Practical Assessment, Research & Evaluation

A peer-reviewed electronic journal.

Volume 28 Number 7, May 2023 ISSN 1531-7714

Kernel Smoothing Item Response Theory in R: A Didactic

Farshad Effatpanah, *Islamic Azad University, Mashhad Branch, Mashhad, Iran*
Purya Baghaei, *Islamic Azad University, Mashhad Branch, Mashhad, Iran*

Item response theory (IRT) refers to a family of mathematical models which describe the relationship between latent continuous variables (attributes or characteristics) and their manifestations (dichotomous/polytomous observed outcomes or responses) with regard to a set of item characteristics. Researchers typically use parametric IRT (PIRT) models to measure educational and psychological latent variables. However, PIRT models are based on a set of strong assumptions that often are not satisfied. For this reason, non-parametric IRT (NIRT) models can be more desirable. An exploratory NIRT approach is kernel smoothing IRT (KS-IRT; Ramsay, 1991) which estimates option characteristic curves by non-parametric kernel smoothing technique. This approach only gives graphical representations of item characteristics in a measure and provides preliminary feedback about the performance of items and measures. Although KS-IRT is not a new approach, its application is far from widespread, and it has limited applications in psychological and educational testing. The purpose of the present paper is to give a reader-friendly introduction to the KS-IRT, and then use the **KernSmoothIRT** package (Mazza et al., 2014, 2022) in R to straightforwardly demonstrate the application of the approach using data of Children's Test Anxiety scale.

Keywords: Non-parametric item response theory, kernel smoothing technique, option characteristic curves, KernSmoothIRT package, Children's Test Anxiety Scale

Introduction

Item response theory (IRT) refers to a theory of testing which describes the relationship between latent continuous variables (unobserved attributes or characteristics) and their manifestations (dichotomous/polytomous observed outcomes or responses) with regard to a set of item characteristics. Unlike the test-level focus of classical test theory (CTT), the name *item response theory* denotes the focus of the framework on individual items, as the unit of measurement (Baker & Kim, 2004). Therefore, IRT models the response of each subject of given ability to each item in a measure. The term *item* is generic and includes all kinds of item formats, such as nominal, partial credit, multiple-choice, multiple-response, and rating scale. IRT is based on the ideal that psychological and educational attributes (e.g., attitudes, knowledge, anxiety/stress, etc.) are abstract latent entities that can be measured when they are elicited through devices called tests (Baghaei & Effatpanah, 2022). More specifically, responses of subjects to items of a measure are considered as observable manifestations of the postulated latent variable.

A variety of IRT models have been developed, and a distinction has been made between parametric IRT (PIRT) models and non-parametric (NIRT) ones to model the encounter of an individual with a specified test item. Regardless of differences among IRT models in terms of number and kinds of parameters, they share a set of important assumptions: unidimensionality,

local independence, monotonicity, and measurement invariance (Hambleton & Swaminathan, 1985). The first assumption is unidimensionality indicating that there is only a single latent variable being measured, that is, all items of a scale should measure one single construct. Local independence assumes that the items of a measure should not be related to each other. That is, different item responses are independent conditioning on the underlying latent variable. The monotonicity assumption states that as the level of the latent variable is increasing, the probability of a correct answer also increases. Finally, measurement invariance posits that item parameters must be invariant in different populations of respondents.

Another common property between PIRT and NIRT models is item characteristic function which predicts responses of a subject to items of a measure in reference to their location on the latent variable continuum and parameters of items. The relation between the latent variable and the probability of getting an item right or endorsing a response option can be characterized by a monotonically increasing function (Rajlic, 2020). This function is known as *item response function* (IRF), graphically shown by an *item characteristic curve* (ICC). For PIRT models, a set of strict assumptions (e.g., unidimensionality, local independence, monotonicity, and measurement invariance) must hold for parameter estimation based on several fit statistics indicating the conformity of the observed responses to model expectations. PIRT models further involve the logistic transformation of observed responses to interval measures, by prescribing a prespecified mathematical shape for ICCs (logistic ogive form, Sigmoid, or S-shaped), with the models which are different in terms of the type of function is employed, the way ICCs are characterized, and the number of item parameters are proposed to specify the ICC (Birnbaum, 1968; Fischer, 1995; Lord, 1980; Rasch, 1960; Samejima, 1969).

However, NIRT models are less restrictive than PIRT models and do not prescribe any mathematical form on the IRFs. They can be of any shape whether logistic or not. ICCs are directly estimated from the data without imposing a particular shape (Junker & Sijtsma, 2001; Molenaar, 1997; Mokken, 1971; Ramsay, 1991; Sijtsma & Molenaar, 2002). The only restriction on the IRFs is the order restriction or monotonicity. The IRFs should be non-decreasing in θ, that is, a positive monotone relationship should exist between the latent variable and the correct response. Moreover, Sijtsma and Meijer (2007) note that in some PIRT models, specific distributions are required for the latent variable, but this is not necessary in NIRT models. For instance, Sijtsma and Meijer (2007) argue that unlike the IRFs of the two-parameter logistic model (2-PL model; Birnbaum, 1968), the IRFs of NIRT models "(1) need not be S-shaped, and need not even be smooth; (2) can have several inflexion points, and may be constant over several ranges of θ; and (3) can have lower asymptotes or minimum values greater than 0 and upper asymptotes or maximum values smaller than 1" (p. 722). The advantage of NIRT models is that many items which do not fit the PIRT models (because of the incongruity of their IRFs with the shape defined by the model) can still be kept in the scale as long as they are non-decreasing. Though they may not be optimal items, many of them may work at certain portions of the latent trait continuum. That is, they may be constant over a portion of the scale, but they may be increasing over other portions. Such items work for those portions of the scale over which they are increasing and thus contribute to test reliability. In other words, the contribution of the items to test reliability does not depend on the parametric shape of their IRFs. They only need to be non-decreasing. Furthermore, keeping more items improves the test coverage which is good for validity and protects the scale against construct underrepresentation (Messick, 1989).

Two most commonly used NIRT approaches are Mokken scale analysis (MSA; Mokken, 1971) and kernel smoothing item response theory (KS-IRT; Ramsay, 1991). Compared to MSA (Firoozi, 2021; Tabatabaee-Yazdi et al., 2021; see Baghaei, 2021, for a comprehensive review of MSA applications), too little attention has been paid to the application of KS-IRT in educational testing. The application of the approach has been limited to few practical (Beevers et al., 2007; Effatpanah & Baghaei, 2022a, 2022b, submitted; Khan et al., 2011; Meijer & Baneke, 2004; Santor et al., 1994; Sijtsma et al., 2008; Sueiro & Abad, 2011) and methodological (Douglas, 1997; Douglas & Cohen, 2001; Pui-wa et al., 2004; Wells & Bolt, 2008) research. The results of these studies have shown that NIRT models can provide valuable insights into the functioning of measures. The limited application of the KS-IRT could be due to the lack of familiarity of applied researchers with the theoretical structure of the

approach and the interpretation of its results. The computer software TestGraf (Ramsay, 2000) was firstly developed to implement the non-parametric estimation of option characteristic curves (OCCs) using kernel smoothing technique and related graphical analyses. The term 'OCCs' is used instead of ICCs to focus on the functioning of response options rather than test items. Recently, Mazza et al. (2014, 2022) developed and presented the convenient R package **KernSmoothIRT**. However, although their paper provides a comprehensive description about both the theoretical aspects of the approach and its applications, it may be difficult to understand for applied researchers who are not familiar with complex statistical concepts. Therefore, the main purpose of this paper is to provide a reader-friendly introduction to the KS-IRT for applied researchers, with the least use of technical terms, and then use the **KernSmoothIRT** package (Mazza et al., 2014, 2022) in R (R Core Team, 2023) to straightforwardly demonstrate the application of the approach using data of a Children's Test Anxiety scale.

Kernel Smoothing Item Response Theory

By proposing regression methods based on kernel smoothing techniques, Ramsay (1991, 1997) introduced what he called "kernel smoothing IRT (KS-IRT)" to provide a non-parametric estimation of OCCs, executed in the TestGraf program (Ramsay, 2000). KS-IRT does not present any numerical values and only provides graphical illustrations of how items of a measure function. This feature allows the KS-IRT, as an exploratory and data-driven technique, to have the diagnostic capacity to identify problems with the performance of the items, with just eyeballing the OCCs, which violate the important assumptions of measurement such as monotonicity, item discrimination across various levels of the expected construct, and measurement invariance and/or differential item functioning (DIF) (Rajlic, 2020). Furthermore, the approach helps researchers to evaluate model fit and select the most optimal parametric model for further data analysis (Lee et al., 2009; Mazza et al., 2014). If the KS-IRT, for instance, indicates that the items have different slopes, the 2-PL might be the more optimal model for the test, or if the items have non-zero lower asymptotes, the 3-PL could be a better modeling strategy for the test (Rajlic, 2020). Therefore, KS-IRT can be considered as an additional spanner in the statistical toolkit of researchers in psychological and educational measurement.

