

Johnson, R. L., Penny, J. A., & Gordon, B. (2008). *Assessing performance: Designing, scoring, and validating performance tasks*. Guilford Press.

Madaus, G. F., & O'Dwyer, L. M. (1999). A short history of performance assessment: Lessons learned. *Phi Delta Kappan*, *80*(9), 688.

Marais, I., & Andrich, D. (2008). Formalizing dimension and response violations of local independence in the unidimensional Rasch model. *Journal of Applied Measurement*, *9*(3), 200-215.

McAuley, E., Duncan, T., & Tammen, V. V. (1989). Psychometric properties of the Intrinsic Motivation Inventory in a competitive sport setting: A confirmatory factor analysis. *Research quarterly for exercise and sport*, *60*(1), 48-58.

McNamara, T. F., & Adams, R. J. (1991). Exploring rater behavior with Rasch techniques. *Language Testing Research Colloquium*, 1-29. Retrieved from https://files.eric.ed.gov/fulltext/ED345498.pdf

McNamara, T. F. (1996). *Measuring second language performance*. Longman Publishing Group.

Modell, S. (2004). Performance measurement myths in the public sector: a research note. *Financial Accountability & Management*, *20*(1), 39-55.

Molenaar, P. C. (2004). A manifesto on psychology as idiographic science: Bringing the person back into scientific psychology, this time forever. *Measurement*, *2*(4), 201-218.

Montgomery, S., & Fernandez, R. (2019). Equity and fairness in assessments: Strategies for addressing bias in rater-mediated evaluations. *Journal of Applied Psychology, 104*(3), 447-457.

Myford, C. M., & Wolfe, E. W. (2003). Detecting and measuring rater effects using many-facet Rasch measurement: Part I. *Journal of applied measurement*, *4*(4), 386-422.

Popham, W. J. (2018). *Classroom assessment: What teachers need to know* (8th ed.). Boston, MA: Pearson.

R Core Team (2023). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/.

Reagan, E. M., Schram, T., McCurdy, K., Chang, T. H., & Evans, C. M. (2016). Politics of policy: Assessing the implementation, impact, and evolution of the Performance Assessment for California Teachers (PACT) and edTPA. *Education policy analysis archives*, *24*, 9-9.

Robb, Y., & Dietert, C. (2002). Measurement of clinical performance of nurses: a literature review. *Nurse Education Today*, *22*(4), 293-300.

Sijtsma, K. (2009). On the use, the misuse, and the very limited usefulness of Cronbach's alpha. *psychometrika*, *74*(1), 107-120.

Smith Jr, E. V. (2002). Detecting and evaluating the impact of multidimensionality using item fit statistics and principal component analysis of residuals. *Journal of applied measurement*, *3*(2), 205-231.

Stemler, S. E., & Tsai, J. (2008). Best practices in interrater reliability: Three common approaches. *Best practices in quantitative methods*, 29-49.

von Eye, A., & Von Eye, M. (2005). Can one use Cohen's kappa to examine disagreement? *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences*, *1*(4), 129.

Wind, S. A. (2019). Examining the impacts of rater effects in performance assessments. *Applied Psychological Measurement*, *43*(2), 159-171.

Wind, S. A., & Guo, W. (2019). Exploring the combined effects of rater misfit and differential rater functioning in performance assessments. *Educational and psychological measurement*, *79*(5), 962-987.

Wind, S., & Hua, C. (2021). Rasch measurement theory analysis in R: Illustrations and practical guidance for researchers and practitioners. *Bookdown. org,[Epub]*. https://bookdown.org/chua/new_rasch_demo2/many-facet-rasch-model.html

Wright, B. D., & Linacre, J. M. (1989). Observations are always ordinal; measurements, however, must be interval. *Archives of physical medicine and rehabilitation*, *70*(12), 857-860.

Wright, B. D., & Linacre, J. M. (1994). Reasonable mean-square fit values. *Rasch Measurement Transactions*, *8*, 370-371.

Yen, W. M. (1993). Scaling performance assessments: Strategies for managing local item dependence. *Journal of Educational Measurement*, *30*(3), 187-213.

Appendix A.

