Psychological Test and Assessment Modeling, Volume 64, 2022 (3), 216-252

# Exploring Rater Quality in Rater-Mediated Assessment Using the Non-parametric Item Characteristic Curve Estimation

*Farshad Effatpanah*[1] *& Purya Baghaei*[2]

**Abstract**

A large number of researchers have explored the use of non-parametric item response theory (IRT) models, including Mokken scale analysis (Mokken, 1971), for inspecting rating quality in the context of performance assessment. Unlike parametric IRT models, such as Many-Facet Rasch Model (Linacre, 1989), non-parametric IRT models do not entail logistic transformations of ordinal ratings into interval scales neither do they impose any constraints on the form of item response functions. A disregarded method for examining raters' scoring patterns is the non-parametric item characteristic curve estimation using kernel smoothing approach (Ramsay, 1991) which provides, without giving numerical values, graphical representations for identifying any unsystematic patterns across various levels of the latent trait. The purpose of this study is to use the non-parametric item characteristic curve estimation method for modeling and examining the scoring patterns of raters. To this end, the writing performance of 217 English as a foreign language (EFL) examinees were analyzed. The results of rater characteristic curves, tetrahedron simplex plots, QQ-plot, and kernel density functions across gender sub-groups showed that different exploratory plots derived from the non-parametric estimation of item characteristic curves using kernel smoothing approach can identify various rater effects and provide valuable diagnostic information for examining rating quality and exploring rating patterns, although the interpretation of some graphs are subjective. The implications of the findings for rater training and monitoring are discussed.

**Keywords:** Non-parametric estimation of item characteristic curves, kernel smoothing, rating patterns, scoring validity

[1] English Department, Tabaran Institute of Higher Education, Mashhad, Iran

[2] English Department, Islamic Azad University, Mashhad Branch, Mashhad, Iran. *Correspondence concerning this article should be addressed to* Purya Baghaei, Islamic Azad University, Ostad Yusofi St., 9187147578, Mashhad, Iran. E-mail: pbaghaei@mshdiau.ac.ir

## 1. Introduction

Educational performance assessments have become increasingly widespread in international high-stakes assessments, such as the Graduate Record Exam (GRE), the Programme for International Student Assessment (PISA), Trends in International Mathematics and Science Study (TIMSS), the Test of English as a Foreign Language (TOEFL), and the International English Language Testing System (IELTS). These assessments are widely used to reflect the ability of students in applying different knowledge and skills to complete a set of educationally meaningful tasks (Lane, 2016). A common type of performance assessment in educational contexts is writing assessment which requires rater judgments (e.g., rater-mediated assessments, Eckes, 2015; Engelhard, 2013). In rater-mediated assessments, human raters typically use some kind of (ordinal or multiple-category) rating scales to assess student responses to multiple complicated tasks and express their interpretation concerning the quality of student performances. As raters should score the performance of students, the central role of raters engenders an extra layer of complexity to the rating process (Kuiken & Vedder, 2014).

A wide array of parametric item response theory (IRT; Yen & Fitzpatrick, 2006) and the various forms of Rasch model (Kubinger, 1989; Wright & Masters, 1982) are typically used to provide a detailed analysis of rater-mediated assessments. The models include the Graded Response Model (Samejima, 1968), the Rating Scale Model (Andrich, 1978A, 1978b), the Partial Credit Model (Masters, 1982), the Many-Facet Rasch Model (Linacre, 1989), the Partial Credit Model versions of Many-Facet Rasch Model (Linacre, 1989), the Generalized Partial Credit Model (Muraki, 1992, 1997), and the Hierarchical Rater Models (De Carlo, Kim, & Johnson, 2011; Lu & Wang, 2006; Patz, Junker, Johnson, & Mariano, 1999). Although these models can calibrate different raters and task characteristics and provide valuable information about these factors, they rely on a set of strict assumptions which require practitioners to have sufficient knowledge in working with these models. However, Lei, Dunbar, and Kolen (2004) argue that as a large number of test developers and practitioners do not receive sufficient training in psychometrics and test theory, they are not able to interpret and explain the numerical values obtained from the analysis of parametric models. For this reason, some researchers have suggested the use of non-parametric IRT models for analyzing rater-mediated assessments (Wind, 2019a; Wind & Engelhard, 2016; Wind & Schumacker, 2017). Graphical displays of non-parametric IRT models could be easier for practitioners to utilize and elucidate.

A neglected method for analyzing rater-mediated assessments is the non-parametric estimation of item characteristic curves based on kernel smoothing approaches developed by Ramsay (1991). This method is only based on visual illustrations of item characteristics (e.g., item difficulty and item discrimination) in a scale and offers diagnostic information about the functioning of the items and the test. The purpose of this study is to apply the non-parametric estimation of item characteristic curves to a writing data to investigate whether this approach would be effective for evaluating

rater-mediated assessments and in particular, analyzing patterns of individual raters in scoring writing performance.

## 2. Background

### 2.1 Rater Effects and Rater Training

As the results of ratings provide rich information about students' ability for further inferences and decisions, a critical concern in educational rater-mediated performance assessments is the quality of rater scorings or the way raters do the ratings (Wind & Engelhard, 2016). It has been well-established that raters should be able to yield an appropriate and a consistent judgmental process of evaluating student responses. However, due to the complexity and subjectivity of the rating process, raters are liable to different sources of random errors, systematic biases, and idiosyncrasies, known as rater effects (Myford & Wolfe, 2003). Put differently, raters are systematically influenced by a number of extraneous irrelevant factors (e.g., construct irrelevant variance) which can distort the pattern of rating, compromise rating validity, and jeopardize the validity of score interpretations and uses (Mesick, 1989; Myford & Wolfe, 2004). The most common types of rater effects are rater severity/leniency, central tendency/extremity, randomness, rater accuracy/inaccuracy, halo, and systematic biases, e.g., differential rater functioning (Myford & Wolfe, 2003). According to Shaw and Weir (2007), "scoring validity is criterial because if we cannot depend on the rating of exam scripts, it matters little that the tasks we develop are potentially valid' (p. 143).

To support the validity of ratings, researchers have frequently used several procedures in operational settings such as rater training and monitoring to mitigate rater effects and increase rating quality (Engelhard & Myford, 2003; Wolfe, Chiu, & Myford, 2000). Prior to rating, most large-scale assessment and test programs often provide interactive training sessions in a variety of forms (e.g., online, webinars, and face-to-face) for potential raters to inform them about rating scales, review different aspects of writing prompts and scoring rubrics, practice rating, and get feedback on rating (Baldwin, Fowles, & Livingston, 2005; Lane & Stone, 2006). Research has shown that rater training can reduce rater variability and biases, and improve rating quality, systematicity in rater behavior, and the intra-rater consistency of the raters (Weigle, 1998). Other studies, however, argued that training sessions are likely to fail to attain inter-rater agreement, and if raters feel impelled to reach an agreement, they may neglect their own experience in the process of scoring, and thus compromise the scoring validity (Eckes, 2015). A number of researchers have further advocated rater monitoring as an effectual method to support rater training and ascertain the accuracy of raters' scores. Experienced raters, or rating leaders, can monitor the quality and the way novice raters assign a rating to students' writing during or after rating, and address potential problems to identify peculiar patterns of scoring, referred to as "*measurement disturbances*" (Wind & Schumacker, 2017, p. 1).

While it is accepted that rater training and monitoring can be effective at decreasing the influence of rater effects, they are unfortunately insufficient for examining rating quality because these methods fail to provide information for scoring leaders to identify raters with unsystematic scoring patterns (Myford & Wolfe, 2009). Instead, a variety of quality indicators relying on quantitative methods have been suggested to monitor deficiencies in rating quality. These methods are utilized to not only notify rater training and monitoring but also ensure that the outcomes of rater-mediated assessments provide accurate information about the reliability, validity, and fairness of the interpretations and uses of scoring results (Engelhard & Wind, 2018).

## 2.2 A Framework for Classifying Measurement Techniques for Writing

In his theoretical framework for evaluating rating quality, Engelhard (2008, 2013) classified measurement techniques into two major approaches, which later were extended by Wind and Peterson (2018) to the context of rater-mediated assessments: (1) the observed ratings tradition, and (2) the scaled ratings tradition. Many of the familiar indices that have been proposed for analyzing rating quality are based on the test-score or observed ratings tradition. The main focus of these methods are to split the observed ratings into true scores and errors, and examine the consistency of observed ratings, where an equal category width or interval for observed ratings is considered for ordinal rating scale categories (Wind & Peterson, 2018). The most common measurement models within this approach, which depend on inter-rater agreement and reliability, include classical test theory, regression models, generalizability theory, analysis of variance, traditional factor analysis, and structural equation modeling. Researchers have argued that although these models are widely applied in the context of rater-mediated assessments and produce important evidence of rater reliability, they are limited in the following ways: (1) the use of various group-level inter-rater consistency and inter-rater agreement, such as kappa coefficient, for the same dataset can induce inconsistent results, (2) the group-level coefficients do not offer sufficient diagnostic information about individual raters, (3) the methods do not offer information on the adherence of raters' interpretation of students performances to the measurement theories, such as invariant measurement, and (4) the existence of high reliability and agreement between raters does not necessarily indicate the accuracy of the ratings (Eckes, 2015; Wind & Peterson, 2018).

As an alternative approach to the observed rating traditions, techniques for exploring the quality of raters' scoring based on the scaled rating traditions focus on the use of latent trait models or item response theory (IRT) models to examine individual rater judgments. Using a scale with equal spaced intervals, the analysis of such models allows analysts to examine rater severity/leniency, measurement invariant, and fairness of rater-mediated assessments in accordance with expected measurement properties of a theoretical measurement framework. In fact, these models can identify those individual raters whose rating patterns diverge from the assumptions of a measurement model, and recognize areas in which further training for raters are required.

A most commonly used method within the scaled rating traditions for detecting rater effects is the application of Many-Facet Rasch Model (Linacre, 1989) which belongs to the family of parametric IRT models. For analyzing scoring patterns, parametric IRT models entail the transformation of (multiple-category or ordinal) observed ratings to interval scales and prescribe a set of strict requirements related to the mathematical shape of the relationship between the locations of individuals on the underlying latent trait and the probability of receiving a higher rating category, referred to as item response function which is graphically represented by an item characteristic curve. For parametric IRT models, several strict assumptions (e.g., monotonicity, unidimensionality, measurement invariance, the logistic ogive form of IRF, and local independence) must hold for parameter estimation in accordance with a set of fit statistics which indicate to what extent observed ratings conform to the model-expected ratings. The violations of these important assumptions yield unreliable and erroneous results for various assessment administrations. Because parametric IRT models involve a set of strict requirement, a number of researchers have argued that these models are inadequate in the measurement procedures of social and behavioral sciences, particularly the logistic transformation of ordinal ratings into interval scales, although this method is mathematically plausible (Junker & Sijtsma, 2001; Molenaar, 2001). As Wind (2020) argued,

> "meaningful application and interpretation of [parametric IRT models] can only occur when there is theoretical and empirical support for imposing a specific mathematical form on the shape of the response function (the relationship between the latent variable and the probability for ratings in each category), when the sample size can support parameter estimation for the full range of rater severity and test-taker achievement included in a particular sample, and when there is a theoretical or practical reason for using interval-level estimates to describe rater severity, test-taker achievement, and other facets in the assessment context" (125).

Several researchers, however, have used non-parametric IRT models to analyze rating quality in performance assessments. Non-parametric IRT models include most of the important requirements of parametric IRT models (e.g., monotonicity, local independence, and unidimensionality) and are less restrictive compared to parametric IRT models since they do not impose a specific shape for item response functions (Sijtsma & Meijer, 2007). Under non-parametric IRT models, item characteristic curves are directly estimated from the data, and item response functions can take any shapes (whether logistic or not). An integral requirement for non-parametric models is monotonicity or the order restriction. Within the context of a rater-mediated assessment, monotonicity refers to the probability of receiving higher rating categories with increasing levels of the expected latent variable, that is, as the levels of individuals' writing ability increase, raters should give non-decreasing ratings to the performance of individuals across increasing levels of writing ability. When raters fulfill monotonicity assumption, it is an indication that raters have interpreted the performance of individuals in the same order (Wind, 2020). According to van der Linden and Hambleton (1997), non-parametric IRT models provide more accurate item characteristic

curves which are closer to the true ones relative to parametric IRT models. Sijtsma and Molenaar (2002) argued that the use of non-parametric IRT models for evaluating rater-mediated assessments is promising because "if an IRT model is used for constructing a test, and the measurement of respondents on an *ordinal scale* is sufficient for the application envisaged, parametric models might be unduly restrictive for this purpose" (p. 15, emphasis in the original). Unlike parametric IRT models whose assumptions are untenable, non-parametric IRT models can provide valuable insight into desirable measurement characteristics such as invariant ordering of items and persons (Meijer, Tendeiro, & Wanders, 2015; Wind, 2015). A great advantage of non-IRT models over their parametric counterparts lies in their potential for exploring unsystematic patterns in the data by means of the analysis of monotonicity assumption and the evaluation of item response functions which allow practitioners to identify misfit and weak items (Sijtsma & Meijer, 2007). Many studies have already utilized Mokken Scale Analysis (Mokken, 1971), as a famous non-parametric IRT model, to evaluate rating quality (Wind, 2019a, 2019b, 2020; Wind & Engelhard, 2016; Wind & Patil, 2018; Wind & Schumacker, 2017, 2018), and their results have supported the effectiveness of non-parametric models, especially the Mokken Scale Analysis, for analyzing rating quality.