OCCs show the relation between the probability of choosing a particular response option for a test item and the ability of a subject. Let's consider the responses of a group of subjects $V = \{V_1, \ldots, V_j, \ldots, V_n\}$ to a set of test items $I = \{I_1, \ldots, I_i, \ldots, I_k\}$. Also, let's consider $O_i = \{O_{i1}, \ldots, O_{il}, \ldots, O_{im_i}\}$ as a set of options for I_i, and x_{il} as the weight ascribed to O_{il}. The observed response of V_j to I_i is shown by indicator variables $y_{il} = \{y_{il1}, \ldots, y_{ilm_i}\}$. If option m is selected by subject v, $y_{ilm} = 1$; otherwise, $y_{ilm} = 0$. In this case, the function of OCC can be expressed as:

$$P_{il}(\theta_v) = P(select\ O_{il} | \theta_v) = P(Y_{il} = 1 | \theta_v) \quad (1)$$

where $i = 1, \ldots, k$; $l = 1, \ldots, m_i$; and $P_{il}(\theta_v)$ is the probability that a subject v with unidimensionality ability level θ chooses option l for item i.

According to Ramsay (1991, p. 615) and Ramsay (2000, pp. 25-26), the estimation of OCCs involves the following steps: (1) *Score*: a value or score is assigned to each subject using different methods including computing the number of correct answers for each subject for multiple-choice items, computing the scale score for scales or mixed item types, which is the sum over items of the numerical weights related to the options selected, and reading in values from a file; (2) *Rank*: subjects are ranked based on the values or scores, with ranks within tied values assigned randomly; (3) *Enumerate*: the ranks are replaced by the quantiles q_v of a certain distribution (mostly standard normal distribution); (4) *Sort*: sort subjects' response patterns $(X_{i1}, \ldots, X_{il})$ by the estimated ability rankings; and (5) *Smooth*: the relationship between item response and the latent variable is estimated by smoothing the relationship between the 0-1 indicator variable values and the standard normal quantiles. Smoothing is implemented at certain selected points, known as evaluation points. Put simply, the probability of a correct response is computed as the observed proportion of people who selected the option at the selected points. Then, the points on the *x*-axis are plotted against the probabilities on the *y*-axis to obtain a trace line. Next, kernel smoothing non-parametric regression is used to smooth the IRF and directly estimate OCCs from the data (Eubank, 1988; Härdle, 1990). In statistics, smoothing is used to

create an approximate curve that attempts to capture important patterns in the data and reduce noise. In the smoothing technique, instead of using all the data points, local averaging is used to estimate the relationship between the latent variable and the probability of choosing an option. "[K]ernel is a weighting function, which assigns weights to the scores, based on their distance from the targeted score" (Rajlic, 2020, p. 373). For kernel smoothing technique, there are various functions that can be selected such as Gaussian, uniform, and quadratic. In addition, for each selected point on θ scale, a constant distance size, referred to as bandwidth *(h)* which controls the width of the kernel around the point, is selected. Then, a weighted average for all the data points that are within the bandwidth and the point is computed. Points closer to the evaluation point get higher weights (Santor et al., 1994). As argued by Rajlic (2020), "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).

Compared to the standard kernel regression methods, the independent variable is latent trait value θ, and the dependent variable is the probability of selecting the option *m* for item *i*, with the actual choices (Rajlic, 2020; Ramsay, 2000). These can be numerically summarized by defining an indicator variable γ_{ilv}, that takes the value of 0 when subjects do not choose the option, and 1 if the option is chosen by the subjects. The probability function $P_{il}(\theta_q)$ is estimated by smoothing the relation between the binary 0-1 values and the subject abilities by local averaging, in which for any proficiency or trait level θ, the probability of choice $P_{il}(\theta_q)$ at that level is a weighted average of the values of γ_{ilv} for subject with proficiency or trait levels close to θ (Rajlic, 2020, p. 373):

$$P_{il}(\theta_q) = \sum_{v=1}^{n} \omega_{vq}\gamma_{ilv} \qquad (2)$$

ω_{vq} is a weight assigned to each subject at each evaluation point *q*

$$\omega_{vq} = \frac{K\left(\frac{\theta_v - \theta_q}{h_i}\right)}{\sum_{r=1}^{n} K\left(\frac{\theta_r - \theta_q}{h_i}\right)} \qquad (3)$$

In equation (3), *K* denotes kernel function, and *h* is the bandwidth parameter. For more detailed information about technical description of KS-IRT and regression methods based on kernel smoothing techniques; interested readers can refer to Eubank (1988), Härdle (1990), Mazza et al. (2014), and Ramsay (1991).

There are a range of outputs specific to KS-IRT. A unique graph of KS-IRT is the dynamic display of item characteristics for items with more than 3 or 4 options. The vector of probabilities can be shown as a point in the probability simplex. As θ varies, due to the assumptions of smoothness and unidimensionality of the latent variable, the vector of probabilities moves along a curve (Mazza et al., 2014). A convenient method is to show points in the probability simplex by the *(reference) triangle* – an equilateral triangle having unit altitude - and by the *(regular) tetrahedron*. At each side of the triangle and tetrahedron, only the options with the highest probabilities are shown; the highest probabilities are normalized to provide a simple representation (Mazza et al., 2014). A good item is one in which the sequence of points begins from the lowest option and terminates at or near the highest option to show all the vector of probabilities along the curve.

Using the response patterns of examinees and the item OCCs, the KS-IRT can also give the relative likelihood or probability of an examinee's true proficiency level (θ) being at various values. The curve is known as relative credibility curve. The θ value with the highest likelihood is taken as the maximum likelihood (ML) estimate of the ability for the respondents. Since the ML estimate of the ability takes into account the respondents' response patterns and the characteristics of the items, it is a more accurate indicator of the latent ability than the sum score (Mazza et al., 2014). In fact, ML estimates are the best estimates of the trait level, and RCCs show how precisely a total score shows the ability of a respondent (Ramsay, 2000). If the θ value with the highest credibility and the actual total score coincide, it means that the total score is an accurate indicator of the latent trait. However, if the total score and the θ do not coincide, it is a sign that the total score is inaccurate and does not represent the actual ability of the examinee.

Another output of the KS-IRT is a principal component analysis (PCA) of correct option ICCs. This plot depicts all of the correct-option characteristic curves or all of the expected item scores (EISs) simultaneously so as to indicate relationships among

items (Ramsay, 2000). This is carried out by a PCA of the values of the curves at each point of curve evaluation. "Prior to the analysis, the average curve is calculated across items, and subtracted from each item characteristic curve. In other words, the principal components analysis is carried out on the centered item characteristic curves" (Ramsay, 2000, p. 41). In this analysis, items play the role of examinees or replications, and evaluation points are similar to the role played by variables in common applications of PCA.

Empirical Example

In this section, for the purpose of illustration, the data of the Persian translation of the Children's Test Anxiety Scale (Shoahosseini & Baghaei, 2020; Wren & Benson, 2004) is used to demonstrate how to estimate and interpret the graphical outputs of KS-IRT using KernSmoothIRT package version 6.4 (Mazza et al., 2022) in R (R Core Team, 2023). The dataset comprises 160 participants (90 girls and 70 boys) aged 8 to 14 years old (*Mean of age* = 12.88, *SD* = 1.96) with Persian as their first language. The scale consists of 30 items scored on a four-point Likert-type scale: almost never (0 point), some of the time (1 point), most of the time (2 point), almost always (3 point). For the dataset used in this study, total scores represent the anxiety level of children; higher scores indicate higher anxiety level of children, and lower scores show their lower anxiety level. The graphs of the KS-IRT are analyzed at both test- and item-level as well as plots for assessing Differential Item Functioning (DIF).

Kernel Smoothing Estimation with the ksIRT() Function

To perform the analyses, the package should be first loaded:

```
> library(KernSmoothIRT)
```

The "**foreign**" package can be used to import the data from different statistical packages, such as SPSS or Stata, into R. The data for the current analysis is saved in the format of tab-delimited text (.txt or .dat) or comma separated values (.csv). To import the data into R, load the package:

```
> library(foreign)
```

Specify the folder where the data file has been saved:

```
> setwd("... file location")
```

The data file is specified and imported by executing the following code:

```
> data<-read.table("data.dat",
header=TRUE)
```

The argument header = TRUE tells R that the first row of the data file contains variable names. The argument header = FALSE should be used when the data file does not have specific names for variables.

To perform kernel smoothing, the ksIRT() function requires responses matrix; rows represent subjects and columns represent items. Columns of items in the data file are specified:

```
> data1<-data [,4:33]
```

For all items of the test, key, as a numeric vector or a scalar, should be identified:

```
> key <-
c(3,3,3,3,3,3,3,3,3,3,3,3,3,3,3,
3,3,3,3,3,3,3,3,3,3,3,3,3,3,3)
```

Note that if key is a vector, the length of key should match the number of items; if key is a scalar, the same value should be used for all items. For multiple-choice and dichotomous items, key should include the correct options for each item. One way to score multiple-response items is to count the correctly classified options. To do this, a preliminary conversion of every option into a separate true/false item is necessary (Mazza et al., 2014). For rating scale items, key should include the largest option value for each item. The weight assigned to each option is thus equal to its option number. In our illustrative data, there are 30 items in which option 3 is the largest response option. For nominal items, key is omitted. Bear in mind that subjects have to be ranked to analyze items or options. This can only be carried out when the test also includes non-nominal items, or when a prior ranking of subjects is given with SubRank (Mazza et al., 2014). Finally, for partial credit items, weighting can be specified in the weights argument. More complicated weighting schemes can be specified using weights instead of both key and format.