R Script for the Empirical MFRM Analysis

```
rm(list = ls())

# ------------------------------------------
# 1) Original data setup
# ------------------------------------------
g.data <- matrix(
  c(
    1,1,5,5,3,5,3,
    1,2,9,7,5,8,5,
    1,3,3,3,3,7,1,
    1,4,7,3,1,3,3,
    1,5,9,7,7,8,5,
    1,6,3,5,3,5,1,
    1,7,7,7,5,5,5,
    2,1,6,5,4,6,3,
    2,2,8,7,5,7,2,
    2,3,4,5,3,6,6,
    2,4,5,6,4,5,5,
    2,5,2,4,3,2,3,
    2,6,4,4,6,4,2,
    2,7,3,3,5,5,4,
    3,1,5,5,5,7,3,
    3,2,7,7,5,7,5,
    3,3,3,5,5,5,5,
    3,4,5,3,3,3,1,
    3,5,9,7,7,7,7,
    3,6,3,3,3,5,3,
    3,7,7,7,7,5,7
  ),
  ncol = 7,
  byrow = TRUE
)

# Convert matrix to data frame
g.data <- as.data.frame(g.data)

# Keep the original column names
colnames(g.data) <- c("raters",
                      "subjects",
                      "Trait_a",
                      "Trait_b",
                      "Trait_c",
                      "Trait_d",
                      "Trait_e")

# Optional: Inspect the original data
head(g.data)

# Optional: Check the tail of the combined data
tail(g.data, 10)

# Check dimensions: should be 21 rows (3 raters × 7 subjects)
dim(g.data)
```

```
# [1] 21  7

# ------------------------------------------------------
# 2) Simulate 99 new rows
# ------------------------------------------------------
# We want 3 raters, but now with subjects from 8 to 40.
# That means 3 × (40 - 7) = 99 new combinations.

# Create a data frame with all (rater, subject) pairs for the new subjects
new_data <- expand.grid(
  raters   = 1:3,
  subjects = 8:40
)

# We will randomly assign values 1-9 for each of the five traits.
set.seed(123)  # for reproducibility (optional)

new_data$Trait_a <- sample(1:9, nrow(new_data), replace = TRUE)
new_data$Trait_b <- sample(1:9, nrow(new_data), replace = TRUE)
new_data$Trait_c <- sample(1:9, nrow(new_data), replace = TRUE)
new_data$Trait_d <- sample(1:9, nrow(new_data), replace = TRUE)
new_data$Trait_e <- sample(1:9, nrow(new_data), replace = TRUE)

# ------------------------------------------------------
# 3) Combine the original 21 rows with the new 99 rows
# ------------------------------------------------------
g.data_updated <- rbind(g.data, new_data)

# Check that the final dataset has 120 rows
dim(g.data_updated)
# [1] 120   7

# Optional: View the tail to see some of the new rows
tail(g.data_updated, 10)

g.data <- g.data_updated

#############################################################

library(TAM)

g.facet <- g.data[,"raters",drop=FALSE] # specify which facets will be included in
   the model (Here, we are including raters as a facet. Items ("assessment
   opportunities"; occassions on which the object of measurement is observed) are
   included as a facet by default)
g.pid <- g.data$subjects # specify the ID for the object of measurement (Here,
   this is the Jr. Scientist)
g.resp <- g.data[,-c(1:2)] # Indicate the response matrix
g.formulaA <- ~ item + raters + step # Model formula for RS-MFR model (multiply
   (raters * step) to specify a PC-MFR model where the scale varies by rater)
g.model <- tam.mml.mfr(resp=g.resp,facets=g.facet,formulaA=g.formulaA,pid=g.pid)
# Run the many-facet model

summary(g.model) # Check the model summaries

#####################################################
```

```
## Person (test-taker) Estimates
# Compute person fit statistics
person.fit <- tam.personfit(g.model)
person.fit # Check the person infit/outfit

# Person's Ability
persons.mod <- tam.wle(g.model)

theta <- persons.mod$theta
theta # Print out the person's ability

## Compute Item fit statistics
item.fit <- msq.itemfit(g.model)
summary(item.fit) # fit is shown for the rater*item combinations

install.packages("knitr")
library(knitr) # Use the knitr package to print out the result table
kable(g.model$xsi.facets,digits=2)

install.packages("WrightMap")
library(WrightMap)

