psych

MDPI

Article

Nonparametric Kernel Smoothing Item Response Theory Analysis of Likert Items

Purya Baghaei [1] and Farshad Effatpanah [2,*]

[1] Research and Analysis Unit, International Association for the Evaluation of Educational Achievement (IEA), 22297 Hamburg, Germany; purya.baghaei@iea-hamburg.de
[2] Research Unit of Psychological Assessment, Faculty of Rehabilitation Sciences, TU Dortmund University, 44227 Dortmund, Germany
* Correspondence: farshad.effatpanah@tu-dortmund.de

Abstract: Likert scales are the most common psychometric response scales in the social and behavioral sciences. Likert items are typically used to measure individuals' attitudes, perceptions, knowledge, and behavioral changes. To analyze the psychometric properties of individual Likert-type items and overall Likert scales, mostly methods based on classical test theory (CTT) are used, including corrected item–total correlations and reliability indices. CTT methods heavily rely on the total scale scores, making it challenging to directly examine the performance of items and response options across varying levels of the trait. In this study, Kernel Smoothing Item Response Theory (KS-IRT) is introduced as a graphical nonparametric IRT approach for the evaluation of Likert items. Unlike parametric IRT models, nonparametric IRT models do not involve strong assumptions regarding the form of item response functions (IRFs). KS-IRT provides graphics for detecting peculiar patterns in items across different levels of a latent trait. Differential item functioning (DIF) can also be examined by applying KS-IRT. Using empirical data, we illustrate the application of KS-IRT to the examination of Likert items on a psychological scale.

Keywords: Likert-type scale; classical test theory; parametric/nonparametric IRT models; kernel smoothing IRT; IRFs

Citation: Baghaei, P.; Effatpanah, F. Nonparametric Kernel Smoothing Item Response Theory Analysis of Likert Items. *Psych* **2024**, *6*, 236–259. https://doi.org/10.3390/psych6010015

Academic Editor: Alexander Robitzsch

Received: 9 January 2024
Revised: 14 February 2024
Accepted: 15 February 2024
Published: 19 February 2024

1. Introduction

A Likert-type scale, originally developed by Rensis Likert in 1932, is a standard psychometric response scale used to assess individuals' attitudes, perceptions, knowledge, and behaviors. As a type of rating scale, Likert scales are prevalent in survey research and are widely used in social sciences research to readily operationalize perceptions and personality traits. On such scales, respondents are required to express their level of agreement or disagreement with a series of declarative statements or items. A Likert item usually contains three to nine response options or categories that are described with descriptors such as 'agree', 'disagree', 'sometimes', 'often', etc. Agreement with positively worded statements indicates a higher level of the trait, and responses to negatively phrased statements are reverse-scored. The scores on all of the items are summed to obtain an overall scale score, which is deemed to indicate the level of a trait in individuals.

The psychometric properties of the individual Likert-type items and overall Likert scales have conventionally been evaluated with the methods of classical test theory (CTT), such as corrected item–total correlations and reliability indices. The drawback of CTT methods is that they are based on the total scale scores. They also fail to directly evaluate how well the response options or categories work across varying levels of the trait being measured [1].

On the other hand, item response theory (IRT) methods for the analysis of Likert items adopt an item-based approach to scale evaluation. IRT models are a family of mathematical models used to define the relationship between individuals' levels of a latent

trait and characteristics of items, called item parameters. The basic idea in IRT models is that educational and psychological attributes (e.g., anxiety/stress, knowledge, attitude, etc.) are abstract latent entities that can be measured if they are elicited through devices called tests [2,3]. In fact, the responses of individuals to items of a test are observable manifestations of the hypothesized latent trait.

2. Assumptions and Properties of Item Response Theory (IRT) Models

IRT models use individual items as the unit of measurement to obtain latent trait/ability scores [4]. A wide variety of parametric and nonparametric IRT models have been developed to describe how individuals respond to items. Although IRT models vary in the numbers and kinds of parameters, common to all IRT models are several key assumptions. These assumptions are (a) unidimensionality, which indicates that only one dominant ability or construct should underlie the responses of an individual to a set of items, (b) local independence, indicating that individuals' responses to items in a test are independent given a certain level of the expected latent trait/ability, and (c) monotonicity, indicating that as the latent trait's level increases, the probability of endorsing a correct answer or a higher response category increases as well.

In addition to these assumptions, there are two properties common to all IRT models. The first property is measurement invariance, indicating that item parameters should be the same in different populations of respondents. The second property is the item characteristic function, which predicts individuals' responses to items of a test based on their position on the latent trait continuum and on the items' parameters. The relationship between the latent trait and items or options can be characterized by the item response function (IRF), which is graphically shown with an item characteristic curve (ICC). IRFs are basically lines that depict the endorsement probability of a correct answer or a category as a function of a latent trait [5]. The shape of the IRF is an important quality criterion for evaluating item effectiveness, and it is more informative than CTT item–total correlations, that is, ICCs show the extent to which items conform to the expectations of a specific IRT model. IRFs visually display the exact locations on a latent trait where an item is effective and the regions where it is not. The steepness of the IRF is an indication of item discrimination and a key feature in evaluating item quality in IRT.

In contrast to CTT, IRT has the potential to estimate option characteristic curves (OCCs) for response categories [5]. OCCs display the relationship between endorsements of particular options as a function of the latent trait. If the probability of choosing a response option changes as a function of the latent trait, the option is effective, that is, it can discriminate between respondents in terms of their latent trait levels and state how the IRT assumptions can be met or even how they are verified. In other words, OCCs show the regions on the latent trait where an option or a category becomes most probable for an individual of a specific level. For an option to function properly, it should be the most probable (to mark) option at a certain level of the trait continuum and become less probable or have zero probability in other regions. If an option is not the most probable option at a certain part of the scale, it is considered to be obsolete and is a candidate for merging with adjacent options.

3. Parametric vs. Nonparametric IRT Models

Parametric IRT models, such as the one-parameter logistic IRT model—also known as the Rasch model [6,7]—and the two-parameter logistic IRT model [8,9], involve the logistic transformation of (ordinal) observed scores into interval measures by imposing a specific mathematical form for modeling the relationship between the probability of a correct answer, the latent trait, and some item parameters. When parametric IRT models are used to analyze scales, a number of strong assumptions and properties, including unidimensionality, local independence, monotonicity, and measurement invariance, should be met. If such assumptions and properties are violated, the outcomes from the models are considered inaccurate and unreliable for different uses of the assessment. As parametric

IRT models require strict assumptions, some researchers have indicated the inadequacy of such models when applied to noisy data in the social and behavioral sciences [10].

On the other hand, many researchers have applied nonparametric IRT models to analyze scales. Although nonparametric IRT models adhere to many assumptions of their parametric counterparts, such as unidimensionality, local independence, and monotonicity, they are less restrictive because they do not impose a specific mathematical form on the IRFs [11]. In nonparametric IRT models, ICCs are directly estimated from the data without assuming any functional shape for them. More specifically, ICCs are constructed from the proportions of individuals at different levels that endorse an item. IRFs can be of any shape, whether logistic or not. The only restriction on IRFs is the order restriction or monotonicity (e.g., any increase in the latent trait level should increase or does not decrease the probability of getting an item correct or endorsing a response option). That is, IRFs should be non-decreasing in θ or the latent trait, or a positive monotone relationship should exist between the latent trait and the correct response. Apart from this, IRFs can take any shape. Additionally, specific distributions for the latent trait are required in some parametric IRT models, but this is not necessary in nonparametric IRT models [11]. As van der Linden and Hambleton [12] argued, ICCs in nonparametric IRT models are more flexible and closer to the true ICCs than those given by parametric IRT models. The use of nonparametric models is promising in situations where assumptions of parametric models are untenable, and an ordinal scale is adequate for the further interpretation and use of test results [13,14]. According to Sijtsma and Meijer [11], nonparametric models offer a significant advantage over their parametric counterparts in graphically diagnosing an peculiarities in data through the examination of IRFs and the evaluation of the monotonicity assumption, which help analysts identify troublesome or misfitting items. Several researchers have already employed Mokken Scale Analysis (MSA; [15]), a well-known nonparametric IRT model, to assess psychological tests, including those with Likert items (see [10] for a comprehensive review of MSA applications). Their results have shown that nonparametric models, especially the MSA, can provide valuable insights into the functioning of scales.

4. Item Characteristic Curves in Parametric and Nonparametric IRT Models

The major goal in IRT is defining a mathematical model to describe the probability of giving a correct answer to an item—or selecting an option—as a function of the underlying latent trait θ. For example, in the two-parameter logistic IRT model, this relationship is specified as follows:

$$P_i(\theta) = \frac{e^{[a_i(\theta-\beta_i)]}}{1+e^{[a_i(\theta-\beta_i)]}} \quad (1)$$

where $P_i(\theta)$ expresses the probability that a person with ability θ gives a correct response to item i, a_i is the discrimination of item i, β_i is its difficulty, and e is a constant equal to 2.718, which is used as the base of the natural logarithm. Once the parameters a_i and β_i are estimated for each item, the probability of a correct response $P_i(\theta)$ for varying levels of the latent trait θ is computed.

In NIRT models, however, there is no mathematical function for estimating the probabilities at different levels of θ. Probabilities are computed directly from the data at certain ability levels for which data are available. To estimate an ICC for an item, restscore groups are first identified (the restscore is the total raw score excluding the item under consideration). Next, the proportions of examinees who have correctly answered the item in each restscore group are computed. These proportions are, in fact, the probabilities of answering the item correctly at different locations of θ. Then, the restscore groups and the probabilities are plotted against each other, and a nonparametric ICC is obtained. If the sample is small or even of average size, there is a chance that the restscore groups are small, and, consequently, the estimated probabilities become very unstable. In such circumstances, adjacent restscore groups are combined. It is expected that as the restscores go up, the probabilities also increase.

5. Kernel Smoothing Item Response Theory

Kernel Smoothing Item Response Theory (KS-IRT; [5]) is a nonparametric IRT approach for estimating IRFs. Unlike parametric IRT models, where the shape of the IRF is specified by item parameters a priori, the IRFs in KS-IRT are data-driven and exploratory. That is, no pre-specified shape is assumed for the IRFs. Nevertheless, the IRFs of the KS-IRT for dichotomous items, just like the IRFs of parametric IRT models, should be monotonically non-decreasing in θ, that is, as the level of θ increases, the probability of giving a correct response should also increase or remain constant. For polytomous items, IRFs illustrate the probability of respondents selecting a certain option on a given scale at various levels of the latent trait.

Ramsay [16] (pp. 25–26) enumerated different steps for estimating OCCs in KS-IRT. The first step is to assign a value or a score to each respondent using various methods, such as adding up the scores from each item to obtain the total score for each respondent for Likert-type items. The second step is to rank respondents based on the values or scores (with ranks within tied values being assigned randomly). The third step is to substitute ranks with the quantiles of a certain distribution (mainly the standard normal distribution). The fourth step is to sort respondents' response patterns according to the estimated ability rankings. Finally, in the last step, the association between the item response and the latent variable is estimated by smoothing the relationship between variable values and the standard normal quantiles. Smoothing is implemented at certain selected points. To put it simply, the probability of a correct response is calculated based on the observed proportion of people selecting the option at the selected points, known as evaluation points. Next, a trace line is generated by plotting points on the x-axis against corresponding probabilities on the y-axis. Kernel smoothing nonparametric regression is then used to smooth the IRF and directly estimate OCCs from the data [17,18]. In statistics, smoothing is utilized to create an approximate curve that attempts to capture important patterns in the data and reduce noise. Instead of using all of the data points, the smoothing technique uses local averaging to estimate the relationship between the latent variable and the probability of choosing an option. According to Rajlic [19], (pp. 373), "kernel is a weighting function, which assigns weights to the scores, based on their distance from the targeted score". Furthermore, for each selected point on the θ scale, a constant distance size, referred to as the bandwidth (h) that controls the width of the kernel around the point, is selected. Then, a weighted average is calculated for all data points falling within the specified bandwidth around the given point. Points in closer proximity to the evaluation point receive greater weights [1]. Rajlic [19] argued that "its [bandwidth] inappropriate selection can lead to over- or under-smoothing of the curve. Selection of bandwidth assumes a trade-off between estimation bias and variance—larger bandwidth for example leads to smaller variance but larger bias" (p. 373).