The basic function that performs KS-IRT is:

```
> Mod1<-ksIRT(data1, key = key,
format = 2, kernel =
c("gaussian"),weights, miss =
c("option"), NAweight = 0,
bandwidth = c("Silverman"),
```

```
RankFun = "sum", thetadist =
list("norm",0,1), groups = FALSE)
```

In the function, there are some elements which should be explained:

- "format" is a vector or a numeric scalar which specifies the type of items. If items are multiple-choice and dichotomous, use format=1. If items are rating scale and partial credit, use format=2. In our example, the data is rating scale. If items are nominal, use format=3. If there are a mixture of different item formats, then format is a numeric vector with length equal to the number of items with entries of 1 for each multiple-choice item and 2 for each rating-scale item. Weights argument is used for more complicated weighting schemes;
- "kernel" specifies the kernel function which must be "gaussian", "quadratic", and "uniform". The default is "gaussian";
- "weights" is an optional list that can be used instead of including key. Determining weights allows for more complicated weighting schemes than the default. Its length must be equal to the number of items, and each entry must be a matrix with option numbers in the first row and option weights in the second row. When weights is removed and format=1, then weights are provided according to key. When weights is removed and format=2, then an option weight equals the option number is provided for each response. If weights is omitted and format=3, then weights are set to zero;
- "miss" specifies the method to handle missing responses. The default value is "option", considering them as a further option (labeled as NA) with zero weight. Such NA option will be added to the plot of the OCCs. As an alternative, a different weight for the NA option may be specified through the NAweight argument. The other approaches to manage missing responses are (a) miss = "random.unif" which substitutes NAs with randomly-chosen options from the possible ones for the items, (b) miss = "random.multinom" which does the same substitution as miss= "random.unif", but each option has a probability of being selected proportional to its relative frequency, and (c) which removes or excludes all the subjects from the analysis with at least one omitted response;
- "NAweight" is a scalar value that determines the weight given to missing responses when miss="option". The default is zero;
- "bandwidth" can be either "Silverman", "CV" (e.g., cross-validation), or a numeric vector specifying, for each item, the bandwidth to use for kernel smoothing. The default value is bandwidth="Silverman", a numeric vector computed following the famous Silverman's rule of thumb (Silverman, 1986, p. 45) with the formula 1.06*sigma.hat*nsubj^(-0.2), where nsubj is the number of subjects and sigma.hat is the standard deviation of the subject summary related to the subjects based on the distribution determined with thetadist. When bandwidth = "CV", the bandwidths is selected for each item through cross-validation (for technical and detailed information for Silverman and CV, see Mazza et al., 2014);
- "RankFun" is a function used to rank subjects. The default value is "sum". The other choice is "mean";
- "thetadist" specifies the distribution of subjects (e.g., ability of θ distribution of subjects). By default a standard normal distribution is used. The other different distributions can be used by specifying the first element of the list as "norm", "beta", "unif", "gamma", etc. where the character string is the same as used in the subjscoresummary function qnorm(), qbeta(), qunif(), qgamma();
- "groups = FALSE" is an optional vector of length equal to the number of subjects containing the group designation of each subject. Including this option allows for comparisons between groups using the DIF tools, which will be explained in the following section.

In addition to these commands used for the analysis of the dataset of the present paper, there are some other arguments. For example, "`itemlabels`" is an optional list of labels for each item. If removed, each item will be labelled based on its numerical order. The labels will be used in plotting; "`nsubj`" is an optional numeric value with the number of subjects; "`SubRank`" is a numeric vector determining the rank of each of the subjects. If undetermined and `format=1` or `format=2`, subjects will be ranked based on the function passed through the argument RankFun. When `format=3`, this argument must be given; "`evalpoints`" is an optional numeric vector that allows users to directly specify evaluation points or the quantiles at which to estimate the OCCs. If undetermined, the default is `nevalpoints` evenly spaced values with end points specified based on the number of subjects and the distribution determined by the `thetadist` argument; and "`nevalpoints`" is an optional scalar value determining the number of evenly spaced points at which curves are estimated. This value is used as an alternative to a user defined vector in the `evalpoints` argument. The default value is 51. The end points are specified based on the number of subjects and to the distribution determined for the thetadist argument. When both `evalpoints` and `nevalpoints` are specified, then `evalpoints` has priority.

If the user tends to perform the kernel smoothing based on default values, the simple function for KS-IRT is:

```
> Mod1 <- ksIRT(data1, key = key, format = 2)
```

The point polyserial correlations (also called point-biserial correlations) can be estimated to show the correlation between each item and the total score (Olsson et al., 1982) as shown in Table 1.

```
> itemcor(Mod1)
```

Table 1. The Point Polyserial Correlations

	Item	Correlation
1	1	0.5489188
2	2	0.6844096
3	3	0.6347367
4	4	0.5879728
5	5	0.6201726
6	6	0.5107299
7	7	0.5257794
8	8	0.4244994
9	9	0.5244451
10	10	0.5288513
11	11	0.6072881
12	12	0.5248562
13	13	0.5897299
14	14	0.5830566
15	15	0.6720744
16	16	0.5113575
17	17	0.5916652
18	18	0.6037813
19	19	0.6268367
20	20	0.4906804
21	21	0.6940162
22	22	0.3330680
23	23	0.6163747
24	24	0.7178708
25	25	0.3907896
26	26	0.4976271
27	27	0.6539635
28	28	0.7596295
29	29	0.6030733
30	30	0.4045336

Plot Methods at Item-level

The function `ksIRT()` creates a `ksIRT` object using kernel smoothing. The plot method for `ksIRT` objects includes a variety of exploratory plots at item- and test-level as well as differential item functioning (DIF), including OCCs, EISs, principal component analysis (PCA), a probability simplex plot for the top 3 or 4 highest probability options of each item, relative credibility curves (RCCs), density plots, expected value plots, standard deviation (SD) or standard error of measurement (SEM), OCCs for each of the different groups (OCCDIF), pairwise expected value comparison plots for each of the different groups (expectedDIF), expected item scores for each of the different groups (EISDIF), and density of observed scores for each of the different groups (densityDIF). The following sections show these plots by means of the `plot()` method.

Option Characteristic Curve (OCC). The following code returns the OCCs for items 2, 7, 8, and 15 of the Children's Test Anxiety scale presented in Figure 1:

```
> plot(Mod1,plottype="OCC",
item=c(2,7,8,15), axistype =
"scores")
```

"`axistype`" specifies the display variable to be used on the *x*-axis. The default is `axistype = "distribution"`, which uses `subjectscoresummary` of the distribution specified in the `thetadist`. The other option is `axistype = "scores"` which shows the expected score. The confidence intervals can be used if `alpha = 0.05`. The default is `alpha = FALSE`.

The OCC graphs show the probability of giving a correct response or endorsing an option (*y*-axis ranging from 0 to 1) for different locations on the latent trait dimension on which individuals are ranked (*x*-axis). On the OCC graphs, the vertical dashed lines indicate the points below which 5%, 25%, 50%, 75%, and 95% of individuals fall with respect to their actual total scores. The position of the vertical lines is identical for all the items. For example, the 75% line is dotted at the score 34 for all the items (2,7,8,15) in Figure 1, indicating that 75% of the respondents fall below the total score of 34 and 25% of the respondents are in the range of scores 34 to 90. This shows that there may exist a relative positive skewness in the data, that is, a large number of the respondents have low total scores, which represent the low test anxiety level.

As illustrated in Figure 1, four curves for each of the scale items are plotted, because there is more than one item response option (e.g., four-point items) in the scale. On the *x*-axis of the OCC graphs, the expected score, ranging from 0 to 90, which represents θ is given. Expected score is the average score that an examinee at a given θ level will achieve. For dichotomous items, it is the sum of the probabilities of a correct response on all the items at a given θ level. For polytomous items, it is the sum of the weighted probabilities of marking all the categories on all the items at a given θ level (Ramsay, 2000). The probability of a correct response or selecting a particular response option at different θ levels is estimated using the kernel smoothing function. According to monotonicity assumption, respondents with higher scores on the latent trait dimension have a higher probability for giving a correct answer to a test item or endorsing an option. In this example, an increase of total scores on the *x*-axis indicates an increase in test anxiety for the respondents. In other words, respondents with higher expected scores on the *x*-axis are more likely to select higher response categories (e.g., Options 2 and 3), and respondents with the lowest level of test anxiety are more likely to select lower response categories (e.g., Options 0 and 1). Therefore, a satisfactory curve for polytomous items is 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 at 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) represent the violation of monotonicity assumption which has a strong effect on the accuracy of measurement (Sijtsma & Molenaar, 2002; Wind, 2020). As can be seen in Figure 1, the four response categories for the items are the most probable category for respondents at certain levels of the anxiety scale. Items 2 and 15 are adequately monotone because

respondents with higher levels of test anxiety (or total scores) have a higher probability to select the higher options; however, the OCC graph for Item 8 indicates a large degree of violation of monotonicity because respondents with lower test anxiety levels have more probability to select higher options. Item 7 also shows a small degree of distortion of monotonicity.