Another neglected method is the non-parametric item characteristic curve estimation using Kernel smoothing approaches (Ramsay, 1991). This method has received too little attention in educational testing and measurement. Similar to other non-parametric models, the non-parametric estimation of item characteristic curves does not prescribe a specific parametric form for item response functions. The method only provides a visual display for identifying any unsystematic patterns across levels of the latent trait.

Schumacker (2015) recently compared empirical and expected item response functions obtained from the Rasch model (Rasch, 1960) to detect any systematic errors or measurement disturbances across response patterns. The results of his study revealed that visual illustrations would be an effective method for exploring measurement disturbances that are unobservable through the use of model fit statistics. In another study, Wind and Schumacker (2017) showed that graphical methods can be used to identify measurement disturbances related to raters in the context of a rater-mediated assessment. They also found the diagnostic value of graphical illustrations for discovering measurement disturbances that are not captured through the use of numerical model-data fit indices. Furthermore, "[t]here seems to be a great reluctance by especially trained psychometricians to use graphs. We often see fit statistics and large tables full of numbers that certainly do not provide more information than graphs" (Meijer, Tendeiro, & Wanders, 2015, p. 89).

## 2.3 Non-parametric Estimation of Item Characteristic Curves

Ramsay (1991) introduced non-parametric estimation of item characteristic curve by suggesting regression methods, on the basis of kernel smoothing approaches. Compared to parametric IRT models, the shape of item response functions in this method are exploratory and data-driven. In fact, the form of item response functions are not assumed a priori. However, similar to the item response functions of parametric models, the item response functions of this method should fulfill two important assumptions: (1) they should be monotonically non-decreasing in θ. This means that the probability of getting an item right or endorsing a higher response option increases or at least remains constant with the increase level of $\theta$; and (2) they should be unidimensional, that is, item response functions should be representable on graphs, with the probability of success on the *y*-axis and the level of $\theta$ on the *x*-axis.

The non-parametric estimation of item characteristic curve offers a visual illustration of item characteristics (e.g., item difficulty and item discrimination) in a scale and gives diagnostic information about the performance of the items and the test. The analysis of various plots and curves obtained from this method can provide convenient preliminary feedback and valuable information about troublesome items in terms of monotonicity, item discrimination across different levels of the latent trait being measured, and differential item functioning (Rajlic, 2020). As a result, the method could be a useful asset in the statistical toolkit of researchers in the context of classical test theory and IRT. A great advantage of the method compared to classical test theory is that the main focus of the model is on the performance of the items and the scale at the level of item, not at the total test scores. Unlike parametric IRT models in which the evaluation of item response functions are usually not taken into consideration and a great deal of valuable information are disregarded, the non-parametric estimation of item characteristic curve only provides graphical illustrations without presenting any statistics and numerical summaries. This feature allows researchers and practitioners not only identify weak items by inspecting the item response functions but also evaluate model fit and then select the most suitable parametric model (Lee, Wollack, & Douglas, 2009; Mazza, Punzo, & McGuire, 2014). The three-parameter logistic model (Lord, 1980), for instance, might be the appropriate model if the results of the non-parametric estimation of item characteristic curve indicate that items of a test possess non-zero lower asymptotes, and also the two-parameter logistic model (2PL; Birnbaum, 1968) could be the best model if the items of a test have various slopes (Rajlic, 2020).

In the non-parametric estimation of item characteristic curve, several equally-spaced points, which are selected from the distribution of standard normal scores, known as *evaluation points*, are used to estimate the probability of getting an item right or endorsing a response category. Put simply, the probability of success is computed as the proportion of individuals who correctly responded to the item or the response option at the evaluation points. These evaluation points on the horizontal axis are then plotted against the probabilities on the vertical axis. When the points are linked, an item

response function is drawn. Kernel smoothing non-parametric regression is then utilized to smooth the form of the item response function and estimate item characteristic curves from data (Eubank, 1988; Härdle, 1990). Smoothing in statistics is typically used to make an approximate curve which tries to decrease noise and seize main patterns in the data. The aim is to estimate the values of $y$ from $x$ using a function $g(x)$. The function $g$ should be defined in such a way that it enables us to estimate the corresponding values of $y$ for any values of $x$, regardless of whether $x$ exists in the available data values or not (Ramsay, 2000). The best way is to estimate a smooth curve from which one can read off values of $y$ from any values of $x$. As Ramsay (2000) argued, smoothing is a type of local averaging used to estimate the relationship between the probability of giving a correct response or endorsing a response option and the level of the intended latent trait. A constant size, referred to as bandwidth that is used to control the size or width of the kernel around the point, for each evaluation point is selected and then a weighted average for all data points within the evaluation point and bandwidth is computed. Higher weights are given to points closer to the evaluation point. For that reason, Santor, Ramsay, and Zuroff (1994) state that the non-parametric estimation of item characteristic curve has great potential to provide better fit to the data relative to parametric models.

## 3. Data

The essays of 217 upper-intermediate English as a foreign language (EFL) university students were analyzed. There were 87 (40.1%) male and 130 (59.9%) female students. The essays were written by the examinees in their final exam in a writing course in the English Department of the Islamic Azad University of Mashhad, Iran. The writing task presented a contemporary social issue on the role of modern technological advances in unemployment and examinees were required to write their opinions in an essay of at least 250 words. The prompt was "Technological advances have replaced humans in the workplace. Technology is increasingly responsible for unemployment. To what extent do you agree or disagree with this statement?" The rating criteria were 'task achievement, 'coherence and cohesion', 'lexical resource', and 'grammatical range and accuracy'. The essays were rated by the two instructors who taught the course using a rating scale containing four rating criteria each rated on a 5-point scale. The raters were experienced EFL university instructors and underwent a 2-hour training session in using and interpreting the scale. Both raters scored all the essays, and thus there were no missing data.

## 4. Data Analysis and Results

Data were analyzed using the "KernSmoothIRT" package version 6.4 (Mazza, Punzo, & McGuire, 2020) in the R statistical software (R Development Core Team, 2019). The non-parametric estimation of item characteristic curve fits non-parametric item

and option characteristic curves using kernel smoothing techniques. The package allows a variety of exploratory plots for both dichotomous and polytomous data at test- and item-level as well as across different subgroups to examine the functioning of a scale, the individual items, and the test takers.

## 4.1 Rater Characteristic Curves

Figure 1 presents the option characteristic curves of the two raters on the five writing components or items. Instead of using the term option characteristic curves, here, the more accurate term rater characteristic curves is used for rater-mediated assessment. On the rater characteristic curve graphs, the *y*-axis represents the probability of assigning rating scores to students' writing, ranging from 0 to 1, and the *x*-axis represents the expected total score on the latent trait dimension on which students are ranked, ranging from 0 to 32. Expected score is defined as the average score a respondent at a particular latent trait level will attain (Ramsay, 2000). The vertical dashed lines, whose positions are equal for all the graphs, show the actual total scores below which 5%, 25%, 50%, 75%, and 95% of examinees fall. As an illustration, the 50% line is located at the score 13 for all the rater characteristic curves, suggesting that 50% of the examinees are below the actual total score of 13 and 50% are within the range of 13 to 32.

As presented in Figure 1, for each writing component or item of the rating scale, five curves in the rater characteristic curve graphs were plotted, representing the five rating categories (e.g., 0 to 4) in the scale. The curves indicate the relation between the latent dimension or students' writing ability and the probability of receiving a rating score. This relation is represented by item response functions, graphically shown by a rater characteristic curve. With regard to the monotonicity assumption, item response functions determine that students with higher writing ability are more likely to have higher probabilities of receiving higher ratings than writers with lower writing ability. In other words, as students' writing ability increases, the probability of receiving higher rating scores increases as well. If monotonicity assumption is satisfied, it is an indication that the probability of assigning higher rating scores for higher level writers is non-decreasing (Wind, 2015). A rater characteristic curve that fulfills the monotonicity assumption, in the case of rater-mediated assessment, indicates consistent ratings. That is to say, with increasing levels of writing ability, the probability of assigning higher ratings monotonically increments, or at least does not decrease. On the contrary, any violations of monotonicity, e.g., a "U-shaped" or a "wave" curve, indicates inconsistent ratings, that is, some writers with higher writing levels have received lower ratings, and writers with lower writing levels have received higher ratings. Research has shown that violations of monotonicity assumption impact the accuracy of measurement (Hambleton & Swaminathan, 1985; Sijtsma & Molenaar, 2002).

With regard to the component of task achievement displayed in Figure 1, except for Category 2, all of the curves show that the monotonicity assumption holds for Rater

1, although Category 4 has a minor distortion in monotonicity in the range of 11 to 14. The satisfied monotonicity assumption indicates that raters have assigned higher ratings as writing ability levels (or total scores) of students increase. Similar to Rater 1, most of the curves indicate the satisfaction of monotonicity for Rater 2, except for Category 2 and Category 1 in which there is a small bump in the right part of the curve. The probability of assigning rating scores to students' writing shows that Rater 1 is more lenient than Rater 2 with respect to task achievement component. As can be seen, the probability of assigning Category 3 is 0.58 for Rater 1 whereas it is 0.42 for Rater 2. For coherence and cohesion, the performance of Rater 1 shows the violation of monotonicity for Category 2 and Category 1 at the lower end of the scale. The performance of Rater 2 on coherence and cohesion component represents an aberrant rating behavior because monotonicity assumption has been seriously violated, indicating that Rater 2 has inconsistently rated the students' writing. That is, some high level examinees have been rated lower than expected whereas low level ones have been rated higher than expected. Similar to task achievement component, Rater 2 is more lenient than Rater 1 because the probability of assigning rating scores is lower for Rater 2 compared to Rater 1. As to the lexical resource component, the performance of both raters indicates the fulfilment of monotonicity since those writers with high writing levels are more likely to receive higher ratings. The comparison of the rater characteristic curves for both raters reveals that Rater 1 has harshly rated on Category 3 while Rater 2 has been more severe on the other rating categories. With respect to the grammatical range and accuracy, Rater 1 showed a consistent rating because students receive higher ratings as their writing ability increases. For Rater 2, rater characteristic curves related to Categories 2, 3, and 4 satisfied monotonicity assumption; however, in Categories 0 and 1, violation of monotonicity is observed at the lower end of the dimension, indicating that students with higher writing ability were given lower ratings and students with lower writing scores received higher scores. Any violations of monotonicity assumption, in performance-based assessment such as writing, on different components are more likely to be ascribed to the multidimensional nature of the latent trait (Baghaei, 2021; Effatpanah & Baghaei, 2021; Effatpanah, Baghaei, & Boori, 2019). In fact, a student may possess higher ability in a particular component, say grammar, but does not have adequate competency in another component, such as vocabulary.

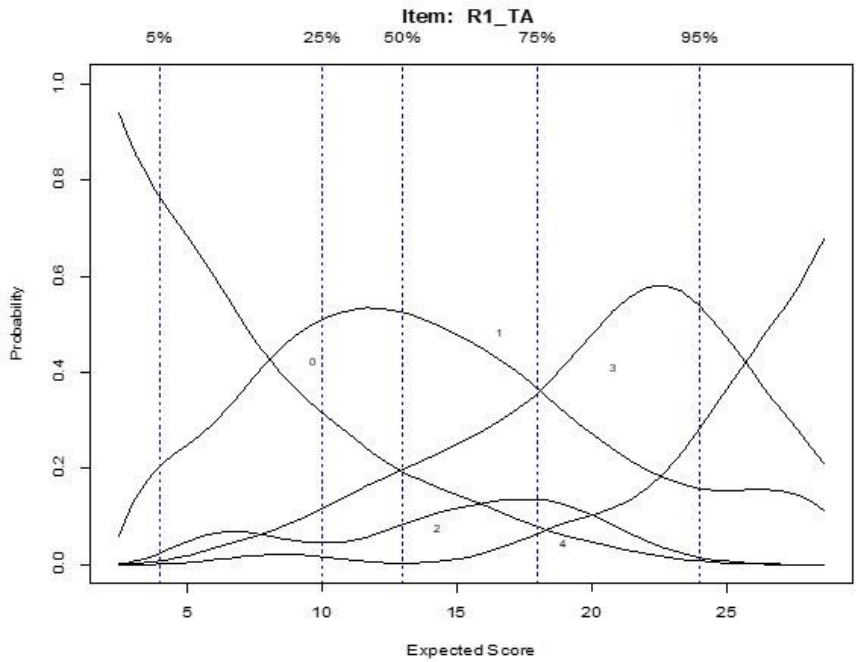
More importantly, each rating category should represent the most probable rating score for writers at particular positions of the writing ability continuum. The probability of receiving higher ratings should increase as the writing ability levels of students increase. For the first rating category (0), it is expected to be the most probable for writers with the lowest writing ability level, and as the writing ability of students increases, the probability of receiving category (0) should decrease. The probability of the lowest category should be near 1 at the lowest end of the writing ability continuum. Category 1 should be the most probable category for low-ability writers and become less probable as the writing ability of the examinees increases. Category 2 should be the most probable for writers with medium levels of writing ability and be less probable for writers below and above this ability range. Category 3 should be the

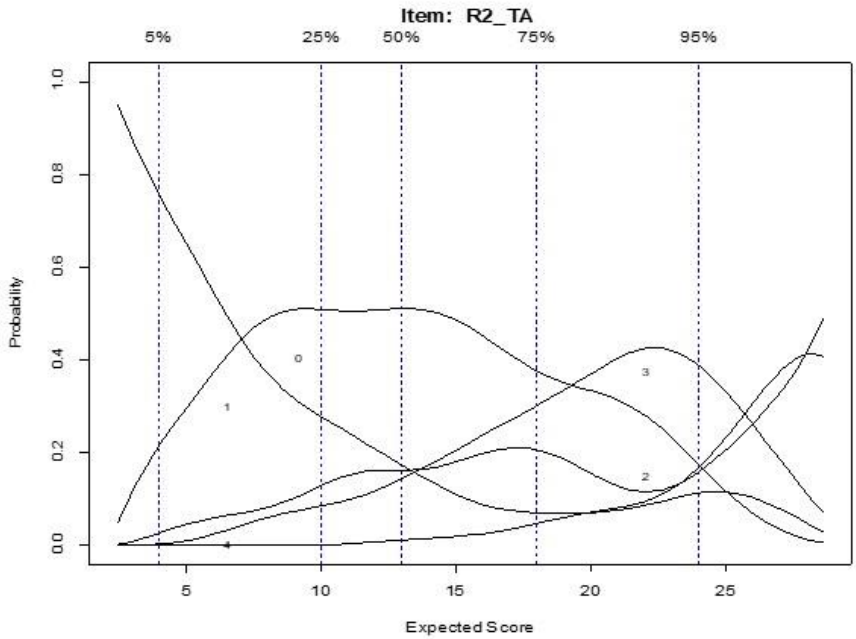
most probable for writers at medium to higher levels of writing ability. And finally, Category 4 should be the most probable for writers with the highest writing ability level. The probability of the highest category should be near 1 at the highest end of the writing ability continuum and should approach zero at the lowest levels of the continuum. Therefore, an ideal rater characteristic curve should resemble several successive dispersed and peaked curves across all levels of the writing ability continuum, each representing a category and a class of writers based on their writing ability levels.