As stated above, the KS-IRT provides a graphical representation of how items function. The inspection of the resulting plots and curves (e.g., OCCs) provides diagnostic information on problematic items with regard to the monotonicity assumption, item discrimination across various levels of the expected construct, and DIF [19,20]. Consequently, KS-IRT can be considered an additional tool in the statistical toolkit of researchers in educational and psychological measurement. The major advantage of KS-IRT over CTT is that it focuses on the functioning of tests at the item level rather than at the level of the total test score. Numerical summaries of item discrimination in CTT do not provide any information about fluctuations in discrimination across the ability continuum, but the IRFs of KS-IRT do. Although parametric IRT models do provide IRFs, the major focus in parametric IRT is on numerical values and the assessment of statistical indices. On the contrary, KS-IRT does not provide any numerical summaries or any statistical indices, and everything is graphical. This helps to diagnose problems with the performance of items and analyze the scale later with the right parametric IRT model. Additionally, it allows practitioners to assess model fit and choose the appropriate parametric model for further data analysis [21,22]. For example, if the KS-IRT shows that items have different slopes, the two-parameter logistic IRT might

be the more appropriate model for the test, or if the items have non-zero lower asymptotes, the three-parameter logistic IRT could be a better modeling strategy for the test [19,23].

Along the same lines, Schumacker [24] showed the power of graphical displays in discovering misfitting items or patterns that cannot be observed through numerical model–data fit indices in parametric IRT models, such as the Rasch model. Wind and Schumacker [25] similarly indicated that a graphical display is a diagnostic way of identifying measurement disturbances. They argued that numerical values (e.g., fit statistics) tend to "mask patterns in residuals", and thus, in some cases, the results of numerical values do not show the correspondence between empirical and theoretical IRFs, and graphical displays become incongruent. Although graphical methods have the potential to provide precise information on exploring measurement disturbances, they are not commonly used in educational measurement. As Meijer et al. [13] noted, "there 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" (p. 89). Lei et al. [26] also stated that the comparison of graphical displays and numerical values allows researchers to better capture the performance of test items. The graphical outputs of nonparametric IRT models in general and KS-IRT in particular, in this case, would be more worthwhile to follow and interpret.

6. A KS-IRT Application

Despite the fact that KS-IRT holds significant promise in offering a purely graphical approach for assessing items and Likert-type scales, especially catering to researchers with limited mathematical expertise, very little research has been devoted to it in the social and behavioral sciences. The application of the KS-IRT has been confined to a narrow range of methodological [26–29] and practical [1,23,30–40] research in educational and psychological testing. Therefore, the main objective of this study is to showcase the usefulness of KS-IRT as a nonparametric approach for graphically examining the effectiveness of Likert items. Researchers have largely used MSA [15] and investigated its application in assessing rating scales [10,41]. However, MSA includes some practical limitations in operational assessments. The advantages of KS-IRT over MSA are that (1) in MSA, all items of a scale should have the same number of response categories, but in KS-IRT, items can have varying numbers of categories, and (2) evaluating DIF or measurement invariance is very cumbersome with MSA. The available computer programs and packages, such as MSP [42] and the R package *mokken* [43], do not accommodate the estimation of separate IRFs across subgroups. If such evaluations are desired for the examination of measurement invariance, IRFs should be estimated in each group separately and then be superimposed on a single graph. However, the DIF examination in KS-IRT is straightforward and simple. KS-IRT offers graphs for DIF analysis that concurrently display the IRFs of two groups. Substantial differences in the shapes of the curves across the groups indicate the presence of DIF. For the purpose of the present study, the following research questions were addressed:

1. How does KS-IRT enhance the analysis of Likert-type scale items?
2. How can KS-IRT be used to detect DIF in Likert-type scale items across different subgroups?

7. Method

7.1. Data

As a demonstration, the data analyzed in this study consisted of item responses of 297 Iranian undergraduate students to a cognitive test anxiety scale. This dataset was previously examined by Baghaei and Cassady [44] to validate the Persian translation of the short form of the revised Cognitive Test Anxiety Scale [45]. The scale included 17 items scored on a four-point ordered response rating scale commonly used in the study of test anxiety (e.g., [46]): 1 = *strongly disagree*, 2 = *disagree*, 3 = *agree*, and 4 = *strongly agree* (See Appendix A, Table A1). No item required reverse scoring, and higher scores represented higher levels of cognitive test anxiety. The total scores representing the anxiety levels of respondents ranged from 19 to

64. The Cronbach's alpha reliability of the scale was 0.75. There were 131 females (44.11%) and 166 males (55.89%). Their mean age was 22.32 (*SD* = 3.73), with Persian as their first language. The students were from two schools of engineering (*n* = 112) and humanities (*n* = 185) in a university located in the Northeast of Iran.

7.2. Data Analysis

The package "KernSmoothIRT" version 6.4 [47] in the R statistical software [48] was used to estimate the KS-IRT model. Using kernel smoothing techniques, KS-IRT was used to fit nonparametric ICCs and OCCs. The package offers a variety of exploratory plots designed for polytomous and dichotomous data at the item and test levels. The plots allow researchers to examine the entire measure, the individual items, the respondents, and different subgroups (see Effatpanah and Baghaei, [23], for a comprehensive tutorial on KS-IRT).

8. Results

8.1. Plot Methods at the Item Level

Figure 1 shows the option characteristic curves (OCCs) for two items of the Test Anxiety Scale. The OCCs illustrate the probability of endorsing an option (*y*-axis ranging from 0 to 1) at various locations on the latent trait dimension, where respondents are ranked (*x*-axis). On the OCCs, the vertical dashed lines indicate the points below which 5%, 25%, 50%, 75%, and 95% of the respondents fell with respect to their total scores. The position of the vertical lines is identical for all items. For example, the 75% line is dotted at the score 41 for the two items shown in Figure 1, indicating that 75% of the respondents fell below the total score of 41, and 25% of the respondents were in the range of 41 to 68. This showed that there may have been a relative positive skewness in the data, that is, a large number of respondents had low total scores, which represented low test anxiety levels.

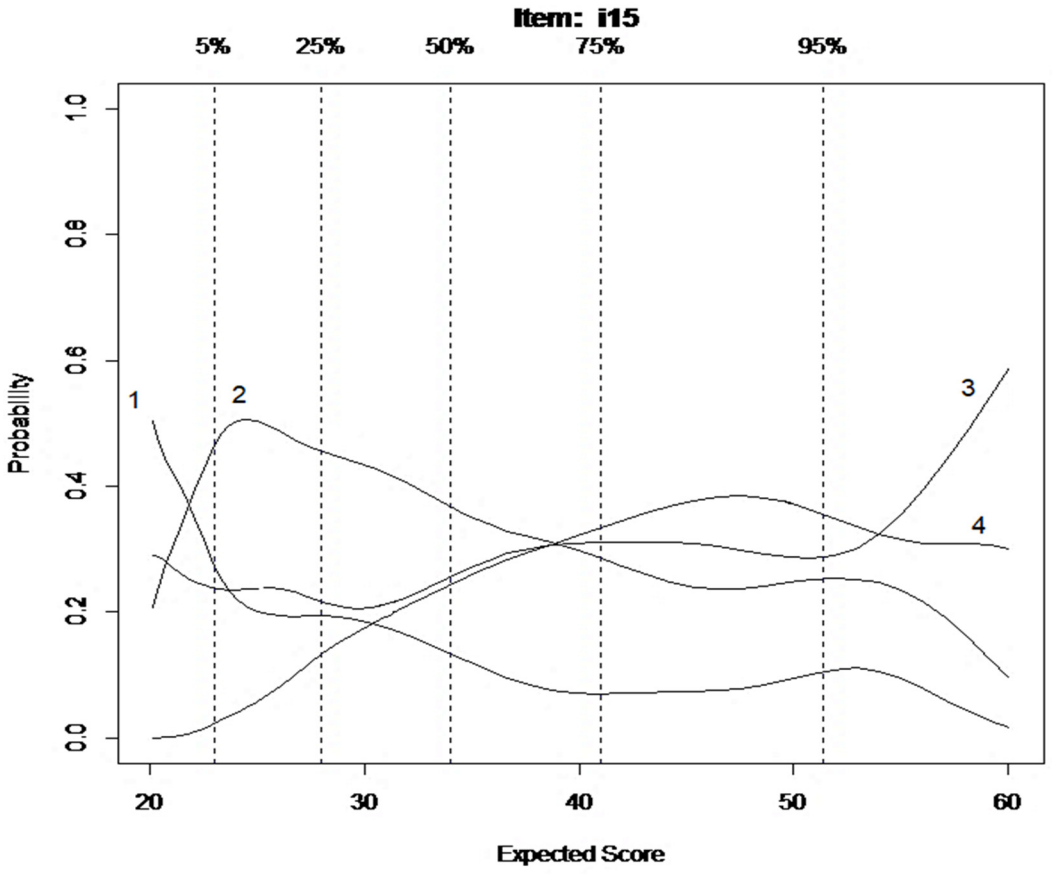


Figure 1. *Cont.*

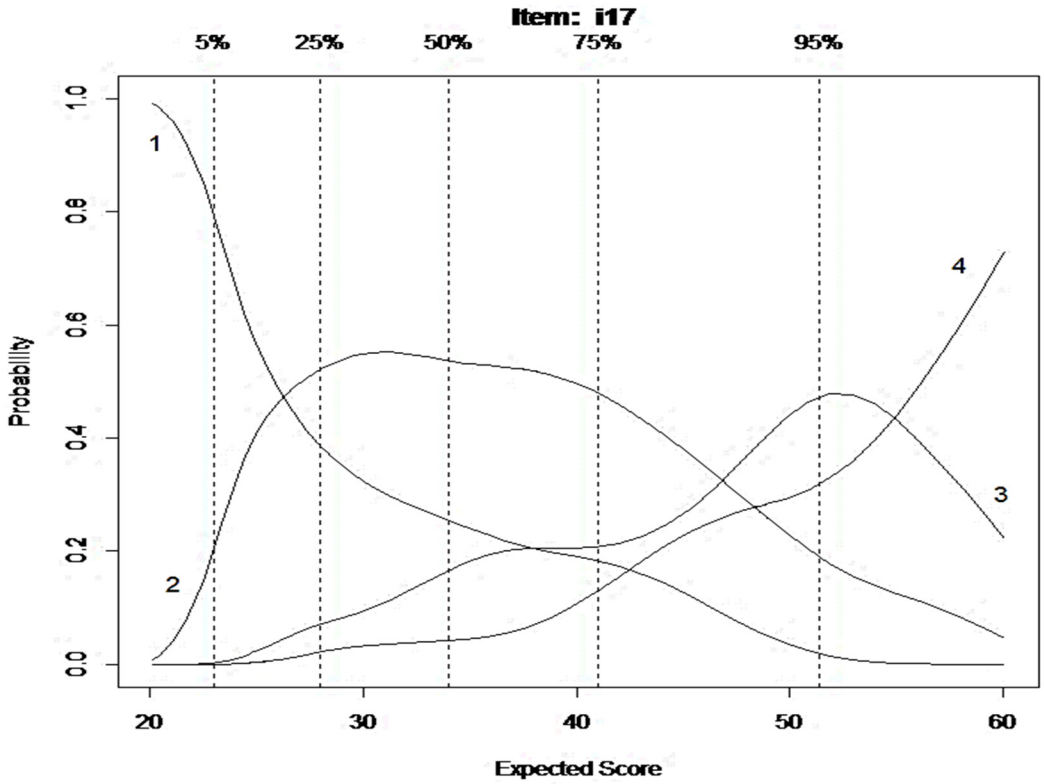


Figure 1. Option characteristic curves (OCCs) for two items of the Test Anxiety Scale.