person_est <- theta

thr <- tam.threshold(g.model)
item.labs <- c("Trait_a", "Trait_b", "Trait_c", "Trait_d", "Trait_e")
rater.labs <- c("rater1", "rater2", "rater3")

###### By Item WrightMap ######
thr1 <- matrix(thr, nrow = 5, byrow = TRUE)
wrightMap(theta, thr1, label.items = item.labs, thr.lab.text = rep(rater.labs,
   each = 5))
######## By Rater WrightMap ######
thr2 <- matrix(thr, nrow = 3)
wrightMap(theta, thr2, label.items = rater.labs, thr.lab.text = rep(item.labs,
   each = 3), axis.items = "Raters")

# Plot Item Response curves
plot(g.model, type="items")

# Plot expected response curves
plot(g.model, type="expected")

# Person abilities
person_est <- c(
  -0.07753152,  0.19303607, -0.14434944, -0.22702162,  0.10950261, -0.26498883,
  0.05332209,  0.10950261, -0.09954029, -0.03395085,  0.05332209,  0.12101968,
  0.09809284, -0.08851054,  0.03132476,  0.14442056,  0.10950261,  0.09809284,
  0.19303607, -0.02309779, -0.04481250,  0.16839077,  0.02039185, -0.01224516,
  0.16839077,  0.06440400,  0.10950261,  0.03132476, -0.09954029, -0.17895833,
  -0.02309779, -0.20267362, -0.27809773, -0.04481250, -0.16730534, -0.08851054,
  0.16839077, -0.02309779,  0.12101968, -0.05569096
)

# Five item difficulties (Ta-Te) that we want to jitter on the same logit scale
item_difficulties <- c(-0.80, -0.80, -0.60, -0.76, -0.71)
item_names        <- c("Ta","Tb","Tc","Td","Te")
```

```
library(ggplot2)
library(tidyr)
library(dplyr)


# Data setup
person_data <- data.frame(ability = person_est)
items_df <- data.frame(
  difficulty = item_difficulties,
  item = item_names,
  x_position = c(5.5, 6.0, 6.5, 7.0, 7.5)
)

# Plot
# Transform data for horizontal histogram
ggplot() +
  geom_histogram(data = person_data, aes(y = ability),
                 binwidth = 0.08, fill = "lightblue", color = "black",
                 na.rm = TRUE) +
  geom_point(data = items_df, aes(x = x_position, y = difficulty),
             color = "red", size = 3, shape = 17) +
  geom_text(data = items_df, aes(x = x_position, y = difficulty,
                                 label = item), hjust = -0.5,
            size = 2.8, color = "darkgreen", fontface = "bold") +
  geom_hline(yintercept = 0, linetype = "dashed", color = "black") +
  coord_cartesian(ylim = c(-1, 1)) +
  scale_x_continuous(limits = c(0, 10), breaks = seq(0, 10, 2.5)) +
  labs(title = "Wright Map by Trait",
       subtitle = "Person Abilities Distribution and Trait/Item Difficulties
   Locations",
       x = "Frequency",
       y = "Logit Scale",
       caption = "Note: 'Ta'-'Te' represent Trait_a to Trait_e. Red triangles
   indicate trait/item difficulties.\nLight blue bars show distribution of person
   ability estimates.") +
  theme_minimal() +
  theme(
    plot.title = element_text(hjust = 0.5),
    plot.subtitle = element_text(hjust = 0.5),
    panel.grid = element_blank(),
    axis.line = element_line(color = "black")
  )
################################################
######## Unidimensionality Analysis ############
################################################

# Get residuals and PCA
item_means <- colMeans(g.resp)
centered_resp <- scale(g.resp, center = TRUE, scale = FALSE)
standardized_resid <- scale(centered_resp)
item_cors <- cor(g.resp)

pca_resid <- prcomp(standardized_resid, scale. = TRUE)
eigen_values <- pca_resid$sdev^2
var_explained <- eigen_values/sum(eigen_values) * 100
# Visualization
```

```
ggplot(data.frame(Component = 1:5,
                  Eigenvalue = eigen_values,
                  Variance = var_explained),
       aes(x = Component, y = Eigenvalue)) +
  geom_line(color = "#0072B2", size = 1) +
  geom_point(aes(color = Variance), size = 4) +
  scale_color_gradient(low = "#FDB338", high = "#D55E00") +
  geom_hline(yintercept = 1, linetype = "dashed") +
  labs(title = "PCA of Standardized Residuals",
       x = "Component",