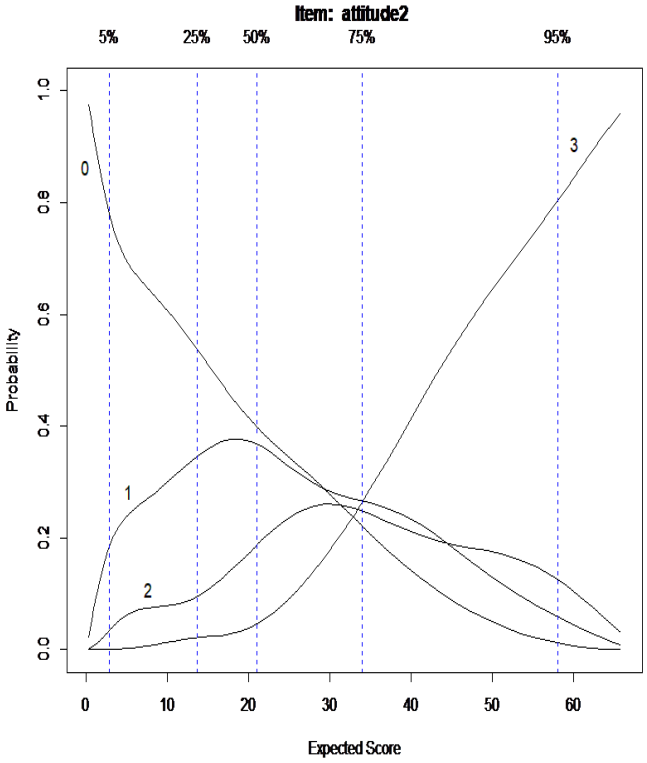
More precisely, each response category should be the most probable option for participants at certain levels of the trait continuum. That is, the first response Category (0) should be the most probable for those with the lowest trait levels (whatever the trait is) and become less probable as the trait increases. The probability of the lowest category should be near 1 at the lowest end of the trait continuum and should approach zero at the highest levels of the trait scale. Category 1 should be the most probable option for those at low to medium levels of the trait and be less probable for those outside this range of the trait. Category 2 should be the most probable for those at medium to high levels of the latent variable and be less probable for those below and above this level. And finally, the highest category (e.g., Category 4) should have a very low probability for those at low and medium levels of the latent variable and be very probable for those with very high trait levels. In summary, ideal OCCs should look like a set of neat successive hills each representing a category and a class of respondents based on the trait levels. Obviously, for multiple-choice (MC) questions we do not expect to see these hills unless each distractor is specifically written to attract respondents from a certain proficiency level. However, we do expect the correct option to have a low probability for low-proficiency examinees and become more probable as θ increases.

For clarification purposes, the performance of Item 15 as shown in Figure 1 is examined. Category 0 is very probable for those with expected anxiety scores between 0 and 10 and becomes very improbable as anxiety increases. For those above 10, Category 1 becomes very probable, and its probability diminishes as anxiety increases. Category 2 does not work well as its probability is lower than the probability of Option 1 across the entire length of the trait continuum. This is indeed a very disturbing finding. Participants with anxiety levels between 10 and 40 are more likely to select Category 1 than 2, and Category 2 never becomes more probable than 1. At anxiety levels above 40, Category 3 becomes the most probable option. This indicates a problem in distinguishing between 4 categories of responses and may call for a reduction in the number of response options. Merging Categories 1 and 2 may solve the problem. This problem also exists for Item 2. In Item 8, Category 0 is the most probable for those between 0 and 60 after which Category 1 becomes more probable. In other words, Categories 2 and 3 do not work, and this item works better with only two response options. Item 7 is also problematic, as category 0 remains the most probable option for a very wide range of the scale. Only after expected score of 40, Category 3 becomes the most probable which makes Options 1 and 2 obsolete.

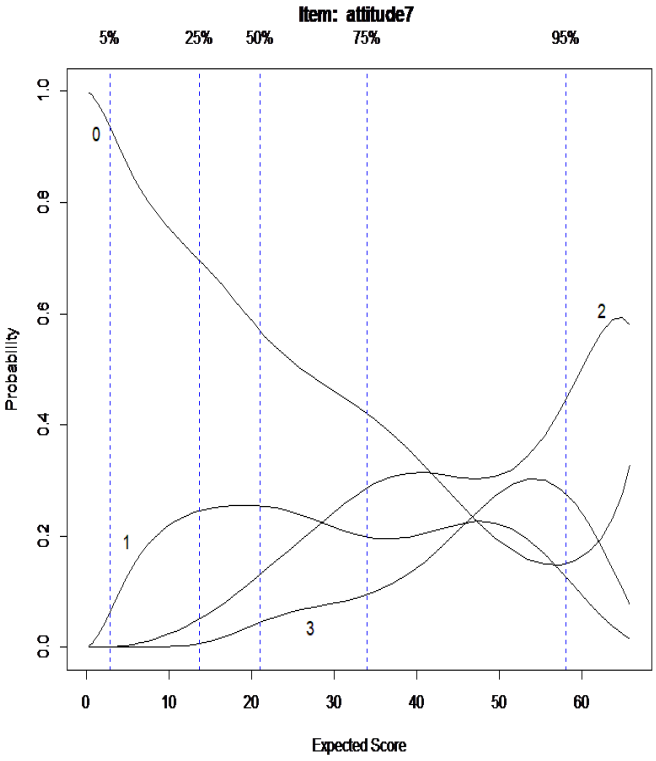
OCC graphs can also represent item discrimination which is defined as the slope or steepness of the OCCs. Item discrimination determines the rate at which the probability of getting an item right or endorsing a response option changes given the latent dimension. As the slope of the curves increases, the better the item can discriminate between the respondents with different trait levels. As presented in Figure 1, Item 2 highly discriminates between the respondents with lower and higher levels of test anxiety, especially with expected total scores ranging from 23 to 45 for Option 3. Items 7 and 15 have also adequate discriminating power with different trait levels. On the contrary, Item 8 displays a weak discrimination item, indicating the inefficiency of the item in differentiating between the respondents with low and high test anxiety level, that is, the respondents with higher level of anxiety have the same probability of endorsing an option with the subjects with lower anxiety levels.

Another important aspect of OCC graphs is the evaluation of the lower and upper asymptote of the curve which refers to the highest and lowest end of a curve. The probability of giving a correct answer to an item or endorsing a response option should approximate 0 in the lowest end of the scale, and should approximate 1 in the highest end of the scale (Rajlic, 2020); otherwise, a set of extraneous, irrelevant, variables may be at work. In Figure 1, the curve of options for all the items approach 0 in the lowest region of the expected score dimension; however, only Items 2 and 15 approach 1 in the highest region of the expected score dimension for the highest option (e.g., Option 3).

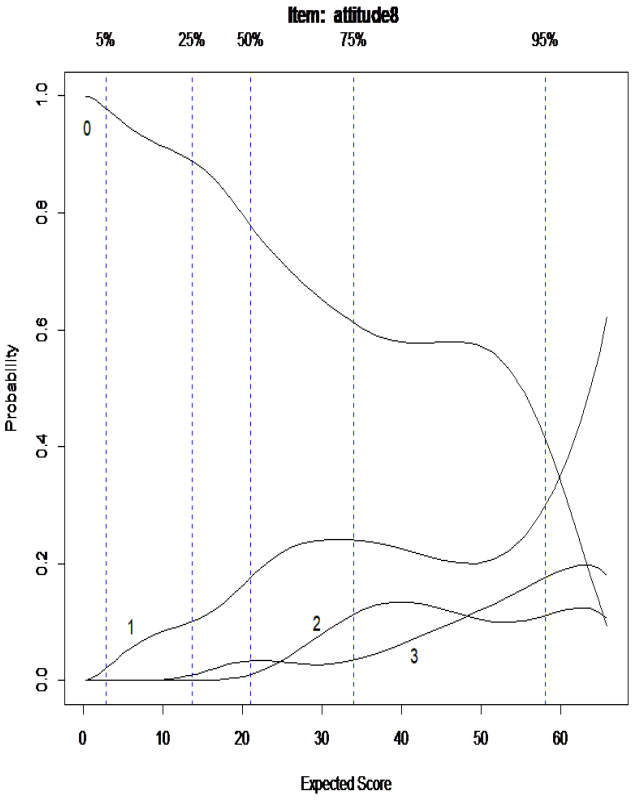
Figure 1. Option Characteristic Curves (OCCs) for Four Items of the Children's Test Anxiety Scale