As illustrated in Figure 1, four curves for each of the scale items were plotted, showing the four response options (e.g., 1 to 4) in the scale. The *x*-axis of the OCCs represents the expected score, ranging from 0 to 68, corresponding to the test anxiety level. The expected item score (EIS) is the average score that a respondent at a given theta level would achieve. For dichotomous items, it was the sum of the probabilities of a correct response on all the items at a given theta level. For polytomous items, it was the sum of the weighted probabilities of marking all of the categories on all of the items at a given theta level [16]. The probability of getting an item right or endorsing a particular category at different theta levels was estimated using the kernel smoothing function. According to the assumption of monotonicity, respondents with higher scores on the latent trait dimension had a higher probability of giving a correct answer to a test item or endorsing an option. In this example, an increase in the total scores on the *x*-axis indicated an increase in test anxiety for the respondents. In other words, respondents with higher expected scores on the *x*-axis were more likely to select higher response categories (e.g., options 3 and 4), and respondents with the lowest level of test anxiety were more likely to select lower response categories (e.g., options 1 and 2). Therefore, a satisfactory curve for polytomous items was expected to show the likelihood of respondents selecting a certain response category on the scale at various levels of the latent trait. In fact, OCCs should indicate the regions on the latent trait where a response category becomes most probable for a respondent of a specific level. An appropriate response category should be the most probable category at a specific level of the latent trait scale and become less probable or have zero probability in other regions. The response category will be inappropriate and a candidate for merging with adjacent options if it is not the most probable category at a specific region of the scale. Any peculiar shapes in the OCCs (e.g., a "wave" or a "U-shaped" curve) flag the violation of the assumption of monotonicity, which has a strong effect on the accuracy of measurement [49,50]. As shown in Figure 1, Item 17 illustrates that each response category was the most probable category for respondents at specific levels of the test anxiety scale, although Category 2 was very probable for a wide range of the trait continuum. However, the OCC for Item 15 indicated

that each response category was not the most probable category for respondents at certain levels of the Test Anxiety Scale.

More precisely, each response category should have been the most probable option for respondents at certain levels of the trait continuum. That is, the endorsement probability of the first category was expected to be high among respondents with the least test anxiety and gradually diminish as test anxiety levels increased. The probability of the lowest category should have been close to 1 at the lowest point of the trait continuum and should have decreased towards zero at the highest levels of the trait scale. Category 2 should have been the most probable option for those at low to medium levels of the trait and less probable for those outside this range of the trait. Category 3 was expected to be more likely for respondents with latent variable levels ranging from medium to high and less likely for those above or below this level. Finally, the highest category (e.g., 4) should have had a very low probability for those at low and medium levels and should have been very probable for those with very high trait levels.

In summary, ideal OCCs should look like a set of neat successive hills, with each representing a category. Each response option or category should exhibit a peak on the curve, indicating that the category is the most likely response for certain regions of the scale. As demonstrated in Figure 1, for Item 17, all of the options fell between the regions in which they were expected to be. That is, Option 2 fell between the areas of Options 1 and 3, and similarly, Option 3 fell between Options 2 and 4. However, the OCCs of Item 15 showed that the response options did not fall within their expected regions. Options 2 and 3 did not fall between Options 1 and 3 and Options 3 and 4, respectively. This indicated a problem with the performance of the scale. Therefore, the item was flagged for further investigation, and other characteristics of the item needed to be analyzed.

For clarification purposes, the performance of Item 17 as shown in Figure 1 was checked. Category 1 was more likely for those with expected anxiety scores between 19 and 26 and became very improbable as anxiety increased. For those above 26, Category 2 became very probable. This category remained highly probable for a very wide range of the scale. Category 3 was probable for a very narrow range of the scale. The OCCs for Item 17 showed that Category 2 was heavily favored. One would prefer to see the curves for Category 2 and Category 3 cross at a lower point on the scale, say at 40. To improve category functioning, it was better to break down Category 2 by adding a new category between Category 2 (*disagree*) and Category 3 (*agree*). A category such as *somewhat agree* could solve the problem. However, this needed to be tested.

Moreover, the steepness or slope of the OCCs offers information on the discriminatory effectiveness of categories and items. Item discrimination determines the rate at which the probability of a correct response or endorsing a category changes given the latent trait. In contrast to parametric IRT models, which provide a single-item discrimination index, KS-IRT allows practitioners to track and monitor variations in item discrimination along the expected latent trait continuum. This also allows for a visual comparison of all items concerning their discriminatory power at different levels of the latent trait, as the KS-IRT does not prescribe a specific shape for curves [19]. As the slope of the curves increases, the better the item can discriminate between respondents with different trait levels. As presented in Figure 1, Item 17 discriminated well between the respondents with lower and higher levels of test anxiety, especially those with expected total scores ranging from 21 to 25 for Option 2 and from 53 to 68 for Option 4. On the contrary, Item 15 lacked discriminatory power, highlighting its inefficiency in distinguishing between respondents with different levels of test anxiety. Specifically, respondents with higher anxiety levels exhibited the same probability of endorsing an option as those with lower anxiety levels.

8.2. Plot Methods at the Test Level

8.2.1. Principal Component Analysis (PCA)

A PCA plot for the items of the Test Anxiety Scale is shown in Figure 2. Items of the scale are represented by numbers inside the plot. The implementation of PCA of the

EISs at each evaluation point allowed us to simultaneously compare items and show the relationships among them [22]. As shown in Figure 2, there were two principal components. On the horizontal axis, the first principal component showed item difficulty in such a way that the easiest items were placed on the left and the most difficult items were placed on the right [22]. The small plots on the left and on the right represented the EISs for the highest option of the easiest and the most difficult items (e.g., the most extreme items). In this example, as can be seen, Item 11 was the easiest item, and Item 9 was the most difficult one. On the vertical axis, the second principal component showed item discrimination in such a way that items at the bottom of the plot tended to have a high positive slope, and items at the top of the plot tended to have a high negative slope [22]. The small plots at the top and the bottom represented EISs for the highest- and lowest-discriminating items. In this example, Item 6 had the highest discriminating power, and Item 15 has the lowest, meaning that it differentiated negatively.

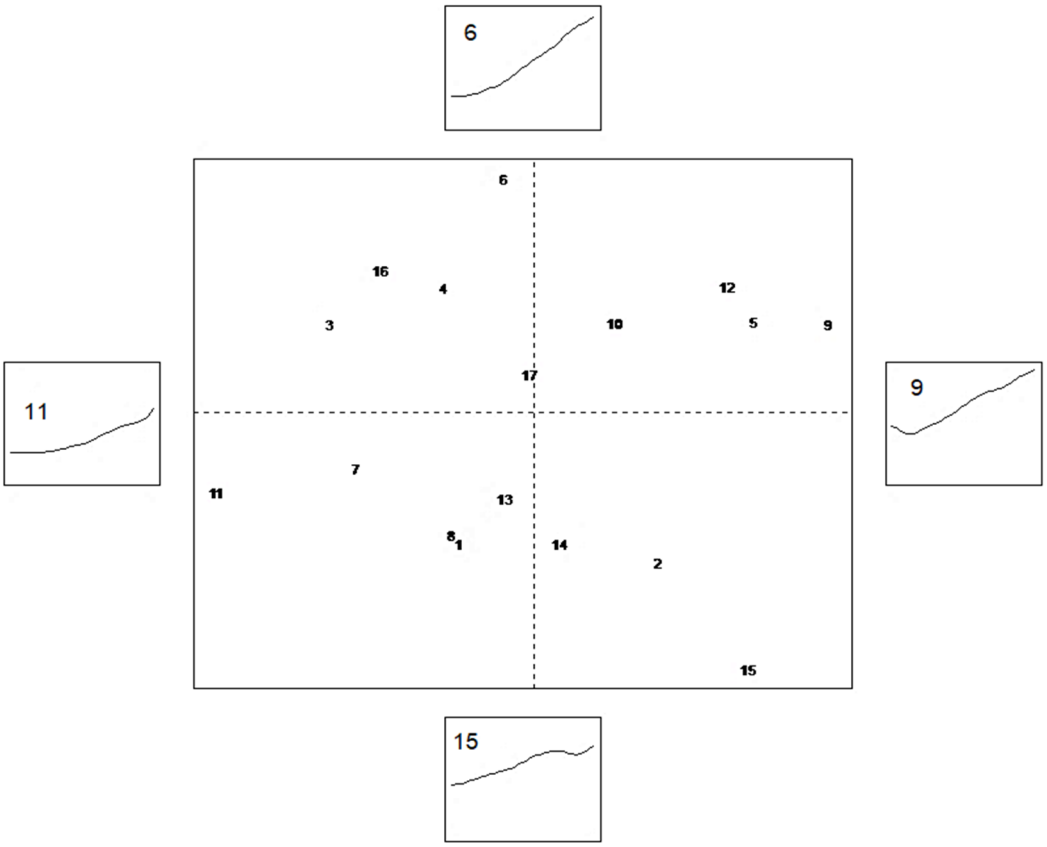


Figure 2. First two principal components for the Test Anxiety Scale.

8.2.2. Relative Credibility Curve (RCC)

Figure 3 illustrates the relative credibility curves (RCCs) for four respondents (i.e., 17, 45, 113, and 221). Using the response patterns of the respondents and the item OCCs, it was possible to compute the relative likelihood of the different values of theta. The theta value with the highest likelihood was considered as the maximum likelihood (ML) estimate of the ability of the respondents. Since the ML estimate of the ability took the respondents' response patterns and the characteristics of the items into consideration, it was a more accurate indicator of the latent ability than the sum score was [22]. RCCs generally indicate how precisely a total score reflects the ability of a respondent [16]. If the theta value with the highest credibility and the actual total score coincide, this means that the total score is an accurate indicator of the latent trait. However, if the total score and the theta do not coincide, it is a sign that the total score is inaccurate and does not represent the actual ability of the examinee. On the RCC plots, the vertical red line represents the actual total

score of the respondent, and the blue vertical dashed lines, similarly to OCCs, show the points below which 5%, 25%, 50%, 75%, and 95% of respondents fell in terms of their actual total scores. The width of the curve also shows the range where the respondent's true latent trait may lie, and the height of the curve with a maximum of 1.0 for a respondent shows the likelihood or the relative credibility of each theta value (e.g., true trait level). The pointier the curve, the more accurate the theta estimate is. If a respondent's actual score is to the right of the ML theta, this means that the respondent should have received a lower total score, whereas if the total score indicator is to the left of the ML theta, it is a sign that the respondent should have scored a higher total score. Furthermore, a bimodal RCC indicates that the respondent answered some hard items but failed some easy items [16]. This is a sign that either some guessing or random answering was involved, or the respondent has a higher level of the trait in some parts but has a lower level of the trait in other parts. Another possible reason for this phenomenon could be multidimensionality.