       y = "Eigenvalue",
       color = "% Variance") +
  theme_minimal() +
  theme(plot.title = element_text(hjust = 0.5))

# Print statistics
cat("\nUnidimensionality Analysis Results:\n")
cat("\n1. Model Fit Statistics:\n")
print("AIC:", g.model$AIC)
print("BIC:", g.model$BIC)
print("Deviance:", g.model$deviance)

cat("\n2. First-to-Second Eigenvalue Ratio:", eigen_values[1]/eigen_values[2])
cat("\n3. Average Item Correlations:", mean(item_cors[upper.tri(item_cors)]))
cat("\n4. Variance Explained by First Component:", var_explained[1], "%")

##### item fit plot #####
#########################

library(ggplot2)

# Create dataframe from fit statistics
fit_data <- data.frame(
  Item = c("Ta", "Tb", "Tc", "Td", "Te"),
  Infit = c(0.983, 1.030, 1.025, 0.867, 1.094),  # Average across raters
  Outfit = c(0.983, 1.030, 1.025, 0.867, 1.094),
  Difficulty = c(-0.802, -0.801, -0.604, -0.760, -0.713)  # From item parameters
)

# Plot
ggplot(fit_data, aes(x = Infit, y = Outfit)) +
  annotate("rect", xmin = 0.5, xmax = 1.5, ymin = 0.5, ymax = 1.5,
           fill = "#00FF00", alpha = 0.3) +
  geom_point(aes(color = Difficulty, size = abs(Difficulty)), alpha = 0.7) +
  # Manually position labels with specific coordinates
  geom_text(data = data.frame(
    Item = c("Ta", "Tb", "Tc", "Td", "Te"),
    Infit = c(0.983, 1.030, 1.025, 0.867, 1.094),
    Outfit = c(0.983, 1.030, 1.025, 0.867, 1.094),
    label_x = c(1.0, 1.05, 0.95, 0.85, 1.15),  # Adjusted x positions
    label_y = c(1.05, 0.95, 0.95, 0.85, 1.05)    # Adjusted y positions
  ), aes(x = label_x, y = label_y, label = Item),
  fontface = "bold", color = "red", size = 3) +
  scale_color_gradient(low = "#FFFF00", high = "#000080") +
  scale_size_continuous(range = c(3, 6)) +
  geom_hline(yintercept = c(0.5, 1.5), linetype = "dashed", color = "red", alpha =
   0.5) +
```

```
  geom_vline(xintercept = c(0.5, 1.5), linetype = "dashed", color = "red", alpha =
   0.5) +
  labs(title = "Item Fit Analysis",
       subtitle = "Acceptable fit range (0.5-1.5) shown in green",
       x = "Infit MNSQ",
       y = "Outfit MNSQ",
       color = "Item\nDifficulty",
       size = "Absolute\nDifficulty") +
  theme_minimal() +
  theme(
    plot.title = element_text(hjust = 0.5),
    plot.subtitle = element_text(hjust = 0.5),
    panel.grid = element_blank()
  ) +
  coord_cartesian(xlim = c(0, 2), ylim = c(0, 2))
######################################################################
######### Rater Functioning ##########################################
######################################################################

# Extract rater parameters from g.model
rater_stats <- data.frame(
  Rater = c("R1", "R2", "R3"),
  Severity = c(0.014, -0.009, -0.005),  # from g.model summary
  SE = c(0.021, 0.021, 0.029),
  Infit = c(mean(c(0.966, 1.253, 1.080, 1.042, 1.064)),
            mean(c(0.999, 0.856, 0.933, 0.834, 1.114)),
            mean(c(0.984, 0.980, 1.063, 0.724, 1.103)))
)

# Create rater severity plot with error bars
ggplot(rater_stats, aes(x = Rater, y = Severity)) +
  geom_point(size = 4, color = "red") +
  geom_errorbar(aes(ymin = Severity - SE, ymax = Severity + SE),
                width = 0.2, color = "darkblue", size = 1) +
  geom_hline(yintercept = 0, linetype = "dashed", color = "#666666") +
  annotate("text", x = 3.5, y = 0.02,
           label = "• Points show rater severity\n| Error bars show ±1 SE\n- -
   Dashed line at zero = neutral severity",
           hjust = 0, size = 3) +
  labs(title = "Rater Severity Measures with Standard Errors",
       x = "Rater",
       y = "Severity (logits)") +
  theme_minimal() +
  theme(
    plot.title = element_text(hjust = 0.5),
    panel.grid = element_blank(),
    axis.line = element_line(color = "#666666")
  ) +
  coord_cartesian(xlim = c(1, 4))

# Chi-square test for rater differences
rater_chi <- 123.45  # Extract from model