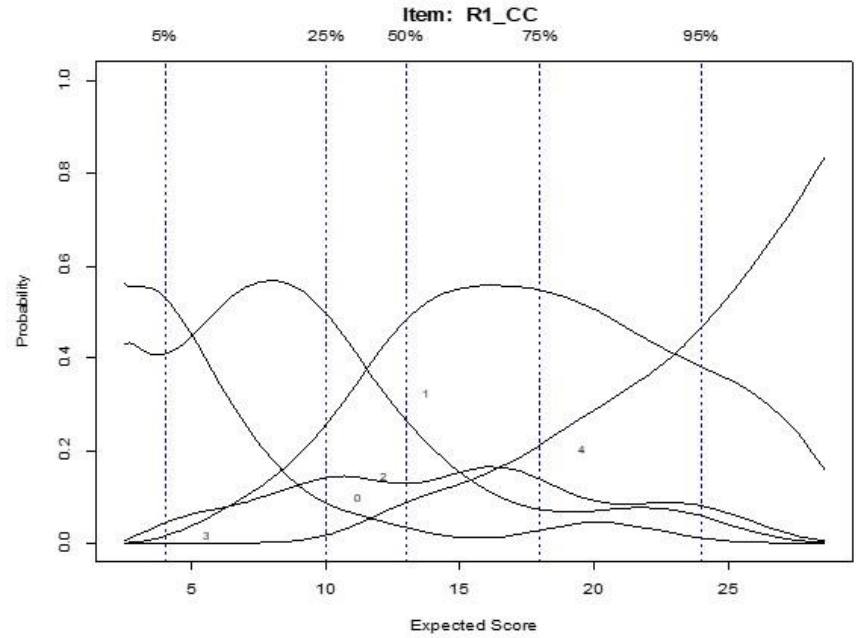
For instance, the performance of Rater 1 and Rater 2 on lexical resource component as illustrated in Figure 1 is examined. For Rater 1, category 0 is the most probable category for writers with expected scores between 1 and 8, and is less probable with the increase of writing ability across the continuum. The probability of category 0 is exactly 1 at the lowest end of the continuum. For writers with expected scores above 8 (e.g., between 9 and 17), Category 1 is more likely and its probability decreases as the level of writing ability increases. The probability of receiving category 2 is more probable for writers with expected scores between 12 and 19. This category is obviously flimsy because its probability is lower than the probability of Category 1 along the continuum. Students with writing ability levels between 9 and 23 have higher probabilities to receive Category 1 than Category 2. Except for the range of 17 to 19, Category 2 is never more likely than Category 1. For expected scores between 14 and 24, Category 3 becomes more probable, and with increased levels of writing ability, its probability decreases. Finally, examinees with expected scores above 25 are more likely to receive Category 4. The probability of category 4 is near 1 at the highest end of the continuum and is near 0 at the lowest end of the continuum. On the other hand, for Rater 2, category 0 is more probable for writers with expected scores in the range of 1 to 6, and becomes less probable as the writing ability increases along the continuum. The probability of category 0 is less than 1 at the lowest end of the continuum. Category 1 has higher probabilities for writers with expected scores between 7 and 15, and as the level of writing ability increases, its probability decreases. Writers with expected scores between 9 and 15 have higher probabilities to receive rating 2 (Category 2). Similar to Rater 1, Category 2 for Rater 2 does not properly operate because its probability is lower than the probability of Category 1 across the entire length of the continuum. It seems that Category 2 is not effective and raters could not distinguish between 4 rating categories. This problem is observed for the other rater characteristic curves on different writing components or items. Therefore, one way for solving this problem is to merge Categories 1 and 2. Category 3 is more probable for writers with expected scores between 10 and 25. After expected score of 25, category 4 becomes the most probable rating category, whose probability is near 0 at the lowest end of the continuum and is near 1 at the highest end of the continuum.

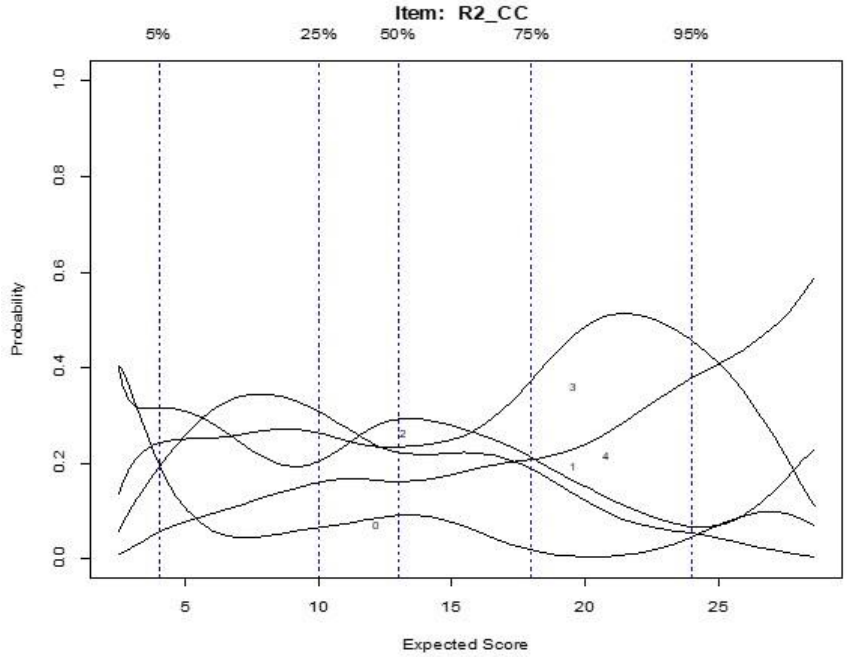
Furthermore, the slope or steepness of the rater characteristic curves give information on discrimination power of items and categories. Unlike parametric IRT models which provide a single item discrimination index, the non-parametric estimation of item characteristic curve allows practitioners to both track the changes in item discrimination across the expected latent trait continuum and visually compare all items with respect to their discrimination power at various levels of the latent trait because

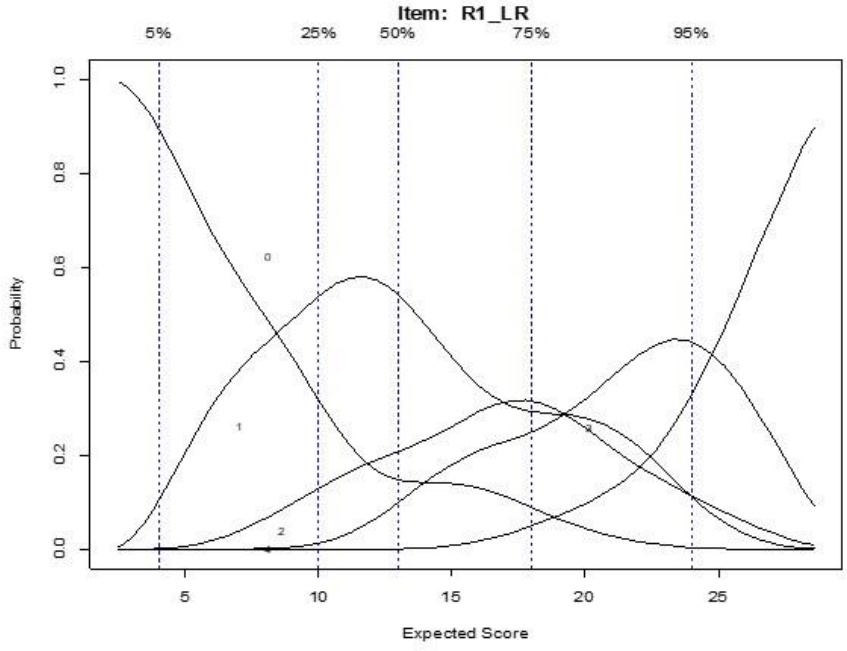
the non-parametric estimation of item characteristic curve does not impose a particular shape for curves (Rajlic, 2020). In the case of rater-mediated assessment, item discrimination indicates the rate at which the probability of assigning rating scores to students' writing performance changes given students' writing ability levels. The steeper the rater characteristic curves, the better the items or the rating categories can differentiate between writers with various writing ability levels. For illustrative purposes, as demonstrated in Figure 1, rater characteristic curves for Rater 1 on grammatical range and accuracy show that the rater could better discriminate between writers with various levels of writing ability in Categories 0, 1, and 4 compared to Rater 1, especially with expected scores ranging from 5 to 9. In contrast, Rater 2 was able to differentiate between writers in Categories 2 and 3, especially with expected scores in the range of 10 to 14.

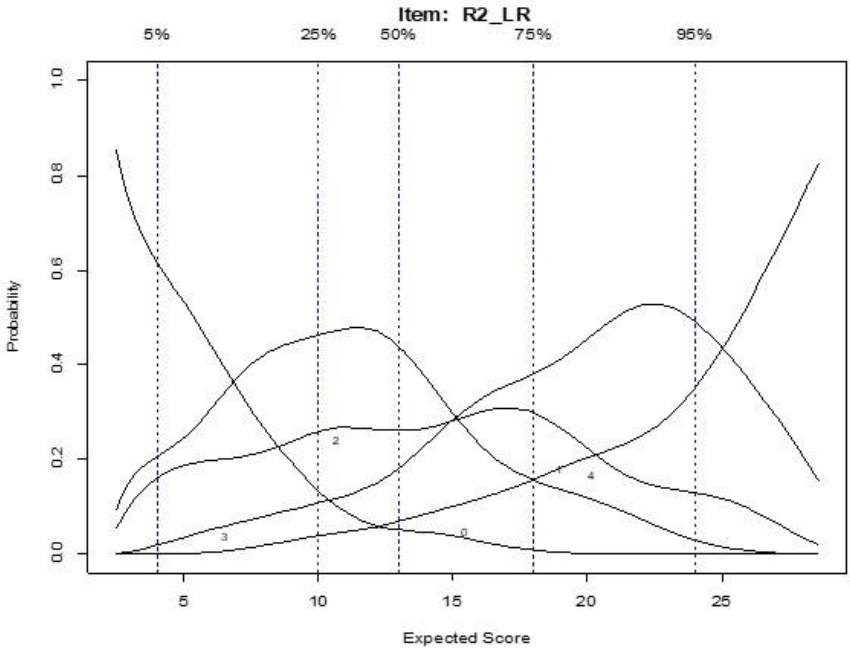


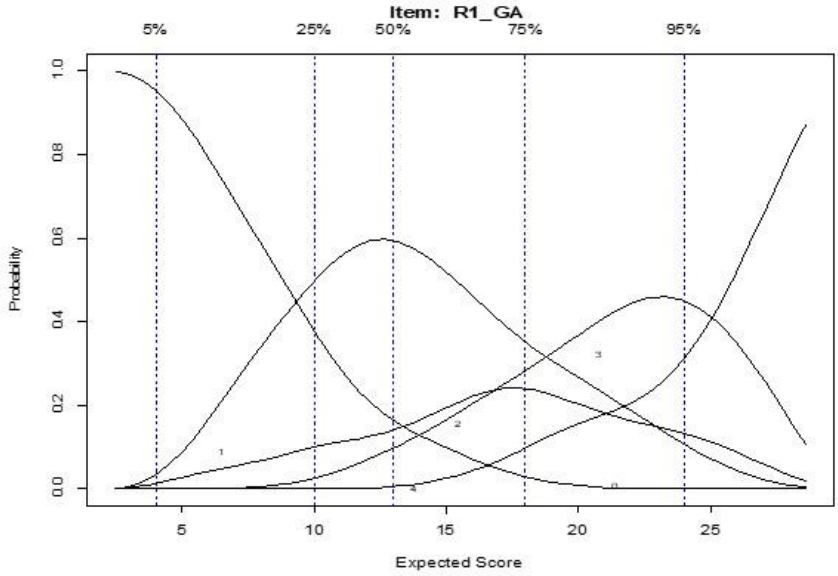

Item: R2_TA
5%
25%
50%
75%
95%
Probability
Expected Score
0
1
2
3
4


Item: R1_CC
5%
25%
50%
75%
95%
Probability
Expected Score
0
1
2
3
4


Item: R2_CC
5%
25%
50%
75%
95%
Probability
0.0
0.2
0.4
0.6
0.8
1.0
0
1
2
3
4
5
10
15
20
25
Expected Score