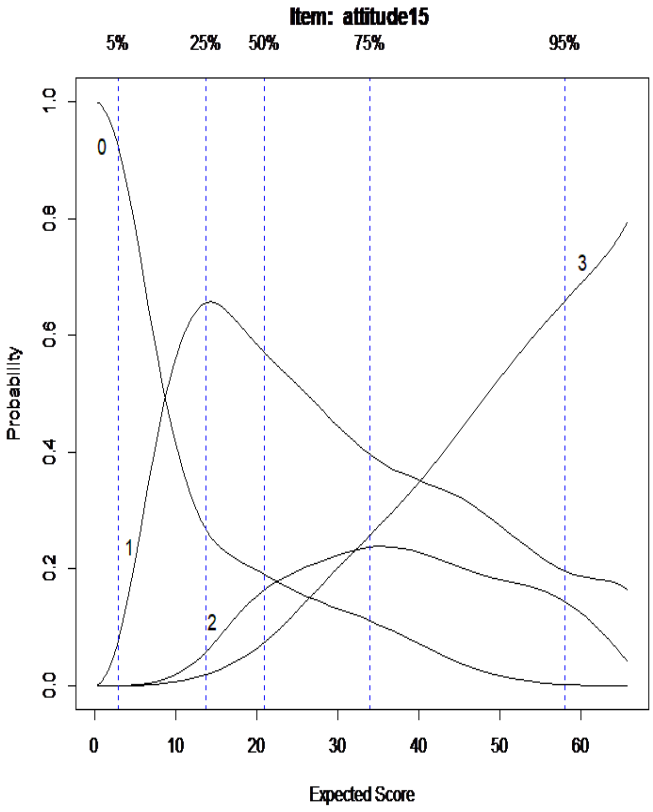
(Item 2)



(Item 7)



(Item 8)



(Item 15)

Expected Item Score (EIS). The code

```
> plot(Mod1,plottype="EIS",
item=c(9,14,23,28), axistype =
"scores")
```

generates the expected item score (IES) graphs for Items 9, 14, 23, and 28 of the Children's Test Anxiety scale displayed in Figure 2. Similar to OCCs, the display variable can be either `axistype = "scores"` or `axistype = "distribution"`, which is the default.

The curves simply show the expected score for the highest option - in this case Option 3 - for different locations of the latent trait. For dichotomous and multiple-choice items, the OCC for the correct option is given. The *x*-axis represents the expected total score on the test, and the *y*-axis represents the expected score on the item. It stands to reason that respondents with higher scores on the overall test also have higher scores on the individual items. Therefore, the IESs are supposed to be monotonically increasing. On the IES graphs, the red dotted lines represents 95% pointwise confidence intervals for only the curve of the highest option (e.g., Option 3), and the points on the graph represent the observed average score for the subjects grouped based on their ordinal ability estimates, which are equally spaced (Mazza et al., 2014). Based on the number of respondents, the intervals show to what extent the curve has been precisely estimated at specific levels of the construct dimension. As can be seen in Figure 2, the narrowest regions are at the low end of the construct dimension, and the widest regions are at the high end of the dimension. When the number of respondents is small (e.g., there is less data) for estimating the curve, the regions get wider which indicates less precision in the estimates. Conversely, the regions get narrower when there is more data for estimating the curve.

Probability Simplex Plots. To produce tetrahedron and triangle simplex plots, run the following codes:

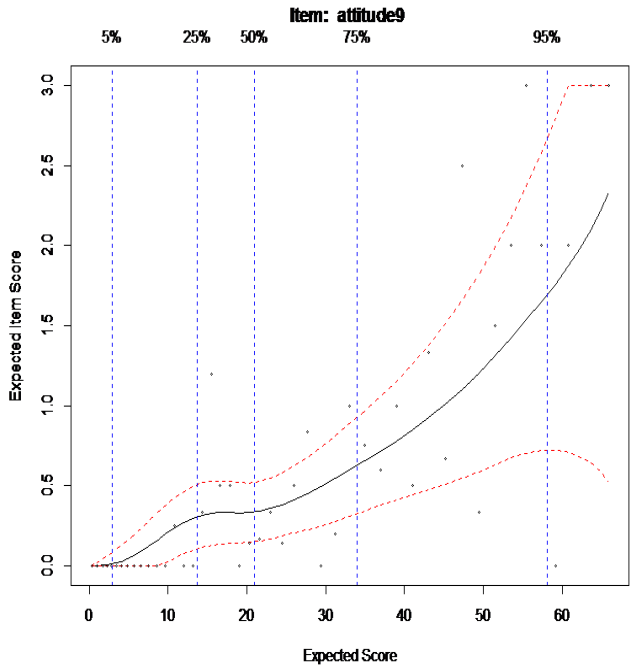
```
> plot(Mod1,
plottype="tetrahedron", items=
c(2,25))
```

```
> plot(Mod1, plottype="triangle",
items= c(23,29))
```

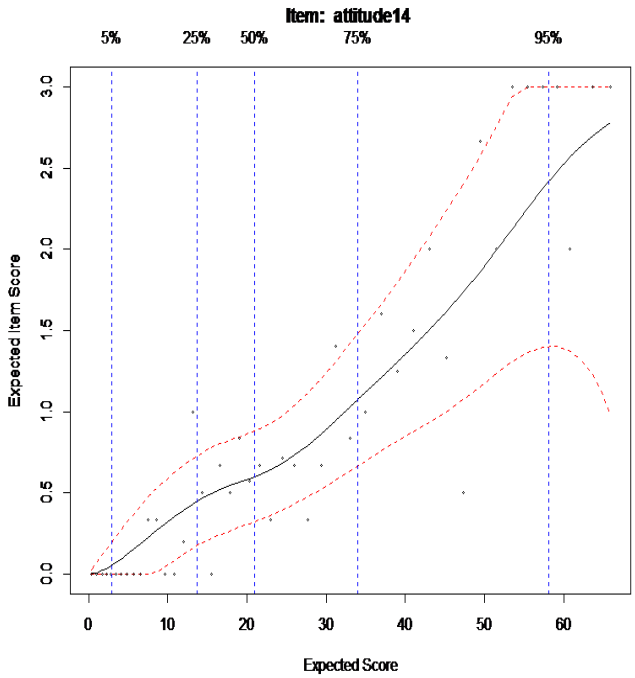
Figures 3 and 4 show (regular) tetrahedron and (reference) triangle simplex plots for four items (e.g., 2, 25, 23, and 29), respectively. These plots are only used for items with more than 3 or 4 options. As given in Figure 3, there is a curve with three colors inside the (regular) tetrahedron. Color points indicate different trait levels which are broken into three equal groups. Low trait levels are recognized by red points, medium trait levels with green, and high trait levels with blue points. Following the ordering of trait levels, a sound item is one in which the sequence of points starts from the lowest option and stops at or near the highest option. In Figure 3, as expected, Item 2 meets this basic requirement because the sequence of points starts from Option 0 (vertex), passes Options 1 and 2, and moves toward Option 3. However, Item 25 is a weak item because the sequence of points do not terminate at or near the highest option. Another issue in the analysis of tetrahedron is the length of the curve. There should be a distance between the respondents with the highest and the lowest trait levels. Unlike Item 2 which is a good item, in Figure 3, because the respondents with the highest levels are far from those with the lowest trait level, Item 25 is a poor item because it has a very short curve which are focused on the lower level options (e.g., Options 0 and 1). Furthermore, the spacing of the points indicate the speed at which the probabilities of response options change. In Figure 3, Item 2 shows a good performance since as test anxiety increases, the probability of endorsing a response option changes; however, in Item 25, the probability of endorsing an option does not change to a great extent with increasing levels of test anxiety, and thus this item needs further examinations.

Figure 4 displays the (reference) triangle simplex plots for two items of the scale. The three sides of the triangle represent the most often chosen options for Items 23 and 29. In particular, for Item 23, the base side of the triangle shows that Option 1 has a much higher probability of being chosen, and as the anxiety level of the respondents changes, the probability of choosing the other options (e.g., Options 0 and 3) changes as well, that is, subjects with higher levels of anxiety have a higher probability to choose higher options and those with lower level of anxiety have a lower probability to select lower options. For Item 29, Option 0 has a much higher probability of being selected. In both triangles for the items, the sequence of points starts from the lowest option and ends at the

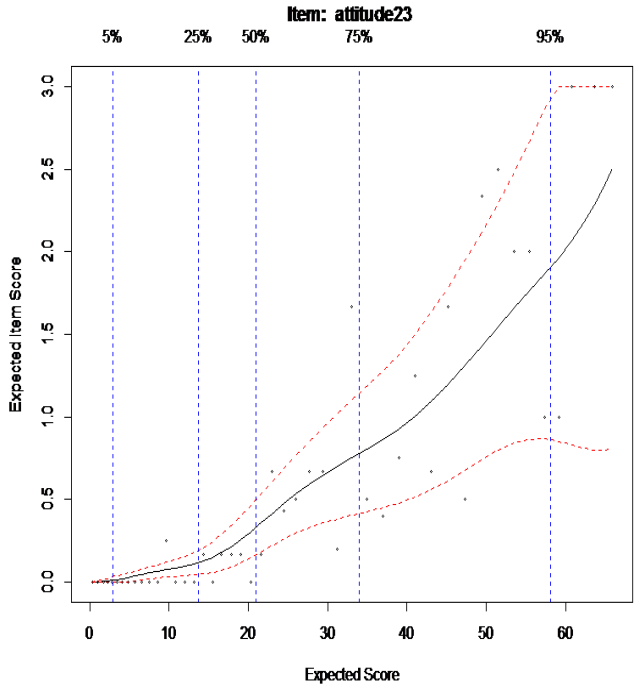
Figure 2. Expected Item Scores (EISs) with 95% Pointwise Confidence Intervals