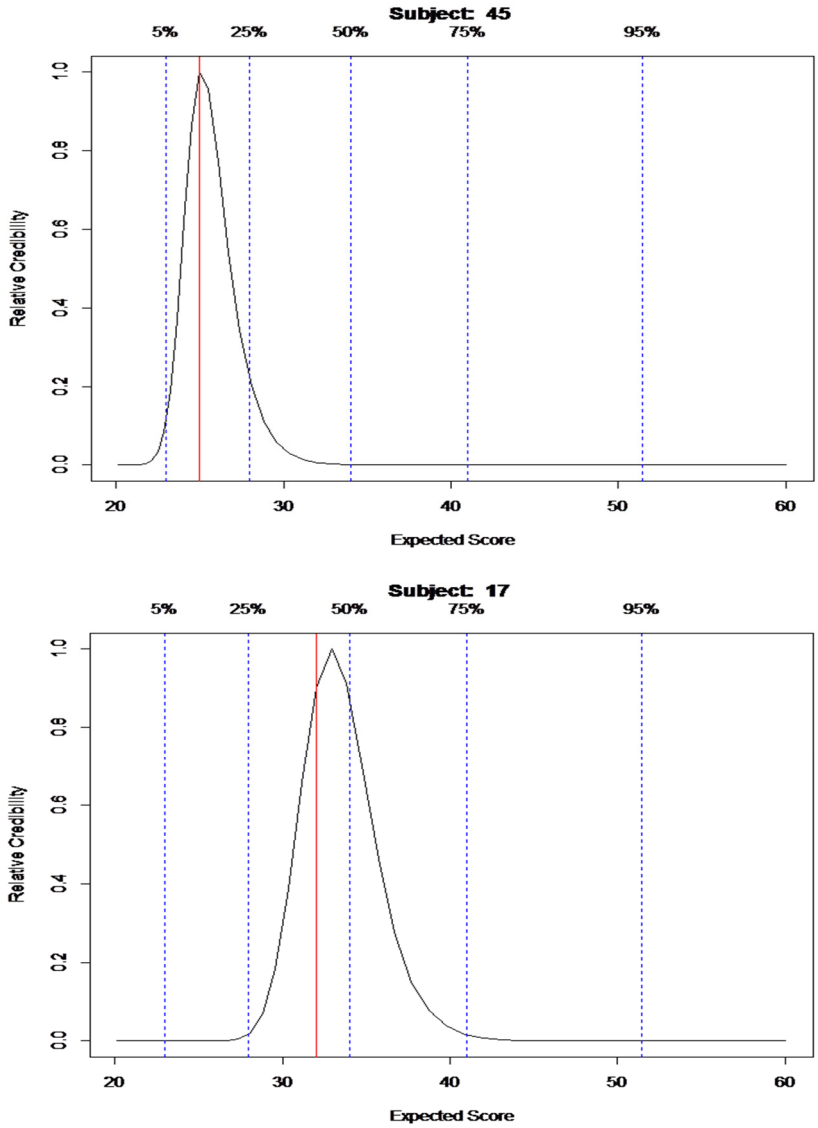


Figure 3. *Cont.*

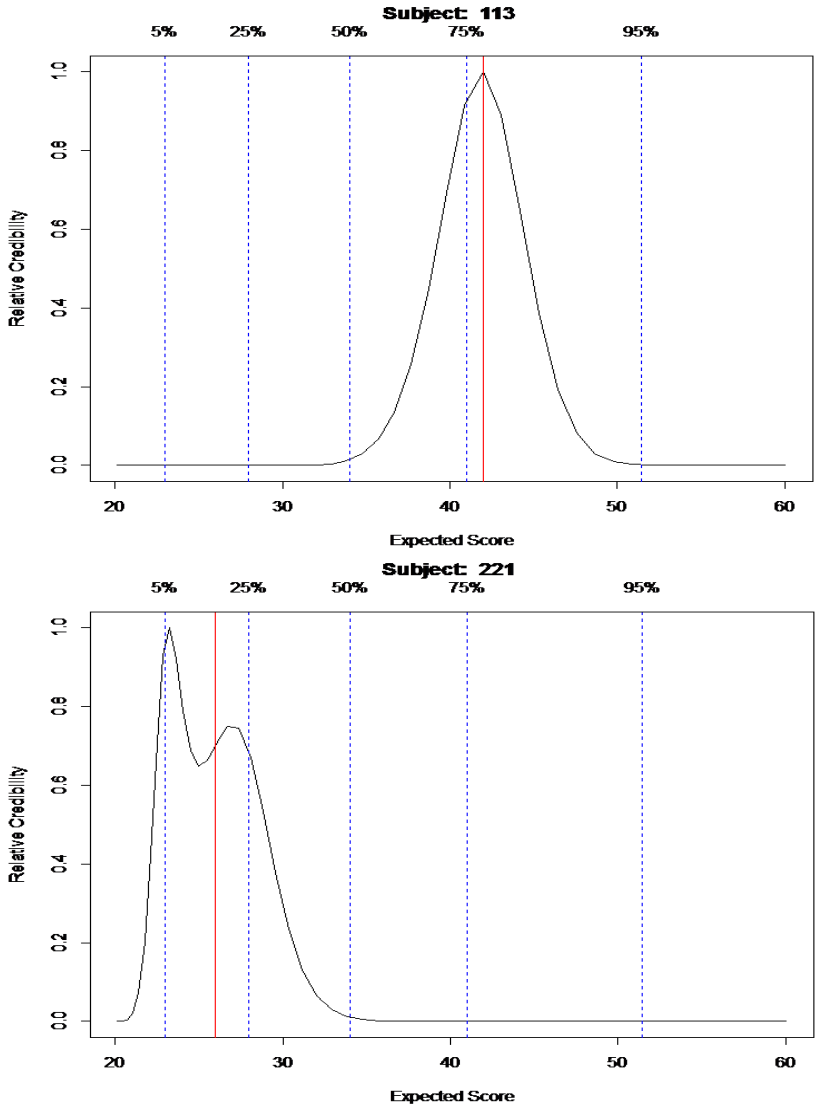


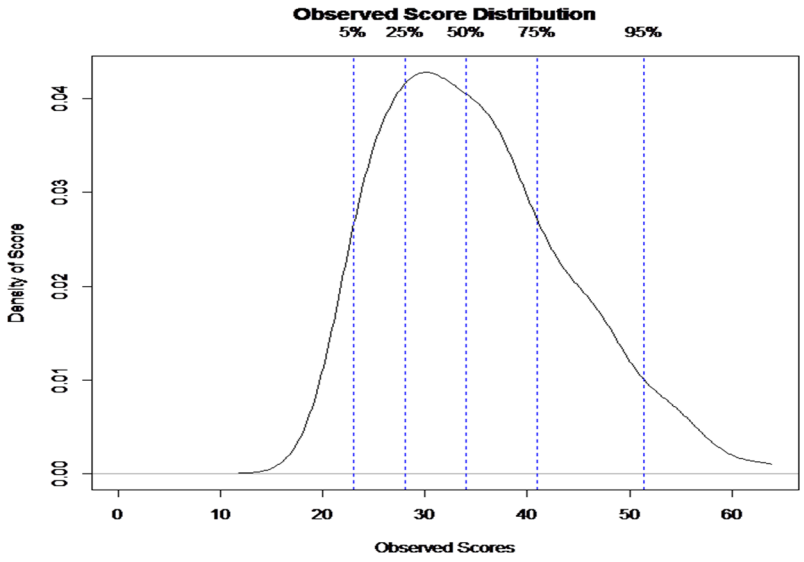
Figure 3. Relative credibility curves (RCCs) for Subjects 17, 45, 113, and 221.

As indicated in Figure 3, there was a substantial agreement between the total scores and the RCCs for Respondents 45 and 113, although the precision of the ML estimate was higher for Subject 45 than for Subject 113 because the RCC of Subject 45 was narrower, and the width of the curve was smaller. For Subject 113, the width of the curves indicated that, on the basis of the subject's total scores, his/her true anxiety was most likely between 34 and 49, while for Subject 45, it was most likely between 22 and 29. For Respondents 17 and 221, however, a substantial difference between the total scores and the maximum of the RCCs was observed. It can be seen that for Subject 17, the true anxiety level of the respondent was less than his/her current anxiety level based on his/her total score, suggesting a lower precision. For Subject 221, there were two peaks in the plot, indicating that the respondent had randomly selected some response categories of the scale, and thus, his/her true test anxiety level was higher than his/her current anxiety level.

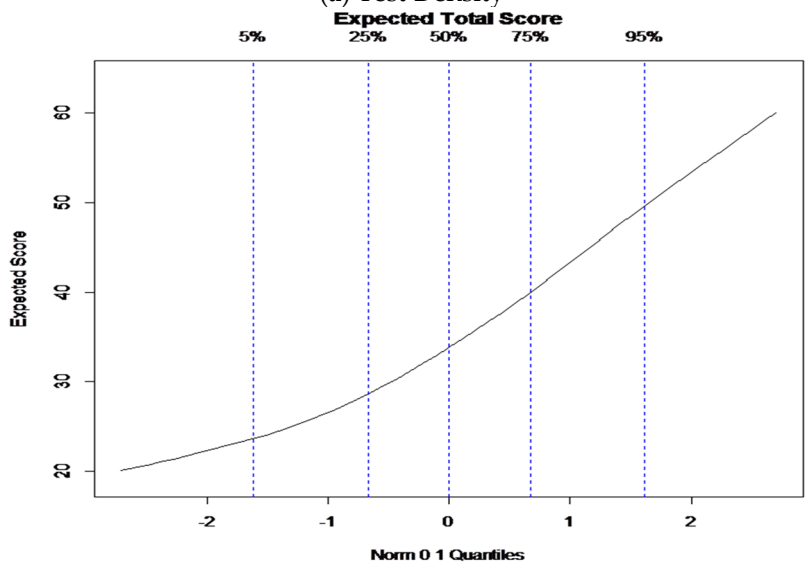
8.3. Test Summary Plots

Figure 4 displays three test-level summary plots for the Test Anxiety scale. A kernel density estimate of the distribution of the actual total score is presented in Figure 4a. This figure shows the extent to which the scores were probable, assuming that they were normally distributed (or bell-shaped). In fact, the density plot simply shows the distribution of the total scores, and based on the shape, it is almost a normal distribution—but not

quite. The density plot in Figure 4a shows that the assumption of normality was not met in the data, and scores in the range of 27 to 35 were most probable for the scale. As most of the observed scores were clustered around the left tail of the distribution, there was a positively skewed distribution in the data, reflecting that most of the respondents possessed low total scores or a low test anxiety levels. In Figure 4b, the expected anxiety scale scores are illustrated in relation to (or as a function of) the quantiles of the standard normal distribution. The curve was expected to be linear or monotonic, indicating that the assumption of monotonicity held at the test level. In this example, the curve was monotonic, suggesting that the monotonic requirement was met for the scale.

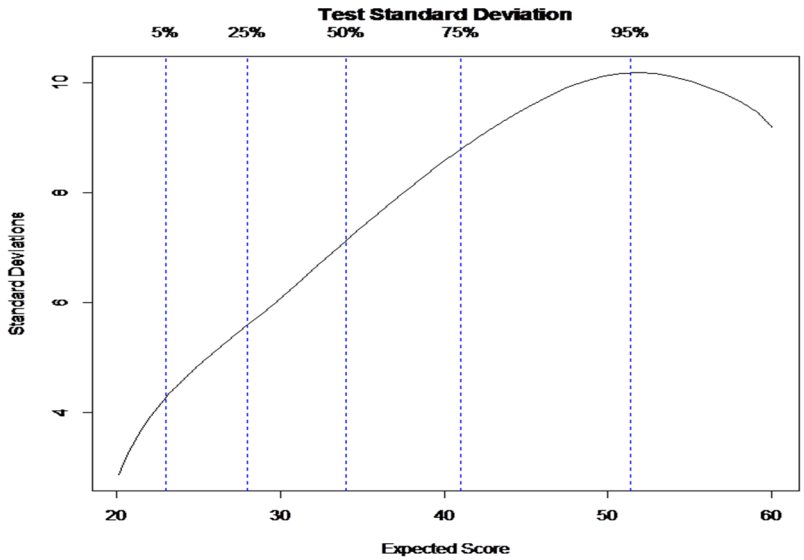


(**a**) Test Density



(**b**) Expected Test Score

Figure 4. *Cont.*



(**c**) Standard Deviation

Figure 4. Test summary plots for the Test Anxiety Scale.