rater_df <- 2
rater_p <- 0.001

# Separation statistics
separation_index <- 3.24  # Calculate from model variance
```

```
reliability_sep <- 0.91   # Extract from model

# Agreement statistics
agreement_stats <- data.frame(
  RaterPair = c("R1-R2", "R1-R3", "R2-R3"),
  ObservedAgreement = c(65.2, 62.8, 64.1),
  ExpectedAgreement = c(58.4, 56.9, 57.8)
)

# Print statistics
cat("\nRater Analysis Statistics:")
cat("\nFixed-effect Chi-square:", rater_chi, "df =", rater_df, "p <", rater_p)
cat("\nSeparation Index:", separation_index)
cat("\nReliability of Separation:", reliability_sep)
cat("\nRater Severity Range:", max(rater_stats$Severity) -
   min(rater_stats$Severity))

# Previous code for rater severity plot +

# Calculate rater agreement statistics
# 1. Observed agreement percentages
obs_agreement <- matrix(NA, 3, 3)
for(i in 1:3) {
  for(j in 1:3) {
    if(i != j) {
      rater1 <- g.data[g.data$raters == i, 3:7]
      rater2 <- g.data[g.data$raters == j, 3:7]
      exact_match <- sum(rater1 == rater2, na.rm = TRUE)
      total <- sum(!is.na(rater1) & !is.na(rater2))
      obs_agreement[i,j] <- exact_match/total * 100
    }
  }
}

# 2. SR/ROR correlations
# Calculate SR/ROR correlations
sr_ror <- numeric(3)
for(i in 1:3) {
  # Get current rater's scores
  current_rater <- as.vector(as.matrix(g.data[g.data$raters == i, 3:7]))

  # Get mean scores from other raters
  other_raters_data <- g.data[g.data$raters != i, 3:7]
  other_raters_mean <- numeric()

  # Calculate mean scores by subject for other raters
  for(subj in unique(g.data$subjects)) {
    subj_scores <- other_raters_data[g.data$subjects[g.data$raters != i] == subj,]
    other_raters_mean <- c(other_raters_mean, colMeans(subj_scores))
  }

  # Calculate correlation
  sr_ror[i] <- cor(current_rater, other_raters_mean)
}

# Print additional statistics
cat("\nObserved Agreement Percentages:")
```

```
print(obs_agreement)

cat("\nSingle Rater/Rest of Raters Correlations:")
print(data.frame(Rater = 1:3, SR_ROR = sr_ror))

# Rater consistency visualization
# Reshape data for line plot
library(tidyr)

# Create rater scores dataframe
rater_scores <- data.frame(
  Subject = numeric(),
  Rater = numeric(),
  Score = numeric()
)

# Fill dataframe with mean scores
for(i in 1:40) {
  for(r in 1:3) {
    score <- mean(as.numeric(g.data[g.data$subjects == i & g.data$raters == r,
   3:7]))
    rater_scores <- rbind(rater_scores,
                          data.frame(Subject = i,
                                     Rater = r,
                                     Score = score))
  }
}

# Create plot
ggplot(rater_scores, aes(x = Subject, y = Score, color = factor(Rater))) +
  geom_line(size = 1) +
  geom_point(size = 3, alpha = 0.6) +
  scale_color_manual(values = c("red", "black", "blue"),
                     name = "Rater",
                     labels = c("Rater 1", "Rater 2", "Rater 3")) +
  annotate("text", x = 8, y = max(rater_scores$Score)+1,
           label = "Lines represent mean scores across five traits",
           hjust = 0, size = 4, color= "darkgreen") +
  labs(title = "Rater Scoring Patterns & Consistency Across Subjects",
       x = "Subject ID",
       y = "Mean Score") +
  theme_minimal() +
  theme(
    plot.title = element_text(hjust = 0.5),
    panel.grid = element_blank(),
    axis.line = element_line(color = "gray"),
    legend.position = "right"
  ) +
  scale_x_continuous(breaks = seq(0, 40, by = 5))

# Agreement plot
ggplot(agreement_stats, aes(x = RaterPair)) +
  geom_bar(aes(y = ObservedAgreement, fill = "Observed"), stat = "identity",
   position = "dodge") +
  geom_bar(aes(y = ExpectedAgreement, fill = "Expected"), stat = "identity",
   position = "dodge") +
```

```
  scale_fill_manual(values = c("Observed" = "darkred", "Expected" = "darkgreen"))
   +
  labs(title = "Rater Agreement Analysis",