Item: R1_LR
5%
25%
50%
75%
95%
Probability
0.0
0.2
0.4
0.6
0.8
1.0
0
1
2
3
4
5
10
15
20
25
Expected Score


Item: R2_LR
5%
25%
50%
75%
95%
Probability
0.0
0.2
0.4
0.6
0.8
1.0
0
1
2
3
4
5
10
15
20
25
Expected Score


Item: R1_GA
5%
25%
50%
75%
95%
Probability
0.0
0.2
0.4
0.6
0.8
1.0
0
1
2
3
4
5
10
15
20
25
Expected Score

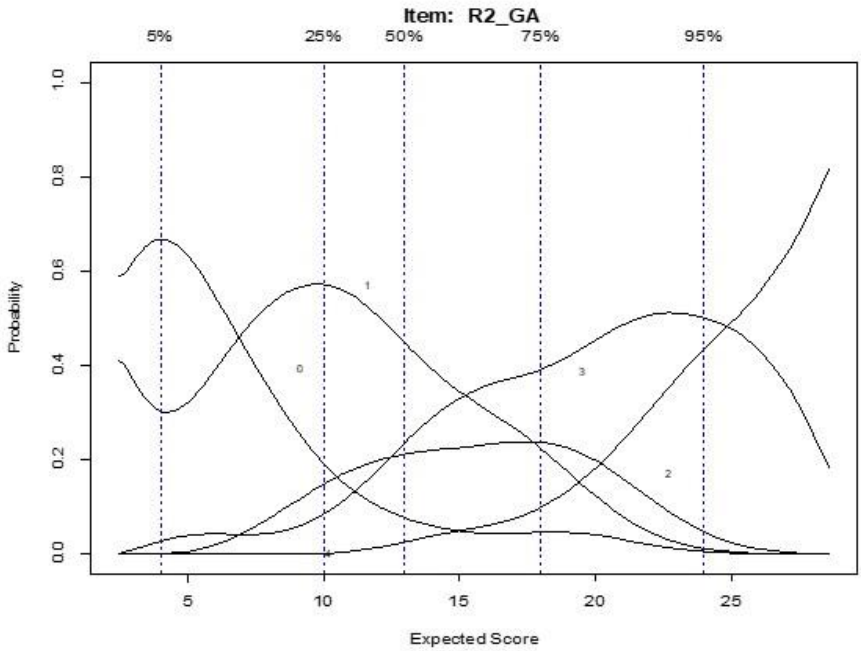



***Figure 1.***
Rater Characteristic Curves (RCCs) for the Two Raters on the Five Writing Components. TA = Task Achievement, CC = Cohesion and coherence, LR = Lexical Resource, and GA = Grammatical Range and Accuracy.

## 4.2 Tetrahedron Plots

Figure 2 gives (regular) tetrahedron simplex plots for the two raters on lexical resource and grammatical range and accuracy components. Tetrahedron plots are only used for items with more than 3 or 4 categories or options. At each side of the tetrahedron, only categories with the highest probabilities are presented, and the highest probabilities are normalized to offer a simple representation (Mazza et al., 2014). A curve with three colors is observed inside the tetrahedron. Each color shows a specific trait level; blue points indicate high trait levels, green points show medium trait levels, and red points represent low trait levels. These trait levels are simply the values of evalpoints divided into three equal groups. As Mazza et al. (2014) note, an essential requirement of a reasonable category or item is that the sequence of points terminates at or near the highest category or the correct response. For space considerations, the tetrahedron plots for only two writing components as illustrated in Figure 2 are presented. The

five categories with the highest probabilities are shown at each side of the tetrahedron and as Category 2 has the least probability, it is not illustrated on the tetrahedron. This can be considered as evidence that the raters could not effectively make a distinction between the rating categories, that is, they did not effectively use Category 2 in their ratings. With regard to the lexical resource and grammatical range and accuracy components, the performance of the two raters shows that the sequence of points starts from category 0 (vertex), passes options 1, 3, and moves toward option 4, suggesting that with the increased levels of writing ability, the probability of receiving higher ratings increases as well. Another issue in the analysis of tetrahedron is the length of the curve. There should be a distance between the examinees with the highest and the lowest trait levels. In Figure 2, the ratings of the two raters on lexical resource and grammatical range and accuracy rating components are satisfactory because the examinees with the highest writing levels are far from those with the lowest writing levels and the length of the curve is not short, but it ends at or near the highest category. Furthermore, the spacing of the points indicate the speed at which the probabilities of rating categories change. As Figure 2 shows, the ratings of the raters on the two components indicate a good performance because as writing ability increases, the probability of assigning higher rating categories increasingly changes.

## 4.3 Differential Item Functioning (DIF) Analysis

In addition to rater characteristic curves, the non-parametric estimation of item characteristic curve can also provide information about differential item functioning (DIF), which is used in educational and psychological measurement to detect bias at item-level. In the non-parametric estimation of item characteristic curve, the presence of DIF is identified by inspecting visual representations of item responses or rating scores across the relevant groups. Any substantial disparities in the shape of the curves across the subgroups and the area between the curves indicate DIF (Rajlic, 2020). In rater-mediated assessments, DIF takes place when rating categories do not function in a similar way for all writers across different subgroups. DIF exists if writers with the identical writing ability levels from different subgroups do not have equal probabilities to receive a particular rating. A biased item or rating category in which the probability of assigning the category is constantly higher for a particular group across the writing ability continuum is called uniform DIF, and when the probability of assigning rating categories vary for the relevant groups across the continuum, it is called non-uniform DIF.

Figure 3a presents the pairwise expected scores or QQ-plot for the distributions of the expected scores for females (on the *x*-axis) and males (on the *y*-axis). The horizontal and vertical dashed lines, similar to rater characteristic curves, represent the actual total scores below which 5%, 25%, 50%, 75%, and 95% of the examinees across the two groups fall. If the two groups have an identical performance, the solid diagonal line should not significantly deviate from the dotted diagonal line, as a reference (Ramsay, 2000); otherwise, the solid line

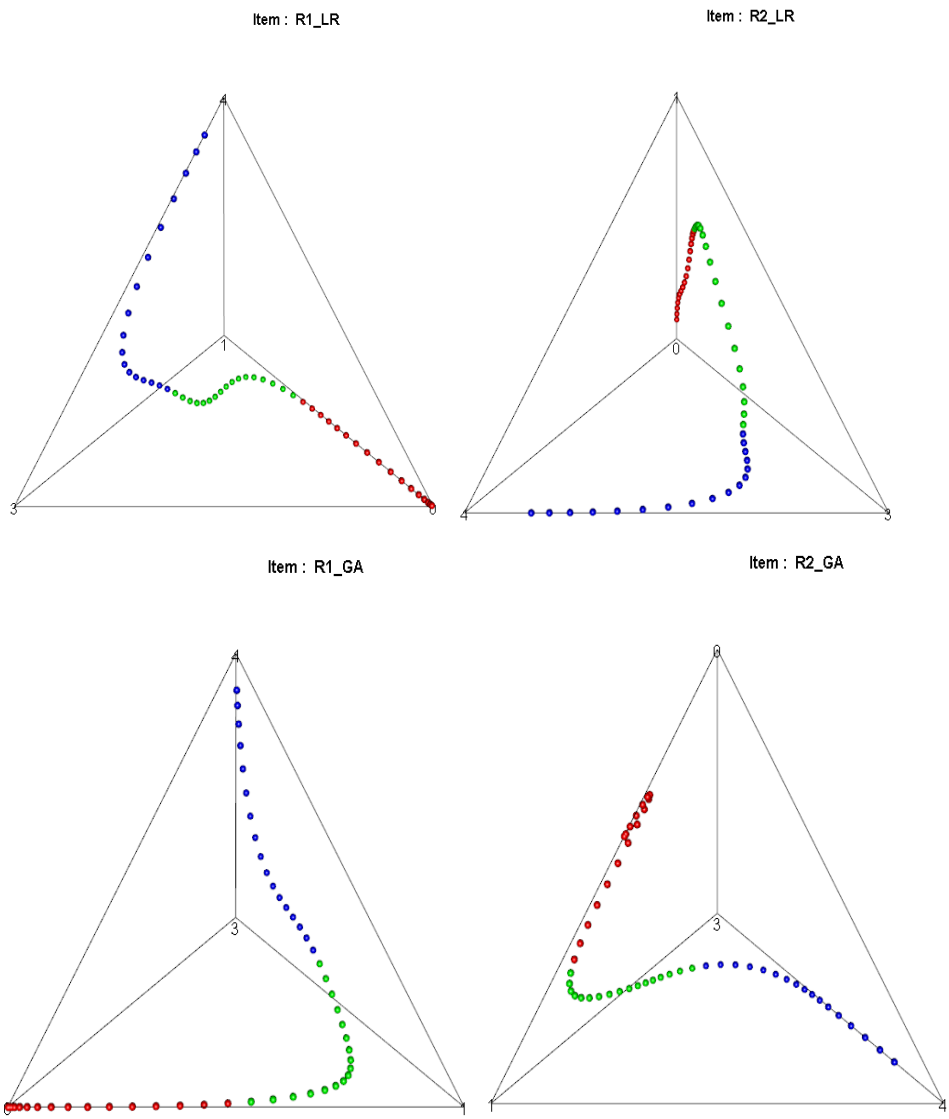



***Figure 2.***

Probability Tetrahedrons for the Two Raters on Lexical Resource and Grammatical Range and Accuracy Components. Low trait levels are plotted in red, medium in green, and high in blue. LR = Lexical Resource and GA = Grammatical Range and Accuracy.

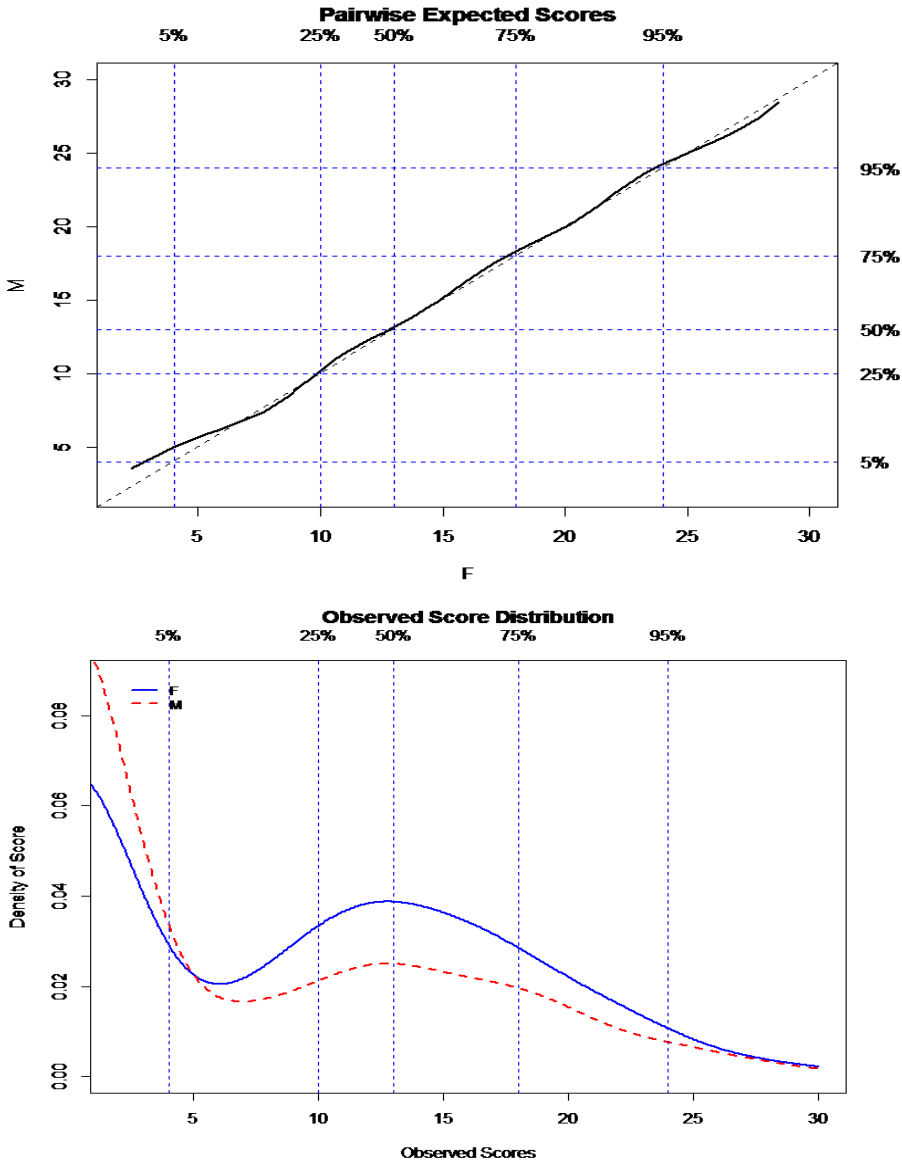
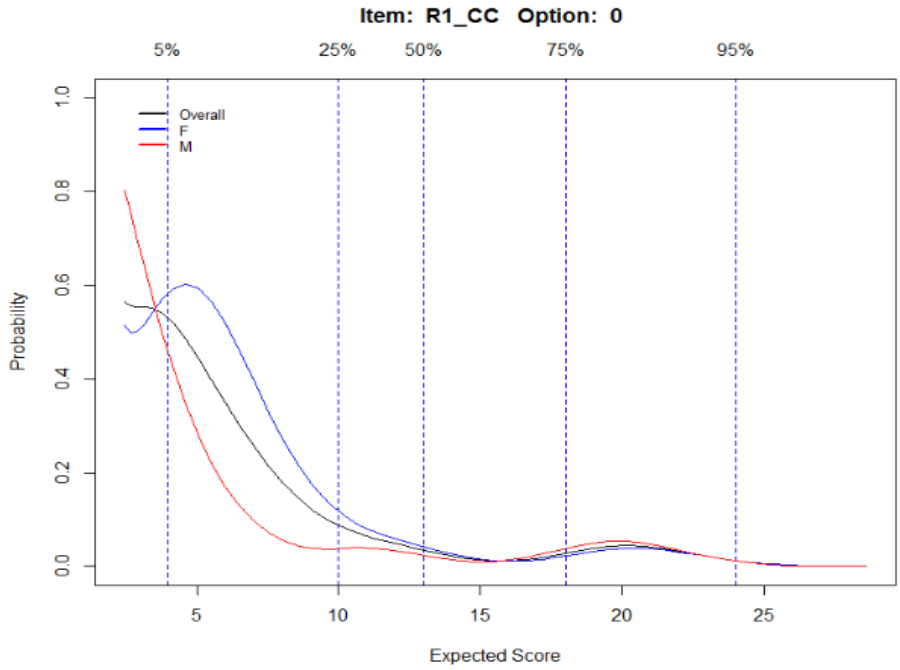