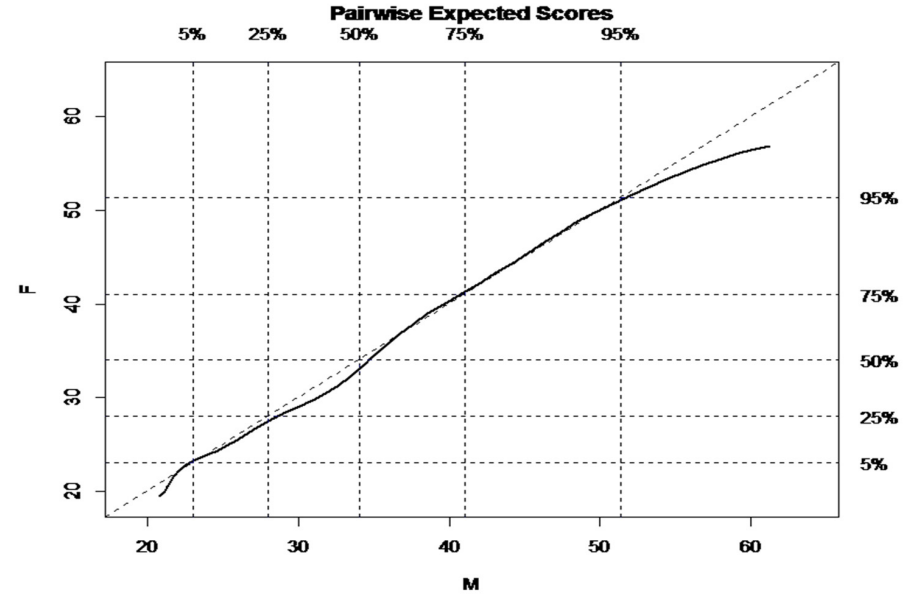
Figure 4c shows a test standard deviation graph or the standard error of measurement (SEM) for different levels of theta. The SEM is, in fact, the standard deviation of scores if a respondent takes a test an infinite number of times. In the literature on CTT, the mean of these repeated tests is called the true score, and their standard deviation is the error of measurement [2]. As can be seen in Figure 4c, the SD or SEM (on the vertical axis) reached the maximum for the respondents at around a total score of 51 (on the horizontal axis), where it was about 10. This translated into 95% confidence intervals of about 31 and 71 for a respondent who had an expected score of 51 ($51 \pm (10 \times 2)$), implying that a respondent with a score value of 51 could be 95% confident that his/her true score was somewhere between 31 and 71. These limits are very wide and, hence, indicate less precision. The graph suggests that the test was more precise for lower levels of test anxiety.

8.4. Plot Methods for Differential Item Functioning (DIF)

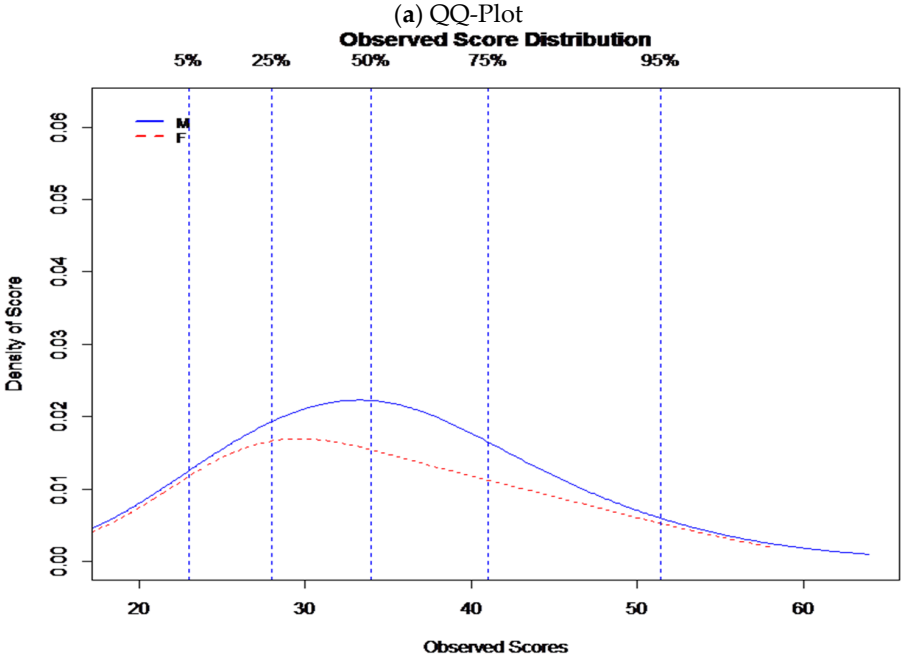
Differential item functioning (DIF) occurs when the items of a scale function differently for or against a particular group over another [51]. In other words, measurement invariance at the item level or DIF is present if respondents with the same level of a trait/ability from different groups have unequal probabilities of giving a correct response to an item or endorsing an option. A distinction is usually made between two types of DIF that may exist in practice: (a) Uniform DIF is a type of DIF in which the probability of getting an item right or endorsing an option is higher for one group than another group across all levels of the trait/ability. In fact, the difference between ICCs for the reference (e.g., the group hypothesized to have an unfair advantage) and focal group respondents (e.g., the group hypothesized to be disadvantaged by the test) remains constant or uniform across levels of the trait/ability. (b) Non-uniform DIF is a type of DIF in which the probability of getting an item right or endorsing an option is different for groups across levels of the trait/ability. In effect, the difference between the ICCs is not constant or uniform across levels of the trait/ability. In the KS-IRT approach, DIF is detected by analyzing curves that produce a visual display of item responses in different groups; that is, DIF or item bias can be examined through pairwise comparison of IRFs across different groups. Any considerable differences in the shapes of the curves across the groups and the sizes of the areas between them could indicate the presence of DIF in the scale [23,29].

Figure 5a shows the pairwise expected scores or QQ-plots of the distributions of the scores for males (on the *y*-axis) and females (on the *x*-axis). In the QQ-plots, the expected number of correct or total score values for any pair of subgroups, corresponding to the various standard normal quantiles, are plotted against each other, and they summarize

the differences in performance between the groups. The horizontal and vertical dashed lines indicate the 5%, 25%, 50%, 75%, and 95% quantiles for the two groups. When the two groups had almost the same performance, the relationship would emerge as a nearly diagonal line (a truly diagonal line is plotted as a reference) [16]. However, the solid line would deviate from the diagonal line if the groups exhibited varied performance. For the Test Anxiety scale, as presented in Figure 5a, there was a discrepancy between the two groups in terms of the distribution of their expected scores, especially towards the end of the trait continuum. By reading off the values on the plot, we found that an expected score of 60 for males corresponded to an expected score of 52 for females. Proper statistical analyses can also be performed to confirm if there really is DIF and to have an idea or estimate of its magnitude.



(**a**) QQ-Plot



(**b**) Kernel Density Function

Figure 5. The pairwise expected scores (QQ-plot) and kernel density functions for males and females on the Test Anxiety Scale.

Figure 5b, i.e., the density functions for the groups, shows differences in the performance of males and females on the anxiety scale. The plot confirms that there was a discrepancy in the behavior of males and females on the test, especially in the range of 28–45. This was an initial indicator of DIF, which required further investigation. Overall, the two plots (Figure 5a,b) suggest a strong disagreement in the behavior of the two groups based on their observed scores, indicating the presence of a substantial difference in the distribution of the scores between the two groups.

To compare the DIF at the item level, the OCCs for different response options for Item 8 of the Test Anxiety Scale across the two groups were examined. As illustrated in Figure 6, the red curves show the score distributions for female respondents, the blue curves show them for male respondents, and the black curves, as the overall curve, show them for all respondents. As can be seen, there was a substantial difference between the groups at the item level. With regard to Option 1, although there was no difference between the performance of the two groups at the higher end of the scale, females had a higher probability than males along most of the latent continuum. Concerning Option 2, the curves indicated that, in the lower range of the scale (e.g., scores between 20 and 28), males were more likely to endorse this option than females were. However, with the increase towards the middle and higher range, the probability of females endorsing the option became higher. For Option 3, the probability of endorsing the option was greater for males when situated in the lower to middle range of the dimension, whereas the probability increased for females as they moved towards the higher end of the scale. Finally, the curves for Option 4 showed that the probability of endorsing this option was higher for males along the dimension. The graphs for Options 1 and 4 suggested the presence of uniform DIF, and those for Options 2 and 3 were indicators of non-uniform DIF, although further statistical analyses are required to substantiate the results of these graphical displays.