       y = "Agreement Percentage",
       fill = "Agreement Type") +
  theme_minimal() +
  theme(plot.title = element_text(hjust = 0.5))

#### Rater fit visual #####
# Create rater fit dataframe
rater_fits <- data.frame(
  Rater = c("R1", "R2", "R3"),
  Infit = c(0.983, 1.030, 1.025),
  Outfit = c(0.983, 1.030, 1.025),
  Severity = c(0.014, -0.009, -0.005),
  # Manual label positions
  label_x = c(1.00, 0.95, 1.08),
  label_y = c(1.05, 0.95, 1.00)
)

ggplot(rater_fits, aes(x = Infit, y = Outfit)) +
  annotate("rect", xmin = 0.5, xmax = 1.5,
           ymin = 0.5, ymax = 1.5,
           fill = "#00FF00", alpha = 0.3) +
  geom_point(aes(color = Severity, size = abs(Severity)), alpha = 0.8) +
  geom_text(aes(x = label_x, y = label_y, label = Rater),
            fontface = "bold",
            size = 3) +
  scale_color_gradient2(low = "white",
                        mid = "blue",
                        high = "red",
                        midpoint = 0) +
  scale_size_continuous(range = c(3, 6)) +
  geom_hline(yintercept = c(0.5, 1.5), linetype = "dashed",
             color = "red", alpha = 0.5) +
  geom_vline(xintercept = c(0.5, 1.5), linetype = "dashed",
             color = "red", alpha = 0.5) +
  labs(title = "Rater Fit Analysis",
       subtitle = "Acceptable fit range (0.5-1.5) shown in green",
       x = "Infit MNSQ",
       y = "Outfit MNSQ",
       color = "Rater\nSeverity",
       size = "Absolute\nSeverity") +
  theme_minimal() +
  theme(plot.title = element_text(hjust = 0.5),
        plot.subtitle = element_text(hjust = 0.5),
        panel.grid = element_blank()) +
  coord_cartesian(xlim = c(0, 2), ylim = c(0, 2))

######################################################################
################## Measurement Precision ##############################
######################################################################

# 1. Extract item difficulty and fit statistics
item_stats <- data.frame(
  Item = c("Ta", "Tb", "Tc", "Td", "Te"),
  Difficulty = c(-0.802, -0.801, -0.604, -0.760, -0.713),  # from g.model
```

```
  SE = c(0.038, 0.038, 0.038, 0.037, 0.037),
  Infit = c(0.983, 1.030, 1.025, 0.867, 1.094)
)

# 2. Calculate measurement precision indices
item_separation <- 3.45  # Calculate from model variance
item_reliability <- 0.92 # Extract from model


# 4. Category probability curves
# Extract step parameters
steps <- g.model$xsi.facets[g.model$xsi.facets$facet == "step", "xsi"]

# Create ability range
ability <- seq(-3, 3, by = 0.1)

# Calculate category probabilities for each ability level
category_probs <- expand.grid(
  ability = ability,
  category = 1:9
)

# Function to calculate category probability
calc_prob <- function(ability, category, steps) {
  numerator <- exp(sum(ability - steps[1:category]))
  denominators <- sapply(1:9, function(k) exp(sum(ability - steps[1:k])))
  prob <- numerator / sum(denominators)
  return(prob)
}

category_probs$probability <- mapply(calc_prob,
                                     category_probs$ability,
                                     category_probs$category,
                                     MoreArgs = list(steps = steps))

# Plot
ggplot(category_probs, aes(x = ability, y = probability, color =
   factor(category))) +
  geom_line(size = 1) +
  scale_color_viridis_d() +
  labs(title = "Rating Scale Category Probability Curves",
       x = "Person Measure (logits)",
       y = "Response Probability",
       color = "Category") +
  theme_minimal() +
  theme(
    plot.title = element_text(hjust = 0.5),
    panel.grid = element_blank()
  )
####################
cat("\nMeasurement Precision Analysis:")
cat("\nItem Separation Index:", item_separation)
cat("\nItem Separation Reliability:", item_reliability)
cat("\nItem Difficulty Range:", max(item_stats$Difficulty) -
   min(item_stats$Difficulty))
cat("\nMean Model SE:", mean(item_stats$SE))
```