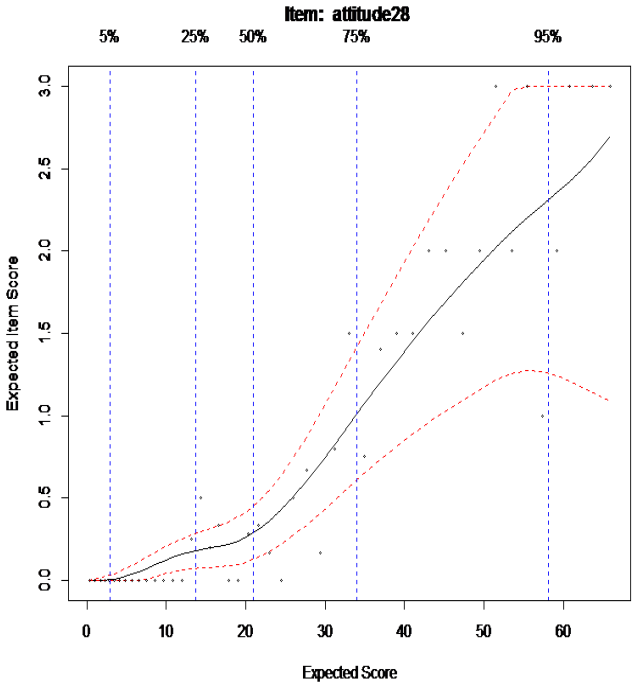
Item (9)



Item (14)



Item (23)



Item (28)

Figure 3. Probability Tetrahedrons for Items 2 and 5 of the Children's Test Anxiety Scale. Low trait levels are plotted in red, medium in green, and high in blue.

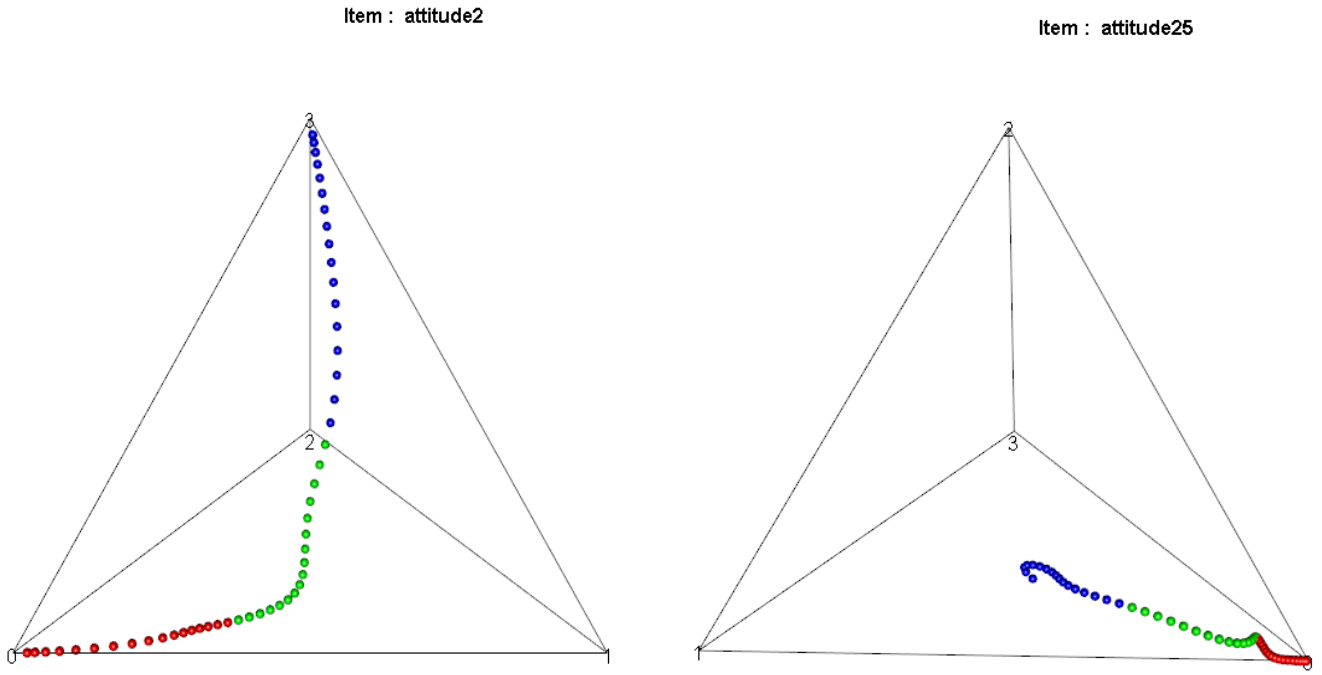
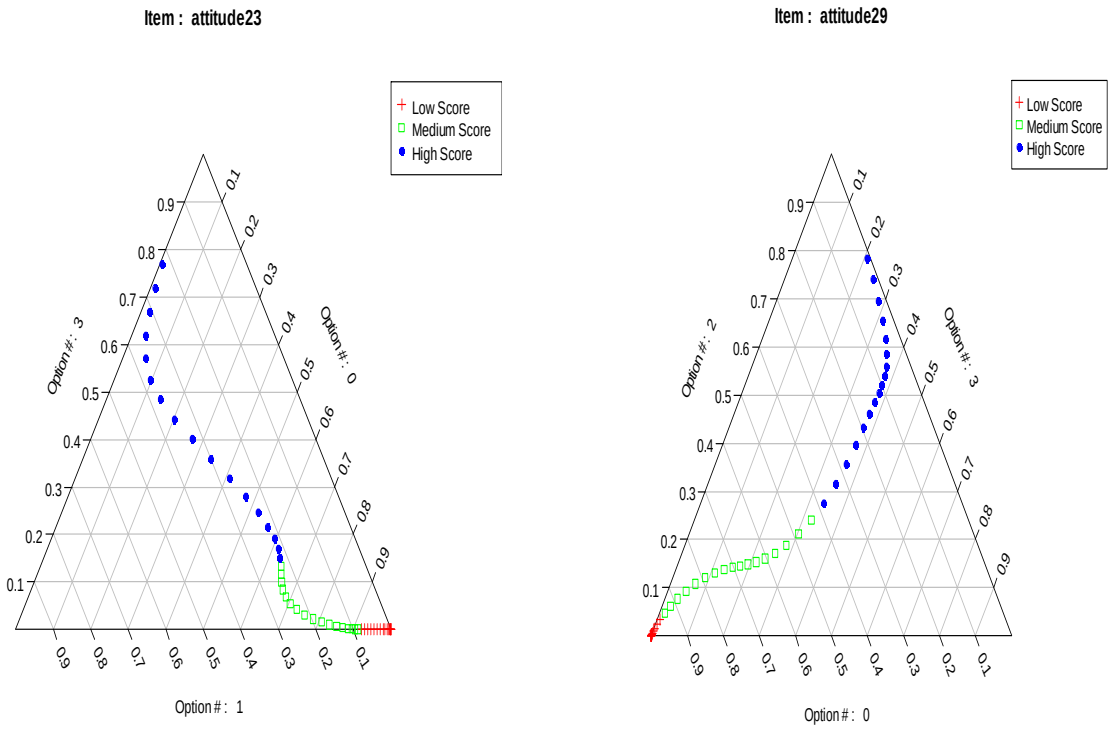


Figure 4 . Probability Triangles for Items 23 and 29 of the Children's Test Anxiety Scale. Low trait levels are plotted in red, medium in green, and high in blue.



highest option, suggesting the reasonable performance of the items.