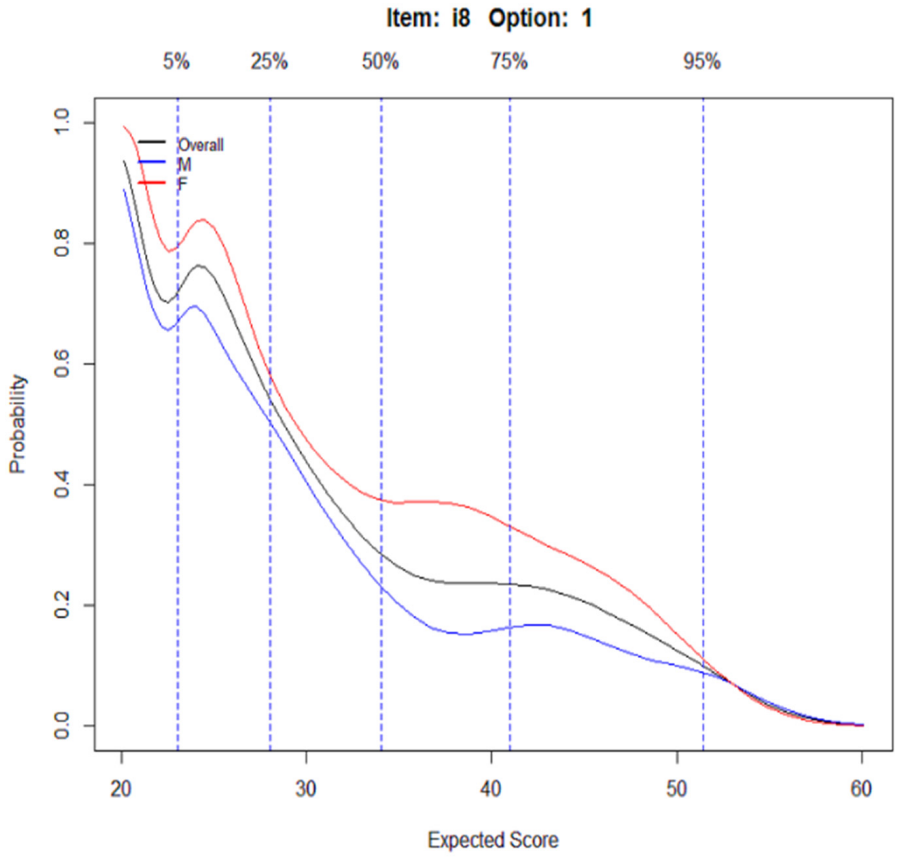


Figure 6. *Cont.*

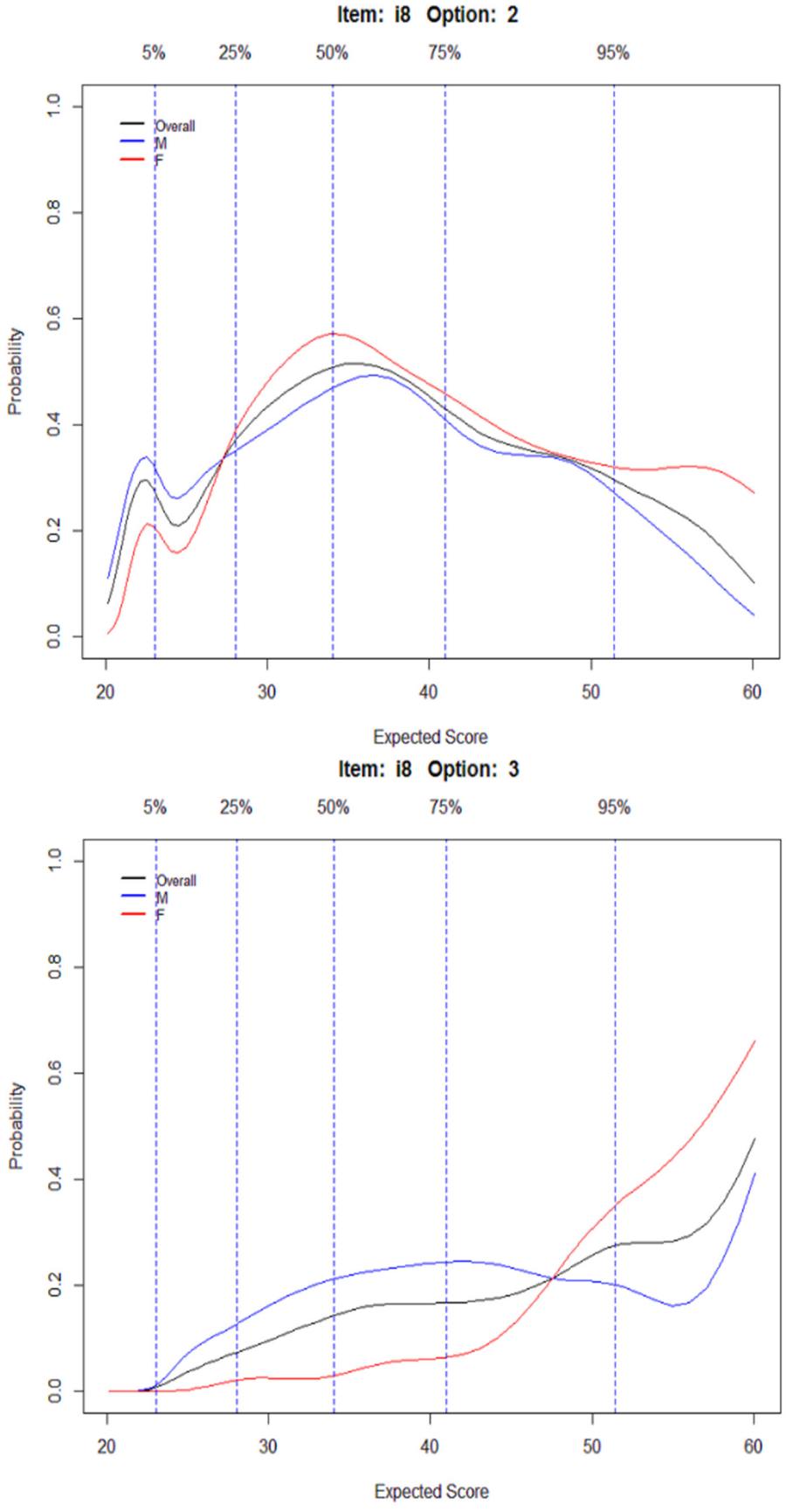


Figure 6. *Cont*.

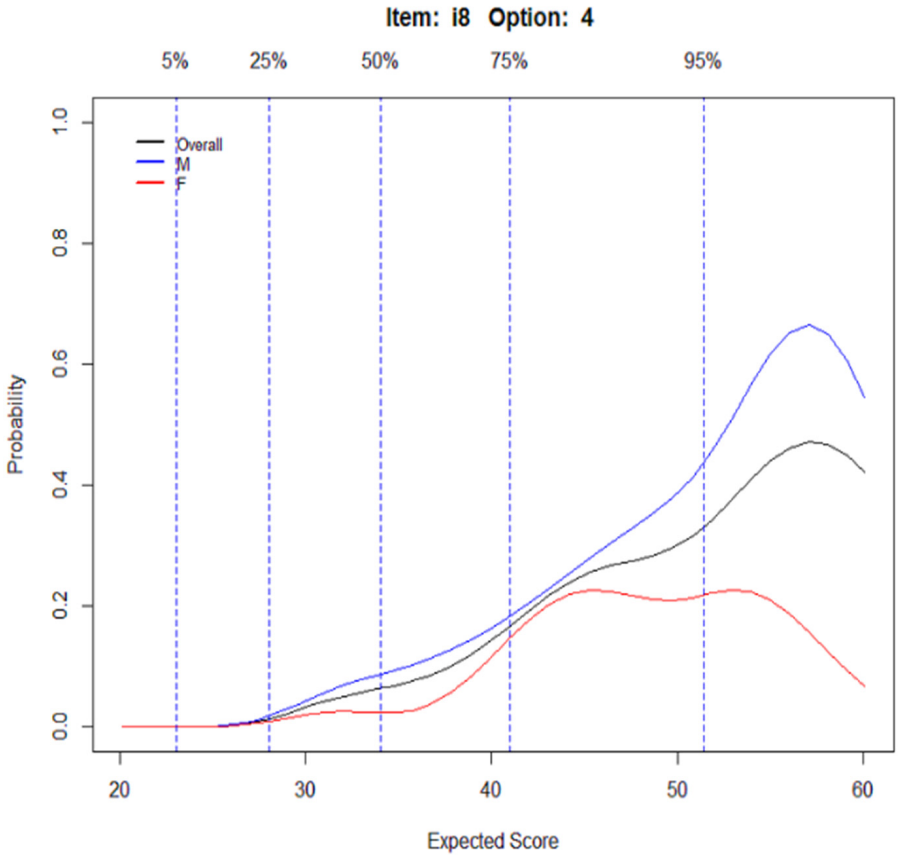
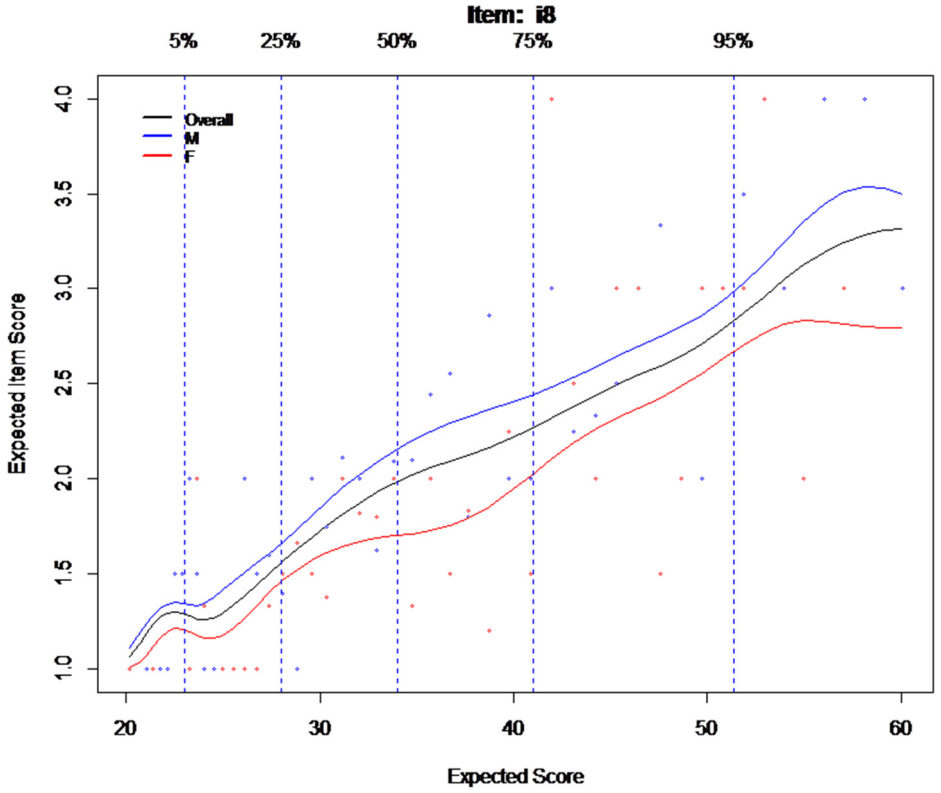


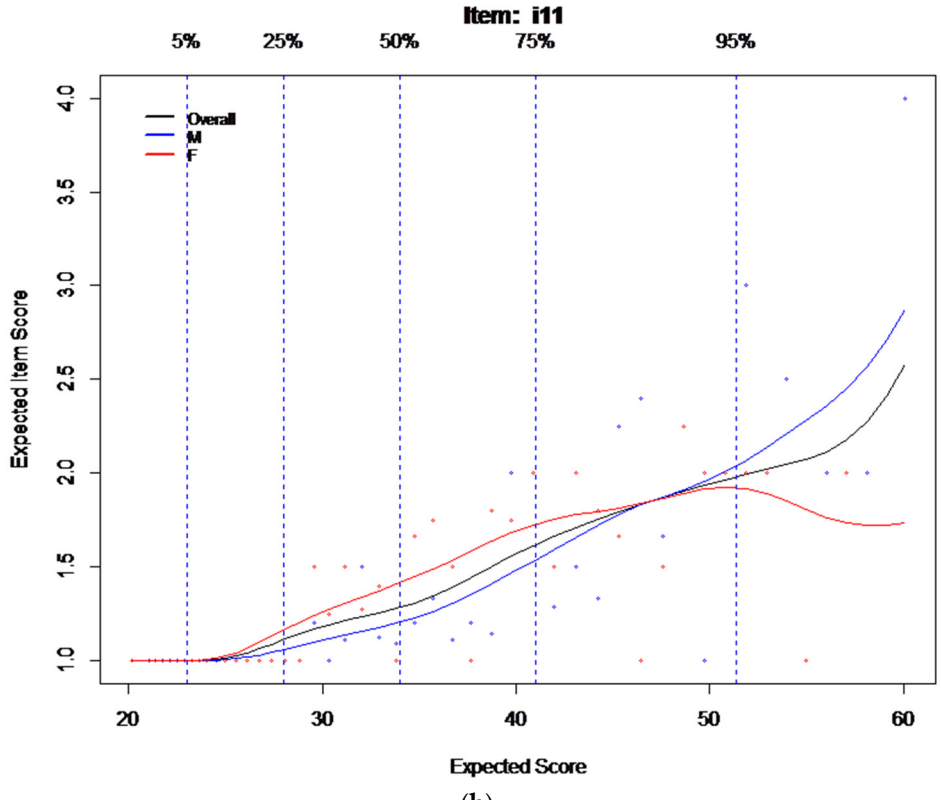
Figure 6. Option characteristic curves (OCCs) for females and males related to Item 8 of the Test Anxiety Scale (the curve for females is in red, and that for males is in blue).

Finally, Figure 7 shows the expected item score (EIS) plot for Items 8 and 11 of the Test Anxiety Scale across the two groups. On the graphs, the blue curve denotes the expected score for male respondents, the red curve denotes that for female respondents, and the black curve, the overall curve, denotes that for all respondents. The vertical dashed lines display the points below which 5%, 25%, 50%, 75%, and 95% of the respondents fell based on their total scores. Also, the differently colored points on the plots indicate how respondents from the groups actually scored on the items. As presented in Figure 7a, male respondents had greater expected scores than those of female respondents along the dimension for Item 8, representing uniform DIF. However, for Item 11 depicted in Figure 7b, the curves show that, from the lower to middle end of the scale, females had higher EISs (i.e., in the range of 26 to 45 expected scores) than those of males, whereas males demonstrated higher EISs at the upper end of the scale, indicating the presence of non-uniform DIF.

Overall, the analysis of the DIF graphs at both the test and item levels corroborated that gender was the main cause of DIF in the Test Anxiety Scale in this practical example. To confirm the results of graphical displays, the use of statistical analyses is further required.



(**a**)



(**b**)

Figure 7. Overall expected item scores (EISs) and EISs of females and males for Items 8 (**a**) and 11 (**b**) of the Test Anxiety Scale (the curve for females is in red, and that for males is in blue).