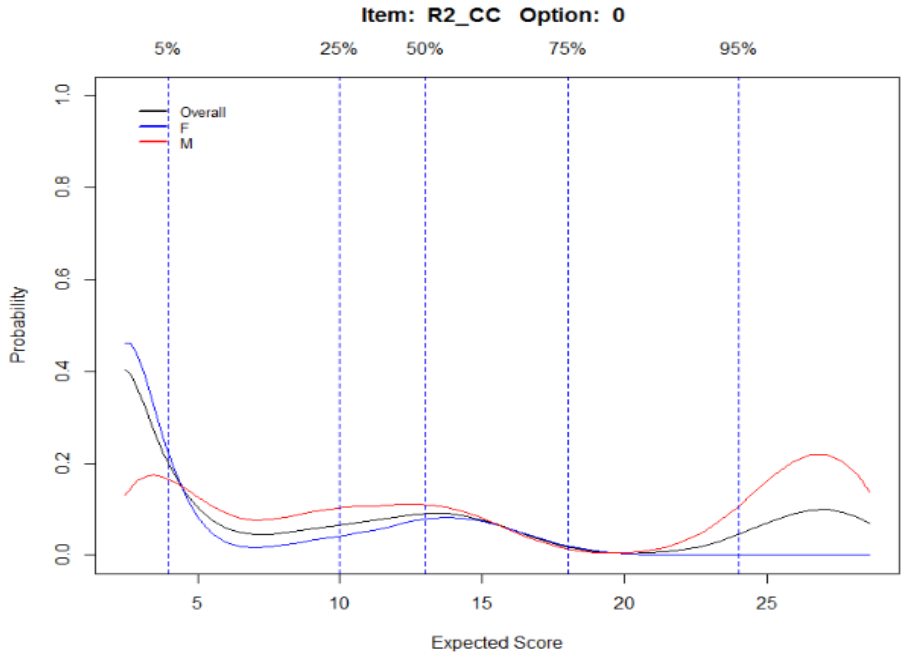

***Figure 3.***

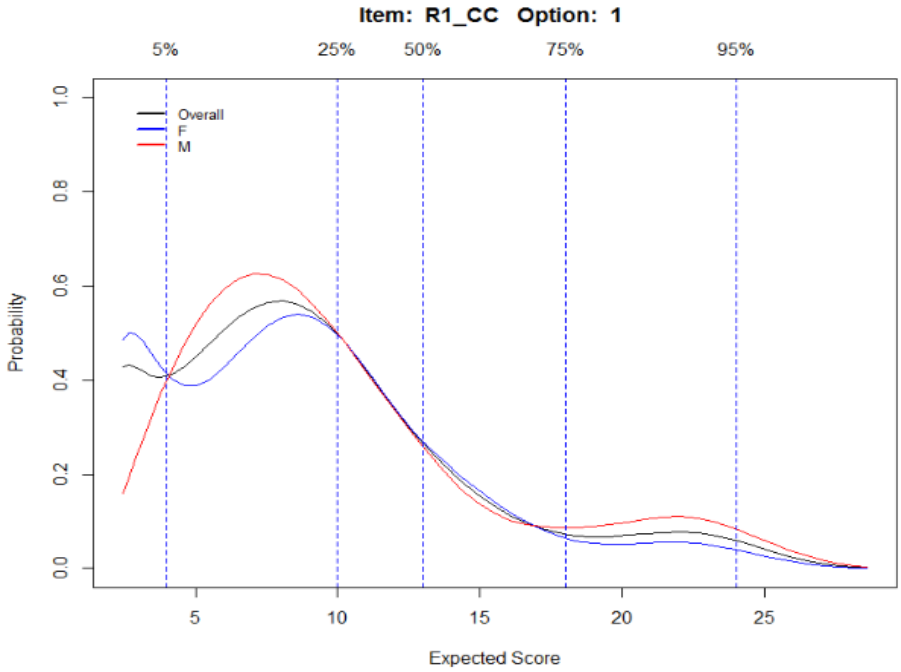
The Pairwise Expected Scores and (QQ-plot) and Kernel Density Functions for Females (blue solid line) and Males (red dotted line) on the Rating Scale

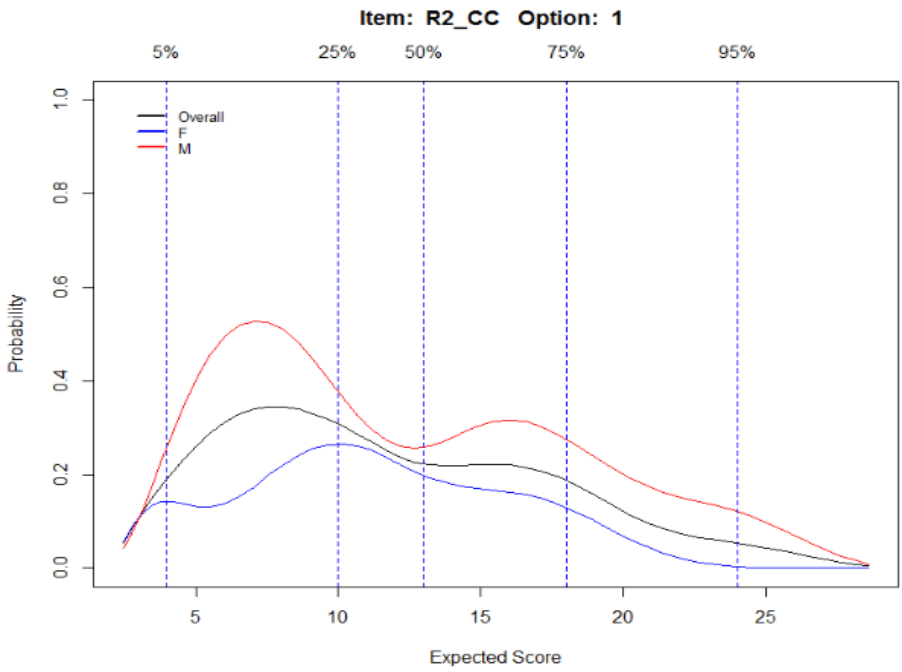
will deviate from the reference line, suggesting greater differences in the two distributions. As can be seen in Figure 3a, although males (M) have higher scores in the range of 2 to 7 compared to females (F), there is a slight deviation from the diagonal line indicating a slight difference in rating performance of the two raters. By marking the plot, it turns out that males with a total score of 5 (on the *y*-axis) received higher ratings about 1 or 2 points over females with total scores of 3 or 4 at the same quantile position (on the *x*-axis), representing a negligible difference between the two groups. Figure 3b further shows the kernel density function plot (e.g., the total score distribution) for the two groups. The solid line shows the observed scores for females and the dotted line for males. The plot shows that there is a disagreement between the distributions of the total scores across the two groups. Females were given higher ratings in the range of 6 to 26 whereas males received higher scores in the range of 1 to 5.

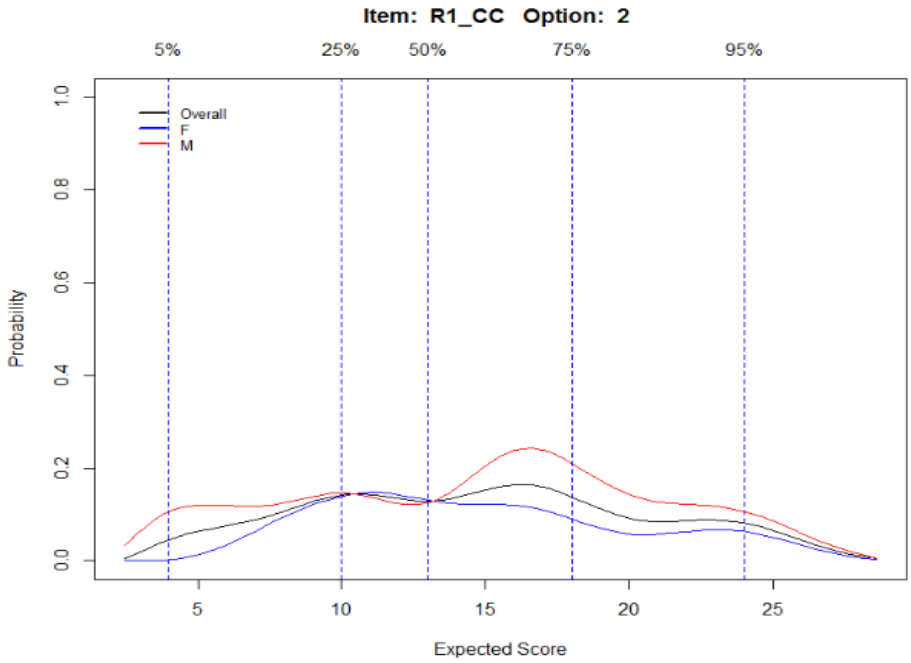
To compare DIF at the item or category-level, the rater characteristic curves for different rating categories of coherence and cohesion component across the two groups as demonstrated in Figure 4 are compared. On the rater characteristic curves, the red curves (M) represent the score distributions for male students, the blue curves for female students (F), and the black curves (the overall curve) for all students. With regard to the rating performance of Rater 1 on Category (0), females have a greater probability than males at the lower end of the scale (e.g., in the range of 4 to 10 expected scores) to receive higher ratings whereas the probability is higher for males than females at the higher end of the dimension for Rater 2. In relation to Category (1), the curves are very similar and close to each other for Rater 1, suggesting the lack of DIF; however, for Rater 2, the probability of receiving higher rating scores is higher for males along the scale, indicating the presence of uniform DIF. As to the Category (2), although the performance of Rater 1 shows a slight difference between the two groups in the middle of the scale (e.g., 15 to 20 expected scores), with higher probabilities for males, the curves are very similar and close to each other. For Rater 2, the probability of assigning higher scores is greater for females in the middle and at the higher end of the scale, representing differences in rating of this rater across the two groups. The performance of Rater 1 on Category 3 shows that males have greater probabilities of obtaining 3 at the lower and higher end of the scale, and females have greater probabilities in the middle of the scale, but these differences are not substantial. For Rater 2, males have lower probabilities than females on Category 3 of cohesion and coherence component, along most of the latent dimension, especially at the higher end of the scale, indicating uniform DIF. That is to say, it shows that Rater 2 tends to be more severe for males than Rater 1 in using this category. With respect to Category (4), the rating function of Rater 1 demonstrates that from the middle part of the scale, the probabilities for females are higher than males, that is, the rater assigned higher ratings to females compared to males. However, the performance of Rater 2 shows that the probability of receiving rating Category 4 is greater for females at the lower end of the dimension, whereas the probability is greater for men at the higher end of the continuum, suggesting the presence of non-uniform DIF.