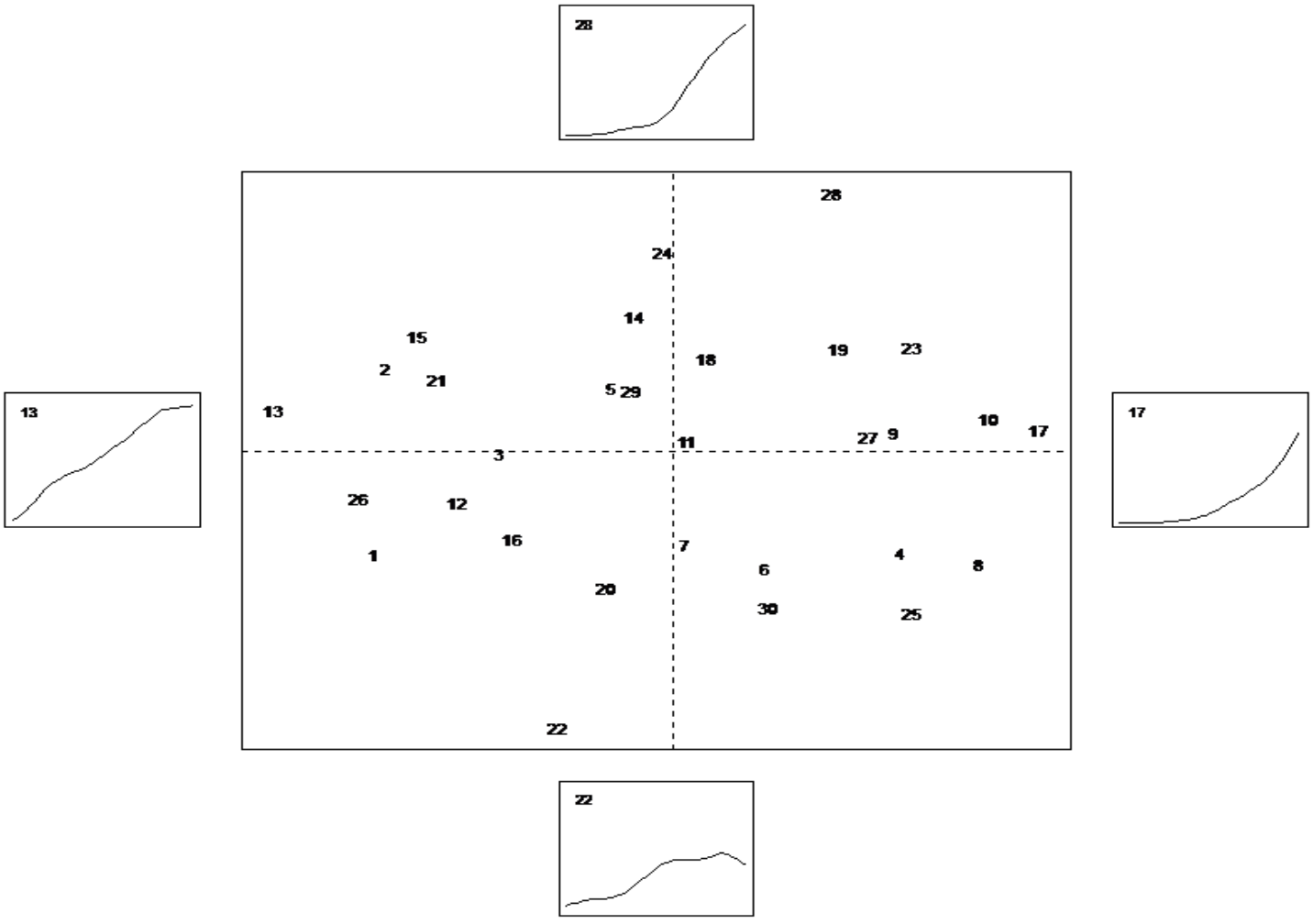
Plot Methods at Test-level

Principal Component Analysis (PCA). The following code can be run to produce principal component analysis (PCA) plot of the test:

```
> plot(Mod1, plottype="PCA")
```

The PCA plot for the Test Anxiety scale items is demonstrated in Figure 5. Items of the scale are represented by numbers inside the plot. PCA plot provides a useful way to compare items all at once and shows the relationship among them. As shown in Figure 5, there are two principal components. On the horizontal axis, the first principal component shows item difficulty; in such a way that the easiest items are placed on the left and the most difficult items on the right. The small plots on the left and right represent the expected item scores (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 in Figure 5, Item 13 is the easiest item, and Item 17 is the most difficult one. On the vertical axis, the second principal component shows item discrimination; in such a way that items high on the plot tend to have a high positive slope, and items low in the plot tend to have a high negative slope. The small plots on the top and the bottom represent EISs for the highest and lowest discriminating items. In this example, Item 28 has the highest discriminating power, and Item 22 has the lowest discriminating power which differentiates negatively.

Figure 5. First Two Principal Components for the Children's Test Anxiety Scale

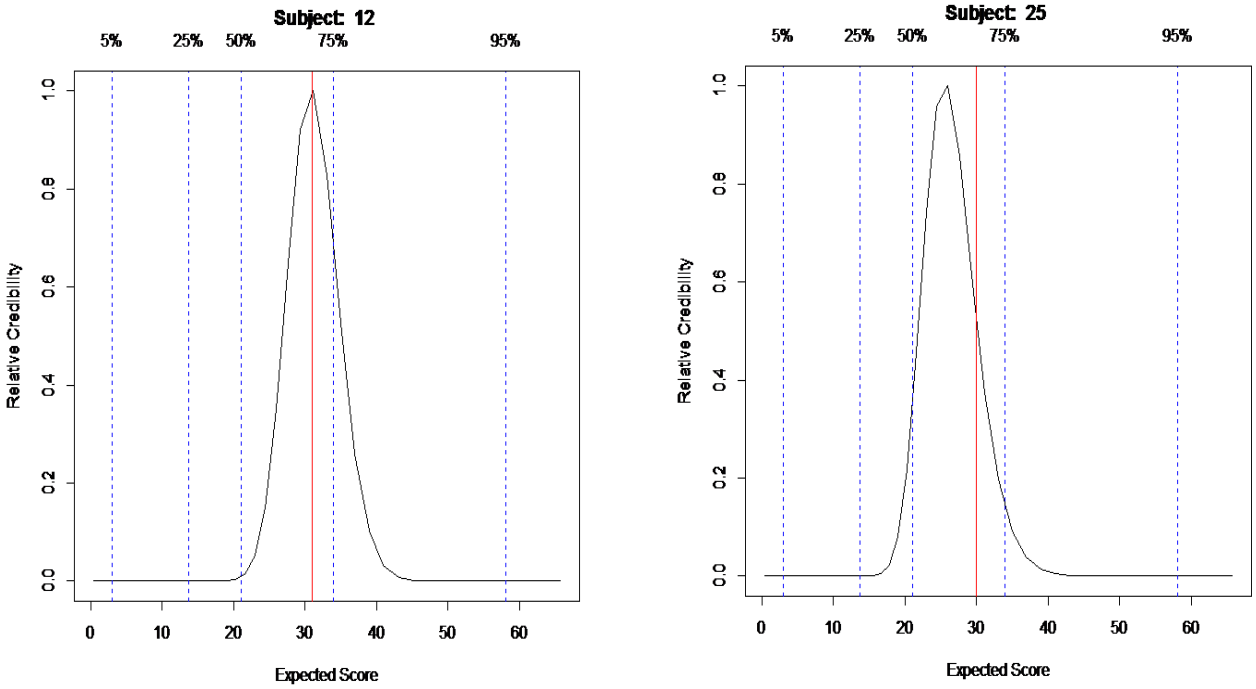


Relative Credibility Curve (RCC). To obtain the RCCs for a number of subjects, run the following code:

```
> plot(Mod1,plottype="RCC",
subjects=c(12,25,27,37))
```

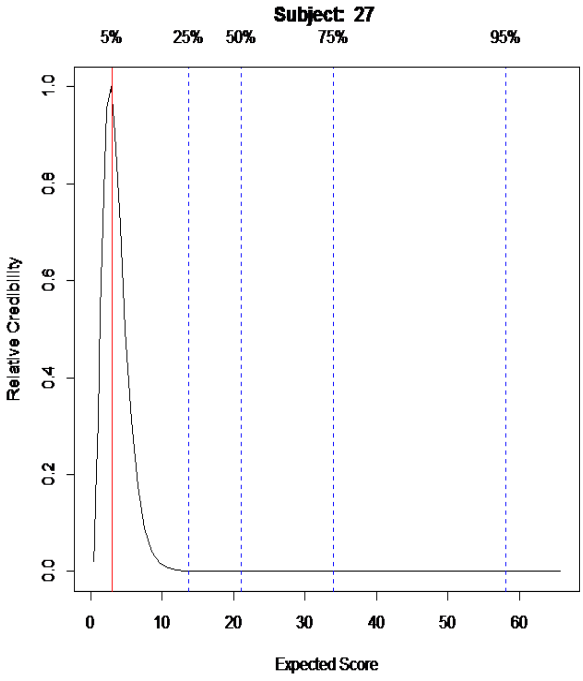
Figure 6 illustrates the RCC plots for four respondents (e.g., 12, 25, 27, and 37) to the Children's Test Anxiety scale items. On the RCC plots, the vertical red line represents the actual total score of the respondent, and