9. Discussion

The present study set out to illustrate and explore the application of the KS-IRT approach to the examination of the quality of Likert-type scales. In the social sciences and in survey research, Likert-type scales are the most popular response types of scales for measuring attitudes and perceptions. Studies investigating the psychometric characteristics of Likert-type scales have traditionally focused on estimates of scale reliability and corrected item–total correlations. These methods are mainly based on all-inclusive statistics that provide a single global (average) measure across levels of individual variation. Much more importantly, they do not take the functioning of the items and the response options across varying levels of the latent trait into account. As an alternative, parametric and nonparametric IRT models, "as response-centered approaches" ([50], p. 124), are used for the evaluation of the quality and score estimation of Likert-type scales [12,52,53]. Researchers typically use parametric IRT models such as the rating scale model (RSM; [54]) and the partial credit model (PCM; [55]) to analyze Likert scales or items with ordered categories. Although parametric models provide a useful framework for evaluating the psychometric quality of Likert scales, they involve a set of strict requirements that are more likely to be unreasonable in many practical assessment contexts. For that reason, many researchers have already used nonparametric IRT models, such as MSA [15], to investigate the quality of Likert items.

A neglected nonparametric IRT approach in educational and psychological measurement is KS-IRT [5]. KS-IRT can help researchers and practitioners in the social sciences to analyze the psychometric properties of measurement instruments without imposing a specific mathematical shape (e.g., the logistic ogive) to identify the expected relationship between the locations of respondents on the latent variable continuum and the probability of getting an item right or endorsing a response option. Similarly to MSA, this approach relies on several exploratory methods that practitioners can employ to realize the extent to which their measurement procedures align with the fundamental ordering characteristics, which are essential for using and interpreting total scores in the social sciences. As an exploratory data-driven IRT approach, KS-IRT has the potential to offer visual information about the functioning of both items and response options in a test. The graphical representations give initial feedback about the functioning of items and options. By analyzing visual displays of items and options, practitioners can identify poorly functioning items and options, check model fit, and find the appropriate parametric model for further data analysis [16,21]. The inspection of plots also allows practitioners to check whether the assumption of monotonicity is satisfied, whether items and options have adequate discrimination across the latent dimension, and if all items and options of a measure function similarly across different subgroups. Therefore, the use of KS-IRT can be a diagnostic tool for researchers within the framework of CTT and IRT to explore response options with unexpected poor behaviors [28,49,56,57]. Additionally, researchers have also indicated that KS-IRT can be utilized as an optimal scoring method. For instance, in an attempt to remove the sum score as an indicator of the intensity of experiences, such as symptom stress, Ramsay et al. [58] introduced a new model that showed "performance as a space with a metric structure by transforming probability into surprisal or information" (p. 347). The results indicated the effective performance of the model (e.g., standard errors of performance estimates were as small as a quarter of those of sum scores).

While acknowledging the useful characteristics of KS-IRT in identifying peculiar response behaviors, several limitations of nonparametric IRT compared to parametric IRT should be taken into consideration. Wind [59] enumerated some shortcomings for nonparametric IRT models, including MSA, which can be extended to KS-IRT in the following ways. First, KS-IRT is not able to parameterize item difficulties. An important advantage of the parametric IRT model is that after calibration, the calibrated items can be used to score the other datasets of the same scale collected from the same population. However, nonparametric IRT models need to be specified for each dataset. Second, the graphical outputs of KS-IRT without providing any numerical values make it difficult for

novice practitioners and researchers to interpret and reach a conclusive decision about the psychometric properties of the scales. More specifically, in terms of DIF analysis, since there are no specific criteria or boundaries for analyzing graphs, it is challenging for researchers to determine whether DIF should be recognized, although nonparametric IRT models help to identify DIF between groups [23]. Nevertheless, with experience and practice, researchers can improve their skills in interpreting graphs. Third, since nonparametric IRT models do not involve a specific parametric shape for IRFs, these models are unable to produce interval-level parameter estimates that are required for equating and devising computer-adaptive assessments, as well as further parametric investigations. Due to such limitations, some alternatives to KS-IRT based on semiparametric modeling that do not require the use of the unweighted sum score (or a transformation of it) in the nonparametric estimation have been proposed [60,61].

With regard to the third limitation of nonparametric IRT models, there are controversies among researchers. Some argue that this interpretation of IRT modeling is inaccurate. According to van der Linden [62], the ability variable θ in both parametric and nonparametric models can undergo arbitrary monotonic transformations. Therefore, it is incorrect to claim that certain models, such as the Rasch or the 2PL model, generate "interval-level" estimates while nonparametric models do not. On the other hand, other researchers have maintained that, in practice, non-parametric IRT models do not provide interval estimates of theta. The only estimates are the unweighted sum scores, which can be interpreted in an ordinal way. Under the additional assumptions that the latent trait follows a normal distribution (or at least has a known distribution) and the item response function has a parametric functional form, parametric IRT models provide numerical theta estimates, which are typically interpreted at an interval level. Grayson [63] and, later, others [64] showed that if the three assumptions of unidimensionality, monotonicity, and local independence hold and if the item scores are dichotomous, the unweighted sum score (X+) and the latent trait (theta) have a monotone likelihood ratio (MLR). A MLR implies that E(theta | X+) is nondecreasing in X+, which means that respondents who have a higher sum score have a higher expected value of theta. In other words, the unweighted sum score gives an expected ordering of theta. So, if for respondent A, X+ = 20, and for respondent B, X+ = 30, then the expected latent trait value of B is higher than the expected latent trait value of A. It is ordinal because we only speak in terms of "A has a lower expected latent trait value than B has" and we do not provide numerical estimates of theta, which is a requirement for an interval scale. In the Mokken scaling literature, it is an oft-cited point that NIRT models, including Mokken scaling, only provide ordinal-level scores (see [49,65,66]).

As a reviewer of an earlier draft of this paper argued, it is widely acknowledged within the psychometric community that the horizontal scale can be monotonically transformed, or "warped", at will, without affecting the crucial features of the graphs. That being the case, the abscissa of the graphs should not be considered as providing an interval scale measurement for the underlying scale. This suggests that while there is a possibility of transforming a graphical continuum into a mathematically linear scale, it is an artificial property only (anonymous reviewer, personal communication, 10 February 2024). As the reviewer suggested, to solve the scale-warping issue, the distance along a probability or any other curve, which is easily computed by adding small increments along the curve, is invariant with respect to monotone transformations of theta, a fact that appears most in fairly advanced calculus texts but can easily be proved. This invariance makes this arc length measure an ideal abscissa for graphical display.

Another point worth mentioning is that there is an exaggerated veneration for mathematical models within psychometric research. The shortcomings highlighted by Wind [59] are only relevant when the model closely aligns with the data. The majority of current mathematical models are designed with probability curves that approach either 0 or 1 at the extremes. However, the graphs presented in this paper suggest otherwise. Similarly, the curves of mathematical models are far smoother than is supported by even the relatively modest amount of data displayed in this paper.

Finally, as articulated by Shannon [67], the transformation of a probability value p by $s(p) = -\log(p)$ converts probability into information, and the distance along the information curve is exactly the interval (it even functions as a ratio scale) for which social scientists have all wished. Information in this case is about whatever the scale is measuring if the scale is sufficiently valid.

Regardless of the above-mentioned shortcomings and discussions, the findings of the present study indicate that KS-IRT provides valuable insights into the quality of Likert-type scales. This information can be beneficial for practitioners in various applications and interpretations that do not hinge on the stringent assumptions of parametric IRT models.

Author Contributions: Conceptualization, P.B. and F.E.; Writing Original Draft, P.B. and F.E.; Formal Analysis, F.E.; Editing the Draft, F.E.; Reviewing and Editing the Draft, P.B.; Supervision, P.B. All authors have read and agreed to the published version of the manuscript.

Funding: The authors received no specific funding for this work from any funding agencies.

Institutional Review Board Statement: As the research involved human participants, this study was conducted in accordance with the Declaration of Helsinki, and the protocol was approved by the Ethics Committee of Islamic Azad University, Mashhad Branch, Iran (institutional review board decision no. ds#910763; date of approval: 3 January 2023).

Informed Consent Statement: Not applicable.

Data Availability Statement: The data presented in this study are available from the corresponding author upon reasonable request.

Conflicts of Interest: The authors declare no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Appendix A

Table A1. The English version of the Cognitive Test Anxiety Scale.

No.	Items	Not at All Typical of Me (1)	Somewhat Typical of Me (2)	Quite Typical of Me (3)	Very Typical of Me (4)
1	Get distracted from studying by thoughts of failing				
2	Difficulty remembering what I studied				
3	Think that I am likely to fail				
4	Not good at taking tests				
5	So nervous that I often can't think straight				
6	Feel defeated before I even start				
7	Freeze up on final exams				
8	Thinking of the consequences of failing				
9	Nervousness causes me to make careless errors				
10	Mind goes blank				
11	I may not be too bright				
12	I forget facts I really know				
13	Not perform well on tests				
14	I feel I am not doing well				
15	My performance does not show how much I know				
16	My test performances make me believe that I am not a good student				
17	Don't have control over my test scores				

References

1. Santor, D.A.; Ramsay, J.O.; Zuroff, D.C. Nonparametric item analyses of the Beck Depression Inventory: Evaluating gender item bias and response option weights. *Psychol. Assess.* **1994**, *6*, 255–270. [CrossRef]
2. Baghaei, P.; Effatpanah, F. *Elements of Psychometrics*, 2nd ed.; Sokhan Gostar Publishing: Mashhad, Iran, 2022.