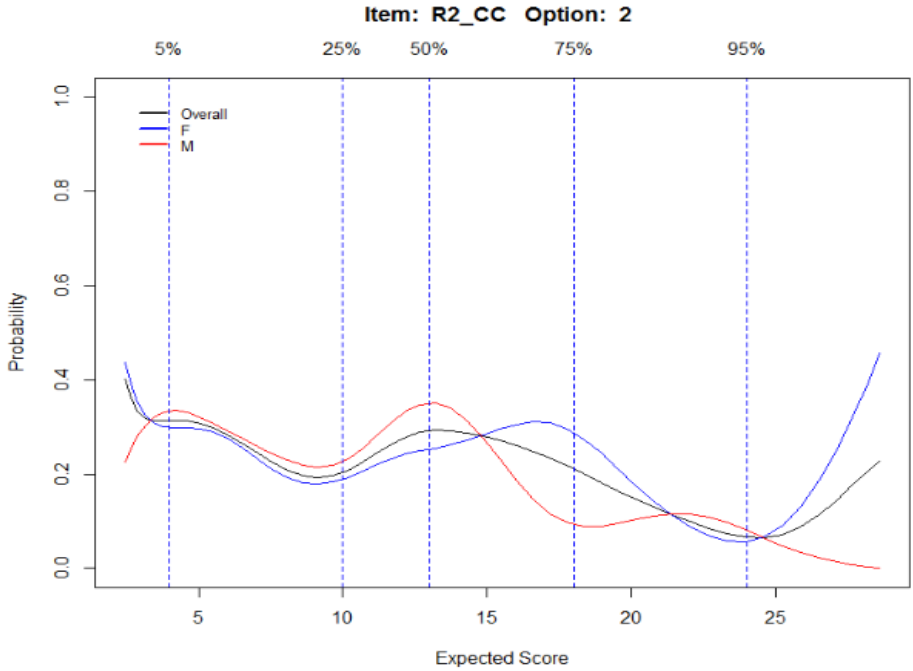

Item: R1_CC Option: 0
5%
25%
50%
75%
95%
Overall
F
M
Probability
0.0
0.2
0.4
0.6
0.8
1.0
5
10
15
20
25
Expected Score

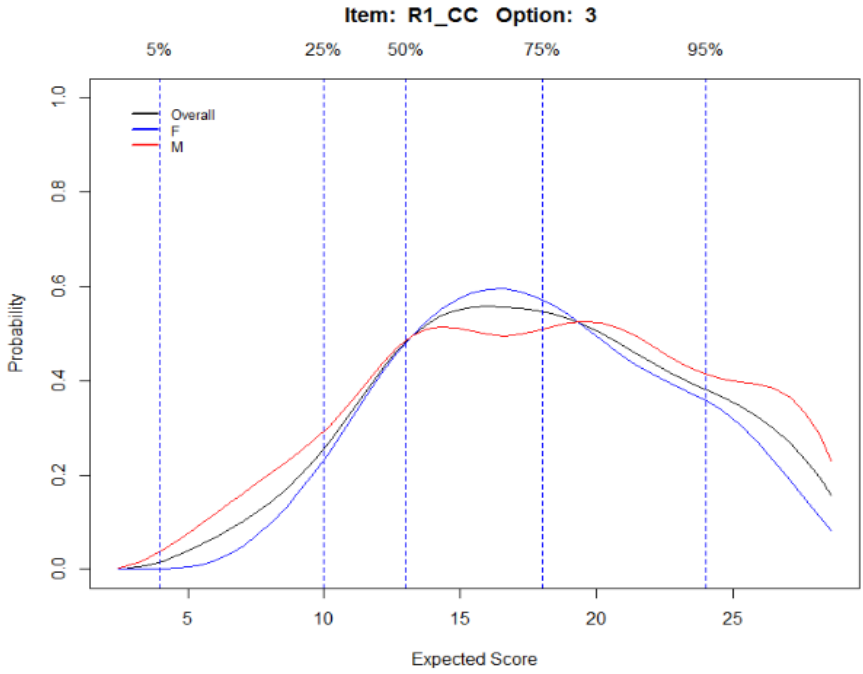

Item: R2_CC Option: 0
5%
25%
50%
75%
95%
Overall
F
M
Probability
0.0
0.2
0.4
0.6
0.8
1.0
5
10
15
20
25
Expected Score

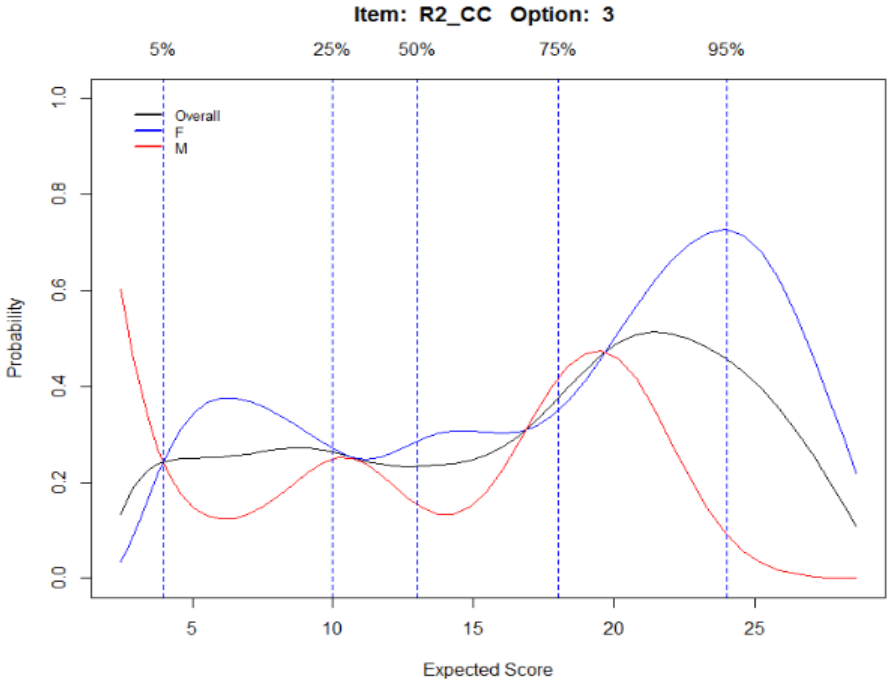

Item: R1_CC Option: 1
5%
25%
50%
75%
95%
Overall
F
M
Probability
0.0
0.2
0.4
0.6
0.8
1.0
5
10
15
20
25
Expected Score


Item: R2_CC Option: 1
5%
25%
50%
75%
95%
Overall
F
M
Probability
0.0
0.2
0.4
0.6
0.8
1.0
5
10
15
20
25
Expected Score


Item: R1_CC Option: 2
5%
25%
50%
75%
95%
Overall
F
M
Probability
0.0
0.2
0.4
0.6
0.8
1.0
5
10
15
20
25
Expected Score


Item: R2_CC Option: 2
5%
25%
50%
75%
95%
Overall
F
M
Probability
0.0
0.2
0.4
0.6
0.8
1.0
5
10
15
20
25
Expected Score


Item: R1_CC Option: 3
5%
25%
50%
75%
95%
Overall
F
M
Probability
0.0
0.2
0.4
0.6
0.8
1.0
5
10
15
20
25
Expected Score


Item: R2_CC Option: 3
5%
25%
50%
75%
95%
Overall
F
M
Probability
0.0
0.2
0.4
0.6
0.8
1.0
5
10
15
20
25
Expected Score

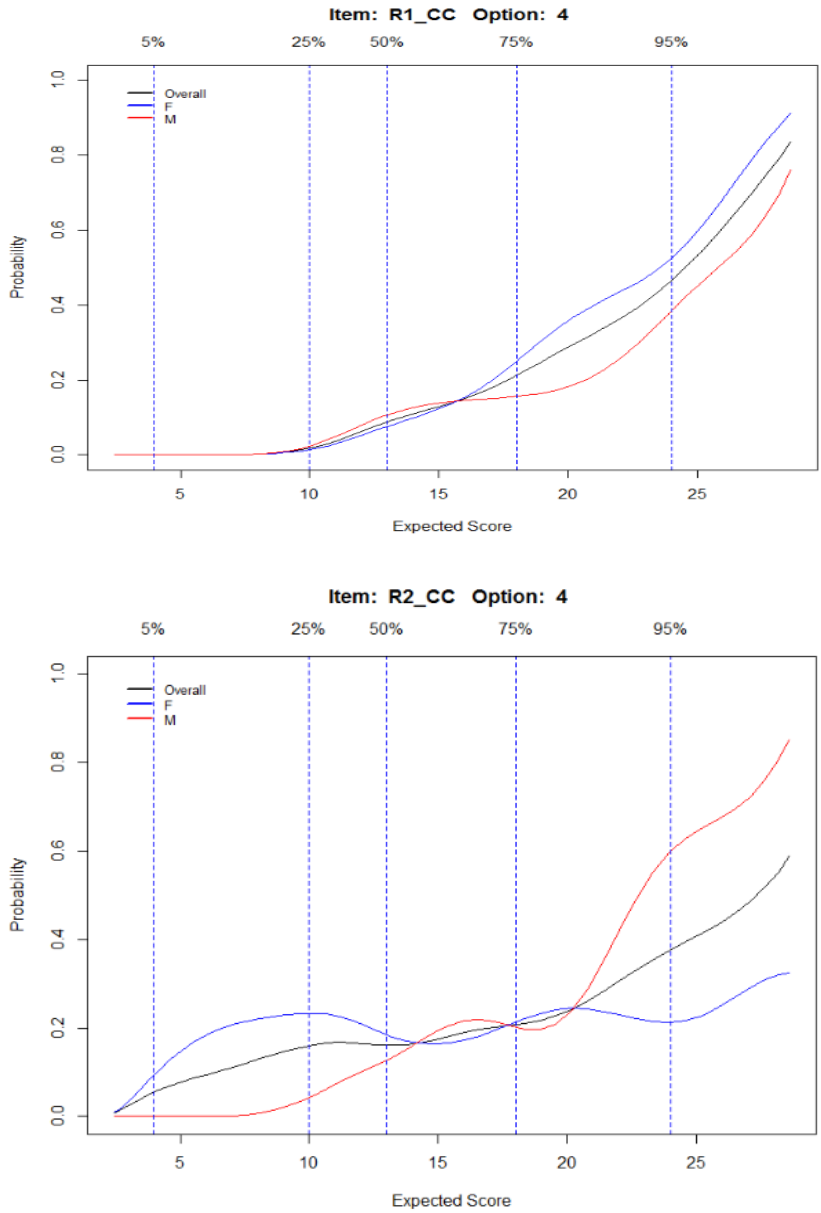
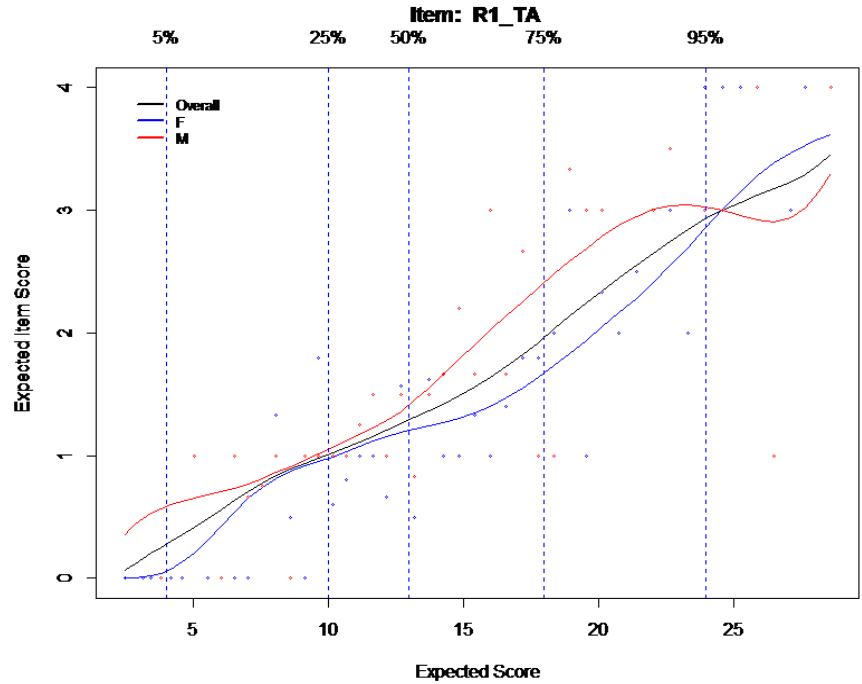
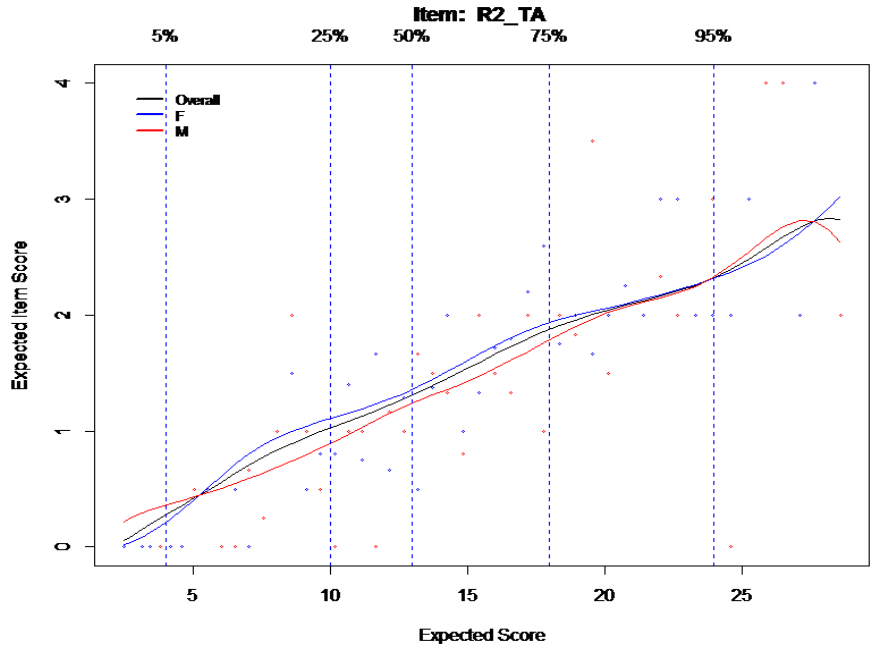



**Figure 4.**
Rater Characteristic Curves (OCCs) for Females and Males related to Coherence and Cohesion (CC) Rating Component

Finally, Figure 5 depicts the expected scores plots for the two raters on rating components of task achievement and coherence and cohesion across the two groups. On the graphs, the expected scores for males are represented by the red curves and for females by the blue curves, and for all the examinees by the black curves (the overall curves). The vertical dashed lines present the points below which 5%, 25%, 50%, 75%, and 95% of the examinees fall based on their total scores, and scattered color points on the plots represent the observed average scores for all the students grouped on the basis of their ordinal ability estimates (which are equally spaced) (Mazza et al., 2014). As can be seen, with regard to the rating performance of Rater 1 on the component of task achievement, male writers have relatively higher expected scores than females across different parts of the writing dimension. The curves at the higher end of the scale show that Rater 1 becomes more lenient for males in rating high ability writers. However, for Rater 2, the curves are very close and similar, indicating the lack of bias. As to the cohesion and coherence component, the performance of Rater 1 shows that females have lower expected scores at the lower end of the dimension, but they have higher expected scores at the upper end of the dimension, suggesting a slight non-uniform DIF. In other words, this indicates that the rater has inconsistently rated across the two groups. For Rater 2, females have greater expected scores than males at the lower end of the dimension, but from the middle part of the continuum, the rating performance becomes more consistent. Overall, the DIF plots at both test- and item-level reveal that gender is a variable causing DIF in the data.