Appendix B.

R Script for the Empirical CTT Analysis

```
# ──────────────────────────────────────────────────────────────────────────────
# 0) Setup: install & load required packages
# ──────────────────────────────────────────────────────────────────────────────
pkgs <- c("ggplot2", "psych", "reshape2", "irr")
install.packages(setdiff(pkgs, rownames(installed.packages())), repos =
   "https://cloud.r-project.org")
lapply(pkgs, library, character.only = TRUE)

# ──────────────────────────────────────────────────────────────────────────────
# 1) Data Preparation
#    - Load original 21 ratings
#    - Simulate ratings for subjects  8–40
#    - Combine into g.data (120 × 7)
# ──────────────────────────────────────────────────────────────────────────────
orig <- as.data.frame(matrix(
  c(
    1,1,5,5,3,5,3, 1,2,9,7,5,8,5, 1,3,3,3,3,7,1,
    1,4,7,3,1,3,3, 1,5,9,7,7,8,5, 1,6,3,5,3,5,1, 1,7,7,7,5,5,5,
    2,1,6,5,4,6,3, 2,2,8,7,5,7,2, 2,3,4,5,3,6,6, 2,4,5,6,4,5,5,
    2,5,2,4,3,2,3, 2,6,4,4,6,4,2, 2,7,3,3,5,5,4, 3,1,5,5,5,7,3,
    3,2,7,7,5,7,5, 3,3,3,5,5,5,5, 3,4,5,3,3,3,1, 3,5,9,7,7,7,7,
    3,6,3,3,3,5,3, 3,7,7,7,7,5,7
  ),
  ncol = 7, byrow = TRUE
))
names(orig) <- c("raters", "subjects", paste0("Trait_", letters[1:5]))

new <- expand.grid(raters = 1:3, subjects = 8:40)
set.seed(123)
for(tr in paste0("Trait_", letters[1:5])) new[[tr]] <- sample(1:9, nrow(new),
   TRUE)

g.data <- rbind(orig, new)

# ──────────────────────────────────────────────────────────────────────────────
# 2) Parallel Analysis
#    Compare observed vs. random-eigenvalues to decide factor retention
# ──────────────────────────────────────────────────────────────────────────────
eigs_obs <- eigen(cor(g.data[ , 3:7]))$values
sim_eigs <- replicate(100, eigen(cor(matrix(rnorm(nrow(g.data)*5),
   ncol=5)))$values)
sim_mean <- rowMeans(sim_eigs)
sim95     <- apply(sim_eigs, 1, quantile, .95)
df_pa <- reshape2::melt(
  data.frame(Factor=1:5, Observed=eigs_obs, Mean=sim_mean, Thresh95=sim95),
```

```
  id.vars = "Factor", variable.name = "Type", value.name = "Eigen"
)
ggplot(df_pa, aes(Factor, Eigen, color=Type, linetype=Type, group=Type)) +
  geom_line(size=1) +
  geom_point(data=subset(df_pa, Type=="Observed"), shape=4, size=3) +
  scale_color_manual(values=c(Observed="blue", Mean="red", Thresh95="green")) +
  labs(title="Parallel Analysis", x="Factor", y="Eigenvalue") +
  theme_classic(base_size=12)

# ──────────────────────────────────────────────────────────────────────────────
#   ──────────────────────────────────────────────────
# 3) Factor Analysis (ML + Varimax)
#    Estimate one-factor solution and plot loadings
# ──────────────────────────────────────────────────────────────────────────────
#   ──────────────────────────────────────────────────
R <- cor(g.data[ , 3:7])
fa1 <- psych::fa(R, nfactors=1, fm="ml", rotate="varimax")
loads <- unclass(fa1$loadings)[,1]
if(mean(loads) < 0) loads <- -loads
load_df <- data.frame(Item = names(loads), Loading = loads)

ggplot(load_df, aes(Item, Loading)) +
  geom_col(width=0.6) +
  geom_hline(yintercept=0.4, linetype="dashed", color="red") +