3. Reise, S.P.; Moore, T.M. Item Response Theory. In *APA Handbook of Research Methods in Psychology: Foundations, Planning, Measures, and Psychometrics*, 2nd ed.; Cooper, H., Coutanche, M.N., McMullen, L.M., Panter, A.T., Rindskopf, D., Sher, K.J., Eds.; American Psychological Association: Washington, DC, USA, 2023; pp. 809–835. [CrossRef]
4. Baker, F.B.; Kim, S.H. *Item Response Theory: Parameter Estimation Techniques*, 2nd ed.; Marcel Dekker: New York City, NY, USA, 2004.
5. Ramsay, J.O. Kernel smoothing approaches to nonparametric item characteristic curve estimation. *Psychometrika* **1991**, *56*, 611–630. [CrossRef]
6. Rasch, G. *Probabilistic Models for Some Intelligence and Attainment Tests (Expanded Ed.)*; University of Chicago Press: Chicago, IL, USA, 1960/1980.
7. Fischer, G.H. Derivations of the Rasch model. In *Rasch Models: Foundations, Recent Developments, and Applications*; Fischer, G.H., Molenaar, I.W., Eds.; Springer: New York, NY, USA, 1995; pp. 15–38.
8. Birnbaum, A. Some latent trait models their use in inferring an examinee's ability. In *Statistical Theories of Mental Test Scores*; Lord, F.M., Novick, M.R., Eds.; Addison-Wesley: Boston, MA, USA, 1968; pp. 397–479.
9. Lord, F.M. *A Theory of Test Scores*; Psychometric Society: Iowa City, IA, USA, 1952.
10. Baghaei, P. *Mokken Scale Analysis in Language Assessment*; Waxmann Verlag: Münster, Germany, 2021.
11. Sijtsma, K.; Meijer, R.R. Nonparametric item response theory special topics. In *Handbook of Statistics: Psychometrics*; Rao, C.R., Sinhary, S., Eds.; Elsevier: Amsterdam, The Netherlands, 2007; Volume 26, pp. 719–747.
12. van der Linden, W.J.; Hambleton, R.K. Item response theory: Brief history, common models, and extensions. In *Handbook of Modern Item Response Theory*; van der Linden, W.J., Hambleton, R.K., Eds.; Springer: Berlin/Heidelberg, Germany, 1997; pp. 1–28. [CrossRef]
13. Meijer, R.R.; Tendeiro, J.N.; Wanders, R.B.K. The use of nonparametric item response theory to explore data quality. In *Handbook of Item Response Theory Modeling: Applications to Typical Performance Assessment*; Reise, S.P., Revicki, D.A., Eds.; Routledge: New York, NY, USA, 2015; pp. 85–110.
14. Molenaar, I.W. Parametric and nonparametric item response theory models in health related quality of life measurement. In *Statistical Methods for Quality of Life Studies*; Mesbah, M., Cole, B.F., Lee, M.L.T., Eds.; Springer: Berlin/Heidelberg, Germany, 2002. [CrossRef]
15. Mokken, R.J. *A Theory and Procedure of Scale Analysis*; De Gruyter: Berlin, Germany, 1971.
16. Ramsay, J.O. TestGraf: A Program for the Graphical Analysis of Multiple-Choice Tests and Questionnaire Data. 2000. Available online: http://www.psych.mcgill.ca/faculty/ramsay/ramsay.html (accessed on 30 September 2022).
17. Eubank, R.L. *Spline Smoothing and Nonparametric Regression*; Marcel Dekker: New York, NY, USA, 1988.
18. Härdle, W. *Applied Nonparametric Regression (Econometric Society Monographs)*; Cambridge University Press (CUP): Cambridge, UK, 1990. [CrossRef]
19. Rajlic, G. Visualizing items and measures: An overview and demonstration of the Kernel Smoothing item response theory technique. *Quant. Methods Psychol.* **2020**, *16*, 363–375. [CrossRef]
20. Yessimov, B.; Hussein, R.A.; Mohammed, A.; Hassan, A.Y.; Hashim, A.; Najeeb, S.S.; Mohammed Ali, Y.; Abdullah, A.; Afif, N.S. Detecting measurement disturbance: Graphical illustrations of item characteristic curves. *Int. J. Lang. Test.* **2023**, *13*, 126–133. [CrossRef]
21. Lee, Y.-S.; Wollack, J.A.; Douglas, J. On the use of nonparametric item characteristic curve estimation techniques for checking parametric model fit. *Educ. Psychol. Meas.* **2009**, *69*, 181–197. [CrossRef]
22. Mazza, A.; Punzo, A.; McGuire, B. KernSmoothIRT: An R package for kernel smoothing in item response theory. *J. Stat. Softw.* **2014**, *58*, 1–34. [CrossRef]
23. Effatpanah, F.; Baghaei, P. Kernel smoothing item response theory in R: A didactic. *Pract. Assess. Res. Eval.* **2023**, *28*, 7. [CrossRef]
24. Schumacker, R.E. Detecting measurement disturbance effects: The graphical display of item characteristics. *J. Appl. Meas.* **2015**, *16*, 76–81. Available online: http://jampress.org/abst2015.htm (accessed on 30 September 2022).
25. Wind, S.A.; Schumacker, R.E. Detecting measurement disturbances in rater-mediated assessments. *Educ. Meas. Issues Pr.* **2017**, *36*, 44–51. [CrossRef]
26. Lei, P.-W.; Dunbar, S.B.; Kolen, M.J. A comparison of parametric and nonparametric approaches to item analysis for multiple-choice tests. *Educ. Psychol. Meas.* **2004**, *64*, 565–587. [CrossRef]
27. Douglas, J. Joint consistency of nonparametric item characteristic curve and ability estimation. *Psychometrika* **1997**, *62*, 7–28. [CrossRef]
28. Douglas, J.; Cohen, A. Nonparametric item response function estimation for assessing parametric model fit. *Appl. Psychol. Meas.* **2001**, *25*, 234–243. [CrossRef]
29. Wells, C.S.; Bolt, D.M. Investigation of a nonparametric procedure for assessing goodness-of-fit in item response theory. *Appl. Meas. Educ.* **2008**, *21*, 22–40. [CrossRef]
30. Beevers, C.G.; Strong, D.R.; Meyer, B.; Pilkonis, P.A.; Miller, I.W. Efficiently assessing negative cognition in depression: An item response theory analysis of the Dysfunctional Attitude Scale. *Psychol. Assess.* **2007**, *19*, 199–209. [CrossRef]
31. Effatpanah, F.; Baghaei, P. Exploring rater quality in rater-mediated assessment using the non-parametric item characteristic curve estimation. *Psychol. Test Assess. Model.* **2022**, *64*, 216–252. Available online: https://www.psychologieaktuell.com/journale/search/ergebnis.html?tx_news_pi1[news]=4662&cHash=f7937aa4c21aa9117c0bdca24faee426 (accessed on 30 September 2022).

32. Effatpanah, F.; Baghaei, P. Graphical kernel smoothing item response theory analysis for rater monitoring: The case of writing assessment. In Proceedings of the 4th Conference on Interdisciplinary Approaches to Language Teaching, Literature, and Translation Studies, Ferdowsi University of Mashhad, Mashhad, Iran, 17–18 May 2022.
33. Gos, E.; Sagan, A.; Raj-Koziak, D.; Skarzynski, P.H.; Skarzynski, H. Differential item functioning of the tinnitus handicap inventory across gender groups and subjects with or without hearing loss. *Int. J. Audiol.* **2023**, *62*, 1–9. [CrossRef] [PubMed]
34. Khan, A.; Lewis, C.; Lindenmayer, J.-P. Use of non-parametric item response theory to develop a shortened version of the Positive and Negative Syndrome Scale (PANSS). *BMC Psychiatry* **2011**, *11*, 178. [CrossRef]
35. Lynch, K. Kernel Smoothing Item Response Theory Approach Applied to a Multiple-Choice Final Exam in Introductory Statistics. Master's Thesis, Ball State University, Muncie, IN, USA, 2020. Available online: http://cardinalscholar.bsu.edu/handle/20.500.14291/202475 (accessed on 30 September 2022).
36. Meijer, R.R.; Baneke, J.J. Analyzing psychopathology items: A case for nonparametric item response theory modeling. *Psychol. Methods* **2004**, *9*, 354–368. [CrossRef]
37. Rosa Franco, V.; Wiberg, M.; Sousa Bastos, R.V. Nonparametric item response models: A comparison on Rcovering true score. *Psico-USF* **2023**, *28*, 685–696. Available online: https://www.scielo.br/j/pusf/a/FdmS4m7gtDCZhh8nxrdYvrp/ (accessed on 30 September 2022). [CrossRef]
38. Sijtsma, K.; Emons, W.H.M.; Bouwmeester, S.; Nyklíček, I.; Roorda, L.D. Nonparametric IRT analysis of Quality-of-Life Scales and its application to the World Health Organization Quality-of-Life Scale (WHOQOL-Brief). *Qual. Life Res. Int. J. Qual. Life Asp. Treat. Care Rehabil.* **2008**, *17*, 275–290. [CrossRef]
39. Sueiro, M.J.; Abad, F.J. Assessing goodness of fit in item response theory with nonparametric models: A comparison of posterior probabilities and Kernel-smoothing approaches. *Educ. Psychol. Meas.* **2011**, *71*, 834–848. [CrossRef]
40. Wallmark, J.; Josefsson, M.; Wiberg, M. Efficiency analysis of item response theory Kernel equating for mixed-format tests. *Appl. Psychol. Meas.* **2023**, *47*, 496–512. [CrossRef]
41. Tabatabaee-Yazdi, M.; Motallebzadeh, K.; Baghaei, P. A Mokken scale analysis of an English reading comprehension test. *Int. J. Lang. Test.* **2021**, *11*, 132–143. Available online: https://www.ijlt.ir/article_130373.html (accessed on 30 September 2022).
42. Molenaar, I.W.; Sijtsma, K. *User's Manual MSP5 for Windows*; IEC ProGAMMA: Groningen, The Netherlands, 2000.
43. van der Ark, L.A. New developments in Mokken scale analysis in R. *J. Stat. Softw.* **2012**, *48*, 1–27. [CrossRef]
44. Baghaei, P.; Cassady, J. Validation of the Persian Translation of the Cognitive Test Anxiety Scale. *SAGE Open* **2014**, *4*, 2158244014555113. [CrossRef]
45. Cassady, J.C.; Finch, W.H. Confirming the factor structure of the Cognitive Test Anxiety Scale: Comparing the utility of three solutions. *Educ. Assess.* **2014**, *19*, 229–242. [CrossRef]
46. Sarason, I.G. Stress, anxiety, and cognitive interference: Reactions to tests. *J. Pers. Soc. Psychol.* **1984**, *46*, 929–938. [CrossRef]
47. Mazza, A.; Punzo, A.; McGuire, B. KernelSmoothIRT: Nonparametric Item Response Theory [Computer Software]. R Package Version 6.4. 2022. Available online: https://cran.rproject.org/web/packages/KernSmoothIRT/index.html (accessed on 30 September 2022).
48. R Core Team. *R: A Language and Environment for Statistical Computing*; R Foundation for Statistical Computing: Vienna, Austria, 2023; Available online: https://www.R-project.org/ (accessed on 30 September 2022).
49. Sijtsma, K.; Molenaar, I. *Introduction to Nonparametric Item Response Theory*; SAGE Publications Inc.: Thousand Oaks, CA, USA, 2002.
50. Wind, S.A. Monotonicity as a nonparametric approach to evaluating rater fit in performance assessments. *Meas. Interdiscip. Res. Perspect.* **2020**, *18*, 124–141. [CrossRef]
51. Zumbo, B.D. Three generations of DIF analyses: Considering where it has been, where it is now, and where it is going. *Lang. Assess. Q.* **2007**, *4*, 223–233. [CrossRef]
52. Hambleton, R.K. Principles selected applications of item response theory. In *Educational Measurement*; Linn, R.L., Ed.; Macmillan Publishing Co., Inc.: New York, NY, USA, 1989; American Council on Education; pp. 147–200.
53. Lord, F.M. *Applications of Item Response Theory to Practical Testing Problems*; Erlbaum: Hillsdale, NJ, USA, 1980.
54. Andrich, D. A rating formulation for ordered response categories. *Psychometrika* **1978**, *43*, 561–573. [CrossRef]
55. Masters, G.N. A Rasch model for partial credit scoring. *Psychometrika* **1982**, *47*, 149–174. [CrossRef]
56. Junker, B.W.; Sijtsma, K. Cognitive assessment models with few assumptions, and connections with nonparametric item response theory. *Appl. Psychol. Meas.* **2001**, *25*, 258–272. [CrossRef]
57. Stout, W. Nonparametric item response theory: A maturing and applicable measurement modeling approach. *Appl. Psychol. Meas.* **2001**, *25*, 300–306. [CrossRef]
58. Ramsay, J.; Li, J.; Wiberg, M. Better rating scale scores with information–based psychometrics. *Psych* **2020**, *2*, 347–369. [CrossRef]
59. Wind, S.A. A nonparametric procedure for exploring differences in rating quality across test-taker subgroups in rater-mediated writing assessments. *Lang. Test.* **2019**, *36*, 595–616. [CrossRef]
60. Falk, C.F.; Cai, L. Maximum marginal likelihood estimation of a monotonic polynomial generalized partial credit model with applications to multiple group analysis. *Psychometrika* **2016**, *81*, 434–460. [CrossRef] [PubMed]
61. Rossi, N.; Wang, X.; Ramsay, J.O. Nonparametric item response function estimates with the EM algorithm. *J. Educ. Behav. Stat.* **2002**, *27*, 291–317. [CrossRef]

62. van der Linden, W.J. (Ed.) Unidimensional logistic response models. In *Handbook of Item Response Theory: Volume One, Models*; CRC Press: Boca Raton, FL, USA, 2016; pp. 11–30. [CrossRef]
63. Grayson, D.A. Two-group classification in latent trait theory: Scores with monotone likelihood ratio. *Psychometrika* **1988**, *53*, 383–392. [CrossRef]
64. Huynh, H. A new proof for monotone likelihood ratio for the sum of independent Bernoulli random variables. *Psychometrika* **1994**, *59*, 77–79. [CrossRef]
65. Koopman, L.; Zijlstra, B.J.H.; Van der Ark, L.A. Evaluating model fit in two-level mokken scale analysis. *Psych* **2023**, *5*, 847–865. [CrossRef]
66. van der Ark, L.A. Mokken scale analysis in R. *J. Stat. Softw.* **2007**, *20*, 1–19. [CrossRef]
67. Shannon, C.E. A mathematical theory of communication. *Bell Syst. Tech. J.* **1948**, *27*, 379–423. [CrossRef]