Item: R1_TA
5%
25%
50%
75%
95%
Overall
F
M
Expected Item Score
Expected Score
0
1
2
3
4
5
10
15
20
25


Item: R2_TA
5%
25%
50%
75%
95%
Overall
F
M
Expected Item Score
Expected Score
0
1
2
3
4
5
10
15
20
25

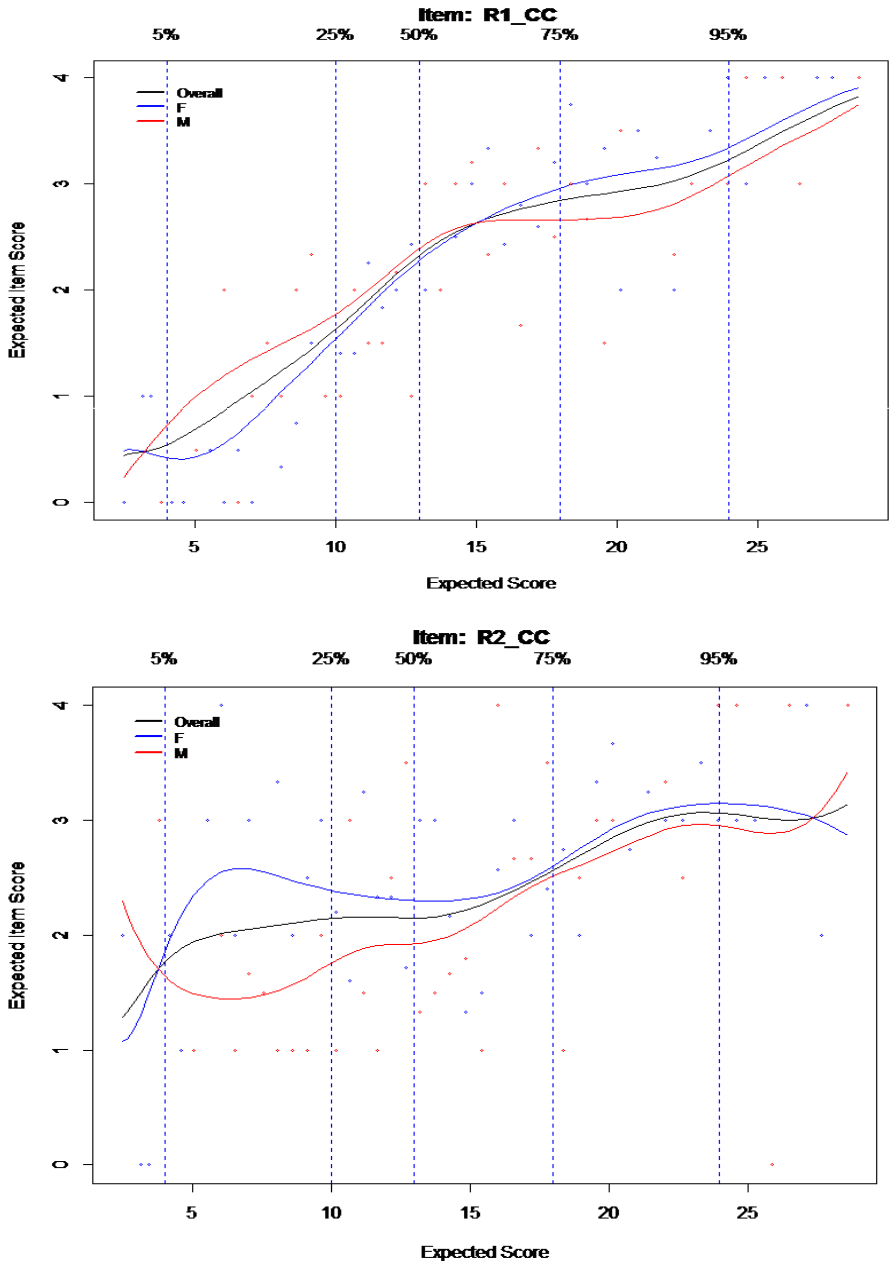



***Figure 5.***
Overall Expected Item Score (EIS) and EIS of Females and Males for the Two Raters on Task Achievement and Coherence and Cohesion Components. TA = Task Achievement and CC = Cohesion and Coherence.

## 5. Discussion

This study set out to use the non-parametric estimation of item characteristic curve using kernel smoothing technique to show the use of visual representations as a diagnostic tool for exploring and modeling the scoring patterns of raters' judgment and detecting measurement disturbances or rater effects in educational rater-mediated performance assessments. In previous studies, numerous researchers have explored the usefulness of graphical methods based on Mokken Scale Analysis (Mokken, 1971) for evaluating rating quality (Wind, 2020; Wind & Patil, 2018; Wind & Schumacker, 2017, 2018) in practical performance assessment settings and illustrated the value of this approach as a complementary technique to other parametric methods, like Many-Facet Rasch Model, for rating quality analysis. This study expands this line of research by using the non-parametric estimation of item characteristic curve. To our best knowledge, this study is the first attempt in the literature that employs the non-parametric estimation of item characteristic curve to model and monitor rating quality to discern idiosyncratic scoring patterns of raters. This research, therefore, makes a contribution to previous studies in applying non-parametric models to rater-mediated educational performance assessments in a number of ways. Firstly, the use of the non-parametric estimation of item characteristic curve has already been limited to some practical (Effatpanah & Baghaei, 2022; Meijer & Baneke, 2004; Santor, Ramsay, & Zuroff, 1994; Sijtsma et al., 2008) and methodological (Douglas, 1997; Douglas & Cohen, 2001; Wells & Bolt, 2008) research in psychological assessments, psychopathology, and quality of life. This study extends the application of the non-parametric estimation of item characteristic curve to evaluate rating patterns, by giving only visual illustrations, and detect rater effects in the context of rater mediated assessments. Secondly, while researchers have largely used Mokken Scale Analysis as a non-parametric approach and explored the use of this model to evaluate rating quality and identify raters with unexpected rating patterns, Mokken Scale Analysis has several practical shortcomings for analyzing rating quality. The first limitation is that in Mokken Scale Analysis, all items of a scale should possess the equal number of response options or categories; however, the non-parametric estimation of item characteristic curve allows items to have different number of categories. It is a common practice in performance assessment to give different weights to different rating criteria. For instance, in writing assessment, 'content' is always given a higher weight—with a rating scale ranging from say, 1 to 5—than 'mechanics' with a scale ranging from say, 1 to 3. This practice makes the scales for items different and unanalyzable with Mokken Scale Analysis. The second limitation of Mokken Scale Analysis is that the analysis of measurement invariance or DIF is demanding. The commonly used computer programs for running Mokken Scale Analysis, including MSP (Molenaar & Sijtsma 2000) and the R package *mokken* (van der Ark, 2012), do not include a procedure for estimating item response functions across different subgroups. To evaluate measurement invariance, item response functions of the subgroups should be separately estimated and then plotted on a single graph. The analysis of DIF in the non-parametric estimation of item characteristic curve, though, is easy to do, and the model

can simultaneously estimate item response functions of subgroups and plot them on a single graph.

Similar to Mokken Scale Analysis, the non-parametric estimation of item characteristic curve has the potential to provide an approach to investigate the rating quality of raters and consider psychometric characteristics as fundamental prerequisites for obtaining sensible interpretations from assessment results. Consequently, any violations of the non-parametric estimation of item characteristic curve assumptions can be used to specify whether the scoring of raters have psychometrically sensible interpretations or require further inspection. To investigate the utility of the non-parametric estimation of item characteristic curve to identify unexpected scoring patterns, different graphs were separately examined based on the scoring of two raters using an analytical scoring rubric measuring four writing criteria on a five–point scale. As the non-parametric estimation of item characteristic curve can provide a variety of exploratory plots at test- and item-level as well as across different subgroups, we focused on rater characteristic curves, (regular) tetrahedron simplex plots, and DIF. According to Wind (2020), since monotonicity is an-easy-to-understand psychometric property to evaluate and interpret with researchers, it can give straightforward information on rating quality. In this study, the interpretation of rater characteristic curves showed that although the raters, in most cases, adhered to the monotonicity assumption, their performance, in some cases, violated the assumption of monotonicity. This indicated frequent inconsistent ratings in which with increased levels of writing ability, the probability of receiving higher ratings did not increase. The violation of monotonicity in educational performance assessment can be due to the complexity and multidimensionality of writing (Baghaei, 2021; Effatpanah & Baghaei, 2021; Effatpanah, Baghaei, & Boori, 2019).

To complement the rater characteristic curves, tetrahedron plots for the raters on two rating components were analyzed. The patterns of raters' scorings at category level indicated that although the raters could not efficiently differentiate between the five rating categories, especially Categories 1 and 2, there was generally a satisfactory functioning of the raters in selecting the appropriate rating category because the assumption of monotonicity held. One possible reason for the inability of raters in drawing a distinction a distinction between Category 1 and 2 may be due to the lack of sufficient discussion and explanation on this category in the training session. In other words, the raters could not capture relevant aspects of students' writing performance regarding the content of Category 2; therefore, as Kuiken and Vedder (2014) noted, they placed more weight on the other categories. Alternatively, the description of Category 2 might overlap with the description of the adjacent categories making the distinction rather difficult.

Furthermore, the performance of raters across the subgroups (males and females) were examined. The QQ-plot and kernel density function plot for analyzing DIF at test-level revealed that the rating quality was invariant across the sub-groups, that is, there were no significant differences in rating function of the raters across females and males examinees. However, further analysis at item or category level on different

rating components showed substantial differences between the rating performance of the raters across the subgroups. The results indicated the existence of both non-uniform and uniform DIF. This finding suggests different levels of severity or leniency in rating performance of raters across the subgroups.

## 6. Conclusion

An attempt was made to demonstrate the usefulness of the non-parametric estimation of item characteristic curve for graphically investigating the rating quality of raters in the context of performance assessments. Overall, findings of the current study highlight the importance and effectiveness of visual methods for examining rating quality. Specifically, the exploratory graphs of the non-parametric estimation of item characteristic curve can present detailed information for further rater training and monitoring procedures. The findings of this study have numerous implications for studies related to evaluating rating quality in rater-mediated assessment. First, the application of the non-parametric estimation of item characteristic curve is useful for analyzing rating quality and identifying peculiar scoring patterns with regard to a set of important psychometric characteristics, such as monotonicity and measurement invariance, without transforming ordinal ratings to interval measures. In fact, prior to employing a parametric model, the use of the non-parametric estimation of item characteristic curve, similar to Mokken Scale Analysis, can be considered as an initial step to examine measurement properties of a rater-mediated assessment. As argued by Meijer et al. (2015, p. 107), "non-parametric approaches are excellent tools to decide whether parametric models are justified. Moreover, given the often no-so-easy-to-interpret fit statistics for parametric models, non-parametric tools provide a nice extension of the parametric toolkit to IRT modeling". Secondly, the results obtained from the non-parametric estimation of item characteristic curve can inform stakeholders of score interpretations and uses for making inferences and decisions about students in educational contexts.

Although the non-parametric estimation of item characteristic curve proved useful in identifying peculiar rating patterns in educational performance assessment involving a set of raters, the findings of the current study have to be considered with respect to some limitations related to non-parametric IRT models. The main limitation of the non-parametric estimation of item characteristic curve is that the mere graphical displays of the method, without giving any numerical values, make a challenge for analysts to investigate and judge the psychometric qualities of a measure. In fact, because there are not specific boundaries or criteria for analyzing graphs, the interpretation and detection of rater effects and/or DIF using the graphic method tend to be subjective or arbitrary to some extent. Furthermore, different plots obtained from the non-parametric estimation of item characteristic curve fail to clearly distinguish between various rater effects, including leniency/severity, centrality/extremity, or accuracy/inaccuracy. It is thus highly recommended for researchers and practitioners to use the non-parametric estimation of item characteristic curve for modeling and evaluating

rating quality along with parametric models. Wind (2019a) also acknowledges the limitations of non-parametric IRT models, including Mokken Scale Analysis, relative to parametric IRT models, which can be extended to the non-parametric estimation of item characteristic curves in the following way:

> "the lack of a parametric form prevents [non-parametric IRT] models from providing interval-level parameter estimates, such as are needed for computer-adaptive assessment procedures, equating, and other parametric analyses. Whereas parametric IRT models result in interval-level estimates that are suitable for such analyses, [non-parametric IRT] models do not. Additionally, [non-parametric IRT] models currently do not include a multi-faceted model similar to the MFR model through which analysts could examine more than two facets in a single analysis" (pp. 18-19).