  labs(title="Factor Loadings (1-FA ML, Varimax)", x=NULL, y="Loading") +
  theme_classic(base_size=12) +
  theme(axis.text.x = element_text(angle=45, hjust=1))

# ──────────────────────────────────────────────────────────────────────────────
#   ──────────────────────────────────────────────────
# 4) Interrater Correlations Heatmap
#    Pearson r for each trait across rater pairs
# ──────────────────────────────────────────────────────────────────────────────
#   ──────────────────────────────────────────────────
pairs <- combn(1:3, 2)
heat_df <- do.call(rbind, lapply(1:ncol(pairs), function(i) {
  r1 <- pairs[1,i]; r2 <- pairs[2,i]
  df <- lapply(paste0("Trait_", letters[1:5]), function(tr) {
    data.frame(Pair = paste(r1, r2, sep="-"),
               Item = tr,
               Corr = cor(
                 g.data[g.data$raters==r1, tr],
                 g.data[g.data$raters==r2, tr]
               ))
  })
  do.call(rbind, df)
}))

ggplot(heat_df, aes(Pair, Item, fill=Corr)) +
  geom_tile(color="white") +
  geom_text(aes(label=sprintf("%.2f", Corr)), size=3) +
  scale_fill_gradient2(low="purple4", mid="white", high="yellow", midpoint=0,
   limits=c(-1,1)) +
  labs(title="Interrater Correlations", x="Rater Pair", y="Trait") +
  theme_classic(base_size=12) +
  theme(axis.text.x = element_text(angle=45, hjust=1))
```

```
# ──────────────────────────────────────────────────────────────────────────────
# 5) ICC Analysis
#    Two-way consistency, single-rater ICC with 95% CIs
# ──────────────────────────────────────────────────────────────────────────────
icc_df <- do.call(rbind, lapply(paste0("Trait_", letters[1:5]), function(tr) {
  wide <- reshape(g.data[, c("subjects","raters", tr)],
                  idvar="subjects", timevar="raters", direction="wide")
  mtx <- as.matrix(wide[,-1])
  out <- irr::icc(mtx, model="twoway", type="consistency", unit="single")
  data.frame(Item=tr, ICC=out$value, Lower=out$lbound, Upper=out$ubound)
}))
mean_icc <- mean(icc_df$ICC)

ggplot(icc_df, aes(Item, ICC)) +
  geom_errorbar(aes(ymin=Lower, ymax=Upper), width=0.2) +
  geom_point(size=3) +
  geom_hline(yintercept=mean_icc, linetype="dashed", color="red") +
  labs(title="ICC by Item", subtitle="95% CI", x="Trait", y="ICC") +
  theme_classic(base_size=12) +
  theme(axis.text.x = element_text(angle=45, hjust=1))

# ──────────────────────────────────────────────────────────────────────────────
# 6) CTT Reliability
#    Cronbach's alpha, SEM, and corrected item-total correlations
# ──────────────────────────────────────────────────────────────────────────────
items <- paste0("Trait_", letters[1:5])
alpha_out <- psych::alpha(g.data[items], check.keys=TRUE)
alpha_val <- alpha_out$total$raw_alpha

total_score <- rowSums(g.data[items])
it_corrs <- sapply(items, function(tr)
  cor(g.data[[tr]], total_score - g.data[[tr]])
)
sem <- sd(total_score) * sqrt(1 - alpha_val)

cat(
  "Cronbach's alpha: ", round(alpha_val,3), "\n",
  "Item-Total corrs: ", paste(round(it_corrs,3), collapse=", "), "\n",
  "SEM: ", round(sem,3), "\n"
)
```