## References

Baghaei, P. (2021). *Mokken scale analysis in language assessment*. Munster, Germany: Waxmann Verlag.

Baldwin, D., Fowles, M., & Livingston, S. (2005). *Guidelines for constructed-response and other performance assessments*. Princeton, NJ: Educational Testing Service.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick (Eds.), *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.

De Carlo, L. T., Kim, Y., & Johnson, M. S. (2011). A hierarchical rater model for constructed responses, with a signal detection rater model: Hierarchical signal detection rater model. *Journal of Educational Measurement*, *48*(3), 333-356. https://doi.org/10.1111/j.1745–3984.2011.00143.x

Douglas, J. (1997). Joint consistency of nonparametric item characteristic curve and ability estimation. *Psychometrika*, *62*, 7-28. https://doi.org/10.1007/BF02294778

Douglas, J., & Cohen, A. (2001). Nonparametric item response function estimation for assessing parametric model fit. *Applied Psychological Measurement*, *25*(3), 234-243.https://doi.org/10.1177/01466210122032046

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

Effatpanah, F., & Baghaei, P. (2021). Cognitive components of writing in a second language: An analysis with the linear logistic test model. *Psychological Test and Assessment Modeling*, *63*(1), 13-44. URL: https://www.psychologie aktuell.com/journale/psychological-test-and-assessment-modeling/currently-available.html

Effatpanah, F., & Baghaei, P. (2022, May 17-18). *Graphical kernel smoothing item response theory analysis for rater monitoring: The case of writing assessment*. 4th Conference on Interdisciplinary Approaches to Language Teaching, Literature, and Translation Studies. Ferdowsi University of Mashhad, Iran. https://doi.org/10.13140/RG.2.2.16421.78566

Effatpanah, F., Baghaei, P., & Boori, A. A. (2019). Diagnosing EFL learners' writing ability: A diagnostic classification modeling analysis. *Language Testing in Asia*, *9*(12), 1-23. https://doi.org/10.1186/s40468-019-0090-y

Engelhard, G. (2008). Historical perspectives on invariant measurement: Guttman, Rasch, and Mokken. *Measurement*, *6*(3), 155-189. https://doi.org/10.1080/15366360802197792

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

Engelhard, G., Jr., & Myford, C. M. (2003). *Monitoring faculty consultant performance in the Advanced Placement English Literature and Composition Program with a many-faceted Rasch model* (College Board Research Report No. 2003-1). New York: College Entrance Examination Board.

Engelhard, G., & Wind, S. A. (2018). *Invariant measurement with raters and rating scales: Rasch models for rater-mediated assessments*. New York, NY: Taylor & Francis.

Eubank, R. L. (1988). *Spline smoothing and nonparametric regression*. New York: Marcel Dekker.

Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory: Principles and applications*. Boston, MA: Kluwer Academic Publishers.

Härdle, W. (1990). *Applied nonparametric regression* (Econometric Society Monographs). Cambridge: Cambridge University Press. https://doi.org/10.1017/CCOL0521382483

Junker, B. W., & Sijtsma, K. (2001). Cognitive assessment models with few assumptions, and connections with nonparametric item response theory. *Applied Psychological Measurement*, *25*(3), 258-272. https://doi.org/10.1177/01466210122032064

Kubinger, K. D. (1989). Aktueller Stand und kritische Würdigung der Probabilistischen Testtheorie [Critical evaluation of latent trait theory]. In K. D. Kubinger (Ed.), *Moderne Testtheorie - Ein Abriß samt neuesten Beiträgen [Modern psychometrics – A brief survey with recent contributions]* (pp. 19-83). Munich: PVU.

Kuiken, F., & Vedder, I. (2014). Rating written performance: What do raters do and why? *Language Testing*, *31*(3), 329-348. https://doi.org/10.1177/0265532214526174

Lane, S. (2016). Performance assessment and accountability: Then and now. In C. Wells & M. Faulkner-Bond (Eds.). *Educational measurement: From foundations to future* (pp.356-372). New York: NY: Guilford Press.

Lane, S., & Stone, C. A. (2006). Performance assessment. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 387-431). Westport, CT: American Council on Education.

Lee, Y. S., Wollack, J. A., & Douglas, J. (2009). On the use of nonparametric item characteristic curve estimation techniques for checking parametric model fit. *Educational and Psychological Measurement*, *69*(2), 181-197. https://doi.org/10.1177/0013164408322026

Lei, P. -W., Dunbar, S. B., & Kolen, M. J. (2004). A comparison of parametric and nonparametric approaches to item analysis for multiple-choice tests. *Educational and Psychological Measurement*, *64*(4), 565-587. https://doi.org/10.1177/0013164403261760

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

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Lawrence Erlbaum Associates, Hillsdale, NJ.

Lu, Y., & Wang, X. (2006). *A hierarchical Bayesian framework for item response theory models with applications in ideal point estimation*. Technical report. Harvard University, Institute for Quantitative Social Science.

Mazza, A., Punzo, A., & McGuire, B. (2014). KernSmoothIRT: An R package for kernel smoothing in item response theory. *Journal of Statistical Software*, *58*(6), 1-34. Retrieved from http://www.jstatsoft.org/v58/i06/

Mazza, A., Punzo, A., & McGuire, B. (2020). *KernelSmoothIRT: Nonparametric Item Response Theory [Computer software]*. R package version 6.4. https://cran.rproject.org/web/packages/KernSmoothIRT/index.html

Meijer, R. R., & Baneke, J. J. (2004). Analyzing psychopathology items: A case for nonparametric item response theory modeling. *Psychological Methods*, *9*(3), 354-368. https://doi.org/10.1037/1082-989X.9.3.354

Meijer, R. R., Tendeiro, J. N., & Wanders, R. B. K. (2015). The use of nonparametric item response theory to explore data quality. In S. P. Reise & D. A. Revicki (Eds.), *Handbook of item response theory modeling: Applications to typical performance assessment* (pp. 85-110). New York, NY: Routledge.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13-103). New York, NY: Macmillan.

Mokken, R. J. (1971). *A theory and procedure of scale analysis*. The Hague, Netherlands: De Gruyter.

Molenaar, I. W. (2001). Thirty years of nonparametric item response theory. *Applied Psychological Measurement*, *25*(3), 295-299. https://doi.org/10.1177/01466210122032091

Molenaar, I. W., & Sijtsma, K. (2000). *User's Manual MSP5 for Windows*. IEC ProGAMMA, Groningen.

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement*, *16*(2), 159-176. https://doi.org/10.1177/014662169201600206

Muraki, E. (1997). A generalized partial credit model. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 153-164). Springer. https://doi.org/10.1007/978-1-4757-2691-6_9

Myford, C. M., & Wolfe, E. W. (2003). Detecting and measuring rater effects using many-facet Rasch measurement: Part I. *Journal of Applied Measurement*, *4*, 386-422. URL:http://jampress.org/abst.htm

Myford, C. M., & Wolfe, E. W. (2004). Detecting and measuring rater effects using many-facet Rasch measurement: Part II. *Journal of Applied Measurement*, *5*, 189-227. URL:http://jampress.org/abst.htm

Myford, C. M., & Wolfe, E. W. (2009). Monitoring rater performance over time: A Framework for detecting differential accuracy and differential scale category use. *Journal of Educational Measurement, 46*(4), 371-389. https://doi.org/10.1111/j.1745-3984.2009.00088.x

Patz, R. J., Junker, B. W., Johnson, M. S., & Mariano, L. T. (1999). The hierarchical rater model for rated test items and its application to large-scale educational assessment data. *Journal of Educational and Behavioral Statistics, 27*(4), 341-384. https://doi.org/10.3102/10769986027004341

Rajlic, G. (2020). Visualizing items and measures: An overview and demonstration of the kernel smoothing item response theory technique. *The Quantitative Methods for Psychology*, *16*(4), 363-375. https://doi.org/10.20982/tqmp.16.4.p363

Ramsay, J. O. (1991). Kernel smoothing approaches to nonparametric item characteristic curve estimation. *Psychometrika, 56*(4), 611-630. https://doi.org/10.1007/BF02294494

Ramsay, J. O. (2000). *Testgraf: A program for the graphical analysis of multiple-choice tests and questionnaire data*. Retrieved from http://www. psych.mcgill.ca/faculty/ramsay/ramsay.html

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

R Core Team (2019). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. Retrieved from http://www.R-project.org/

Samejima, F. (1968). Estimation of latent ability using a response pattern of graded scores. *ETS Research Bulletin Series, 1*, i-169. https://doi.org/10.1002/j.2333-8504.1968.tb00153.x

Santor, D. A., Ramsay, J. O., & Zuroff, D. C. (1994). Nonparametric item analyses of the beck depression inventory: Evaluating gender item bias and response option weights. *Psychological Assessment, 6*(3), 255-270. https://doi.org/10.1037/1040-3590.6.3.255

Schumacker, R. E. (2015). Detecting measurement disturbance effects: The graphical display of item characteristics. *Journal of Applied Measurement*, *16*, 76-81. URL:http://jampress.org/abst2015.htm

Shaw, S. D., & Weir, C. J. (2007). *Examining writing: Research and practice in assessing second language writing*. Cambridge: Cambridge University Press.

Sijtsma, K., & Meijer, R. R. (2007). Nonparametric item response theory and special topics. In C. R. Rao & S. Sinhary (Eds.), *Handbook of statistics: Psychometrics* (*Vol. 26*, pp. 719-747). Amsterdam: Elsevier.

Sijtsma, K., & Molenaar, I. W. (2002). *Introduction to nonparametric item response theory*. Thousand Oaks, CA: Sage.

Sijtsma, K., Emons, W. H., Bouwmeester, S., Nyklíček, I., & Roorda, L. D. (2008). Nonparametric IRT analysis of quality-of-life scales and its application to the world health organization quality-of-life scale (WHOQOL-Bref). *Quality of Life Research*,m*17*, 275-290. https://doi.org/10.1007/s11136-007-9281-6

van der Ark, L. A. (2012). New developments in Mokken scale analysis in R. *Journal of Statistical Software, 48*(5), 1-27. https://doi:10.18637/jss.v048.i05

van der Linden, W. J., & Hambleton, R. K. (1997). Item response theory: Brief history, common models, and extensions. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 1-28). Springer.

Weigle, S. C. (1998). Using FACETS to model rater training effects. *Language Testing, 15*(2), 263-287. https://doi.org/10.1177/026553229801500205

Wells, C. S., & Bolt, D. M. (2008). Investigation of a nonparametric procedure for assessing goodness-of-1t in item response theory. *Applied Measurement in Education*, *21*, 22–40. https://doi.org/10.1080/08957340701796464

Wind, S. A. (2015). Evaluating the quality of analytic ratings with Mokken scaling. *Psychological Test and Assessment Modeling, 57*(3), 423-444. URL:https://www.psychologie-aktuell.com/fileadmin/download/ptam/3-2015_20150925/07_Wind.pdf

Wind, S. A. (2019a). A nonparametric procedure for exploring differences in rating quality across test-taker subgroups in rater-mediated writing assessments. *Language Testing, 36*(4), 595-616. https://doi.org/10.1177/0265532219838014

Wind, S. A. (2019b). Nonparametric evidence of validity, reliability, and fairness for rater mediated assessments: An illustration using Mokken scale analysis. *Journal of Educational Measurement, 56*(3), 478-504. https://doi.org/10.1111/jedm.12222

Wind, S. A. (2020). Applying Mokken scaling techniques to incomplete rating designs for educational performance assessments. *Measurement: Interdisciplinary Research and Perspectives, 18*(1), 23-36. https://doi.org/10.1080/15366367.2019.1644093

Wind, S. A., & Engelhard, G. (2016). Exploring rating quality in rater-mediated assessments using Mokken scale analysis. *Educational and Psychological Measurement, 76*(4), 685-706. https://doi.org/10.1177/0013164415604704

Wind, S. A., & Patil, Y. J. (2018). Exploring incomplete rating designs with Mokken scale analysis. *Educational and Psychological Measurement, 78*(2), 319-342. https://doi.org/10.1177/0013164416675393

Wind, S. A., & Peterson, M. E. (2018). A systematic review of methods for evaluating rating quality in language assessment. *Language Testing, 35*(2), 161-192. https://doi.org/10.1177/0265532216686999

Wind, S. A., & Schumacker, R. E. (2017). Detecting measurement disturbances in rater-mediated assessments. *Educational Measurement: Issues and Practice, 36*(4), 44-51. https://doi.org/10.1111/emip.12164

Wind, S. A., & Schumacker, R. E. (2018). Exploring within-rater category ordering: A simulation study using adjacent-categories Mokken scale analysis. *Educational and Psychological Measurement, 78*(5), 887-904. https://doi.org/10.1177/0013164417724841

Wolfe, E. W., Chiu, C. W. T., & Myford, C. M. (2000). Detecting rater effects with a multifaceted Rasch rating scale model. In M. Wilson & G. Engelhard Jr. (Eds.), *Objective measurement: Theory into practice* (Vol. 5, pp. 147–164). Stamford, CT: Ablex Publishing.

Wright, B. D., & Masters, G. N. (1982). *Rating scale analysis*. Chicago, IL: MESA Press.

Yen, W. M., & Fitzpatrick, A. R. (2006). Item response theory. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 111–153). Westport, CT: American Council on Education/Praeger.