Figure 6. Relative Credibility Curves (RCCs) for Subjects 12, 25, 27, and 37

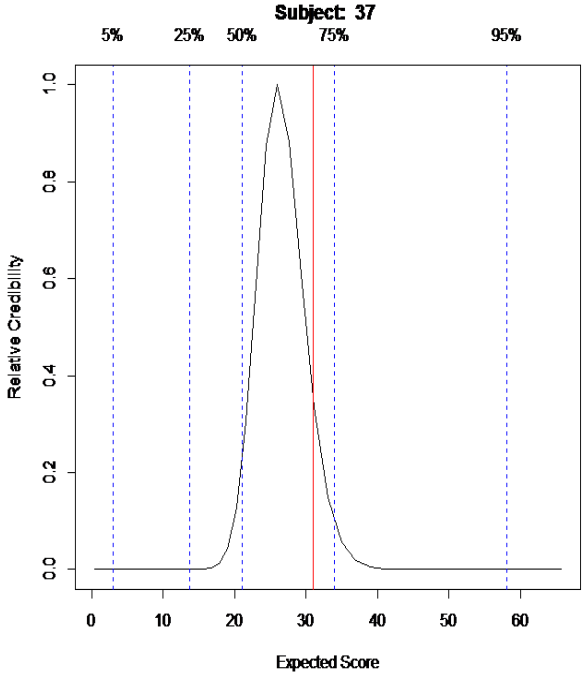


Subject (12)

Subject (25)



Subject (27)



Subject (37)

the blue vertical dashed lines, similar to OCCs, show the points below which 5%, 25%, 50%, 75%, and 95% of individuals fall in terms of their actual total scores. The width of the curve also shows the range where the examinee's true ability 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 θ value (e.g., true trait level). The pointier the curve, the more accurate the θ estimate is. If the total score line (the red vertical line) is to the right of the ML θ, it means that the examinee should have received a lower total score. If the total score line is to the left of the ML θ, it indicates that the examinee should have scored a higher total score. Furthermore, a bimodal RCC indicates that the examinee answered some hard items but failed some easy items (Ramsay, 2000). This is a sign that either some guessing or random answering was involved, or the examinee has a good command of some parts but is less proficient in other parts. Another reason for this phenomenon is multidimensionality.

As indicated in Figure 6, there is a considerable agreement between the total scores and the RCCs for Subjects 12 and 27, although the precision of the ML-estimate is higher for Subject 27 than Subject 12 because the RCC of Subject 27 is more spiky (having sharp point), and the width of the curve is smaller. For Subject 12, the width of the curves indicates that, on the basis of subjects' total scores, his/her true anxiety is most likely between 22 and 40 while, for Subject 27, it is most likely between 0 and 8. For Subjects 25 and 37, however, a difference between the total scores and the maximum of the RCCs is observed. It can be seen that for Subjects 25 and 37, the most likely value, where the curve reaches 1.0 is about 26, indicating that their true latent trait or test anxiety levels are about 6 and 5 points, respectively, lower than their observed total scores. Put it simply, the true anxiety level of Subjects 25 and 37 is lesser than their current anxiety level based on their total scores, suggesting a lower precision.

The following codes respectively produce a vector containing the observed total score of each subject and a vector containing the maximum likelihood estimate of the ability of each subject:

```
> subjscore(Mod1)
  [1] 21  0 22 26 11 61 25 27 60 15 19 31 11  6 42 39 14 20 14 54 39 12 44 28 30
 [26]  7  3  8  0  6 18 58 37 17 36 26 31 17 18 51 20  4 15 46  5 52 51 17 26 29
 [51] 17 15 22 24  8 32 19 23 36  3 23 10 48 26 14  3 23 30 18  6  5  9 18 45 23
 [76]  0 23  4 38  8 19 31 11 15 67 17 28 35  1 15 17  5 59 39 22 32 34  2 20 23
[101] 47  6 13 37 16 39  0 35 31 47 16 55 15 34 13 70  9  5  5 14 32 26 44 22  4
[126] 15  0 64 21 10 18 15 22 37 27 50 18 47 20 15 18 68 16  6 23 45  8 30  3 28
[151] 35 22 34 62 18  1 19 16 30 45
```

```
> subjscoreML(Mod1)
  [1] 20.2818275  0.3459245 22.9947624 27.6654119 10.8035919 63.6433970
  [7] 26.0261668 26.0261668 59.1277277 15.4660874 19.0244064 31.1732905
 [13] 11.9634667  4.8598563 40.9993226 36.9224540 14.3010404 20.2818275
 [19] 13.1326303 53.4949242 40.9993226 13.1326303 43.0847865 27.6654119
 [25] 26.0261668  6.5714423  2.8191766  9.6683087  0.3459245  6.5714423
 [31] 15.4660874 55.4626861 36.9224540 20.2818275 34.9495927 26.0261668
 [37] 26.0261668 16.6330514 15.4660874 49.3921560 17.8138289  4.1130500
 [43] 13.1326303 45.1872540  4.8598563 51.4649163 49.3921560 15.4660874
 [49] 24.4688665 31.1732905 15.4660874 13.1326303 26.0261668 24.4688665
 [55]  8.5748475 33.0309492 17.8138289 22.9947624 36.9224540  2.2681873
```

Test Summary Plots. To obtain an overall assessment of the test, use the following functions:

```
> plot(Mod1,plottype="density",
axistype = "scores")
> plot(Mod1,plottype="expected",
axistype = "scores", lwd = 2)
> plot(Mod1,plottype="sd")
```

Figure 7 displays three test-level summary plots for the Children's Test Anxiety Scale. A kernel density estimate of the distribution of the actual total score is presented in Figure 7a. This figure shows to what extent scores are probable assuming that they are normally distributed (or be bell-shaped). The density plot in Figure 7a shows that the scores in the range of 17 to 20 are most probable for the scale, and the normality assumption is not met in the data. As the most observed scores are clustered around the left tail of the distribution, there is a positively skewed distribution in the data, reflecting that most of the respondents possess low total scores or low test anxiety level.

In Figure 7b, the expected test scores (ETSs) for the scale in relation to (as a function of) the quantiles of the standard normal distribution is illustrated. The curve is expected to be linear or monotonic, indicating if the monotonicity assumption is satisfied at the test level. In this example, the curve is monotonic, that is, the monotonic requirement is met for the scale.

Test standard deviation graph shows the standard error of measurement (SEM) for different levels of θ. SEM is in fact the standard deviation of scores if an examinee takes a test an infinite number of times. In CTT literature, the mean of these repeated tests is called true score, and their standard deviation is the error of measurement (Baghaei & Effatpanah, 2022). As can be seen in Figure 7c, the SD or SEM (on the vertical axis) arrives at the maximum for respondents at around a total score of 58 (on the horizontal axis), where it is about 9. This translates into 95% confidence intervals about 40 and 76 for a respondent who has an expected score of 58 ($58\pm(9 \times 2)$), implying that a respondent with a score value of 58 can be 95% confident that his/her true score is somewhere between 40 and 76. These limits are very wide and hence indicate less precision. The graph suggests that the test is more precise for lower levels of test anxiety.

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 (Zumbo, 2007). In other words, measurement invariance at the item level or DIF is present if respondents with the same level of the trait/ability from different groups have unequal probabilities to give a correct response to an item or endorse an option. A distinction is usually made between two types of DIF that may exist in practice: (a) Uniform DIF is the type of DIF when the probability of endorsing an item is higher for one group than another group across all levels of the trait/ability. In fact, the difference between ICCs for 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; and (b) Non-uniform DIF is the type of DIF when the probability of endorsing an item is different for groups across levels of the trait/ability. In fact, the difference between the ICCs is not constant or uniform across levels of the trait/ability.

In KS-IRT approach, DIF is detected by analyzing curves which produce a visual display of item responses in different groups. Any considerable differences in the shape of the curves across the groups and the size of the areas between them could indicate the presence of DIF in the scale (Rajlic, 2020). To perform DIF analysis using the person variable "Gender", a new object must be created, provided that groups arguments is specified. The following code introduces the column of gender in the data file as the grouping variable:

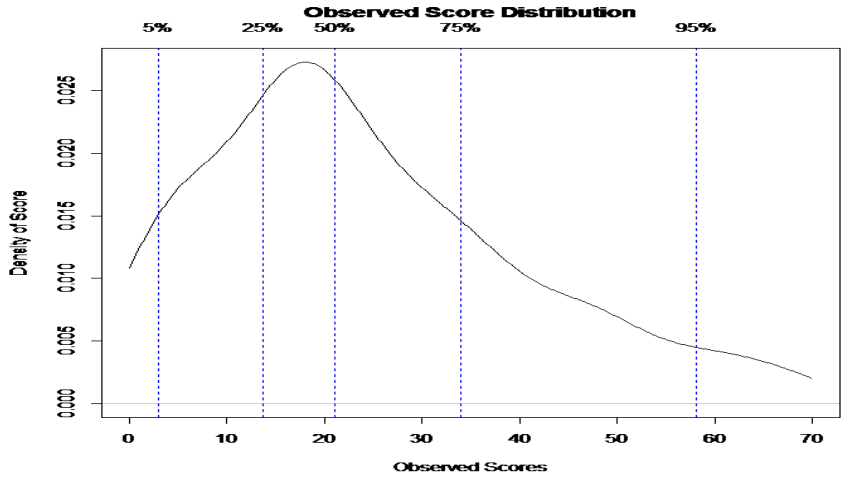
```
> gender <-data [,2]
```

which means that the variable gender is in the second column of the dataset.

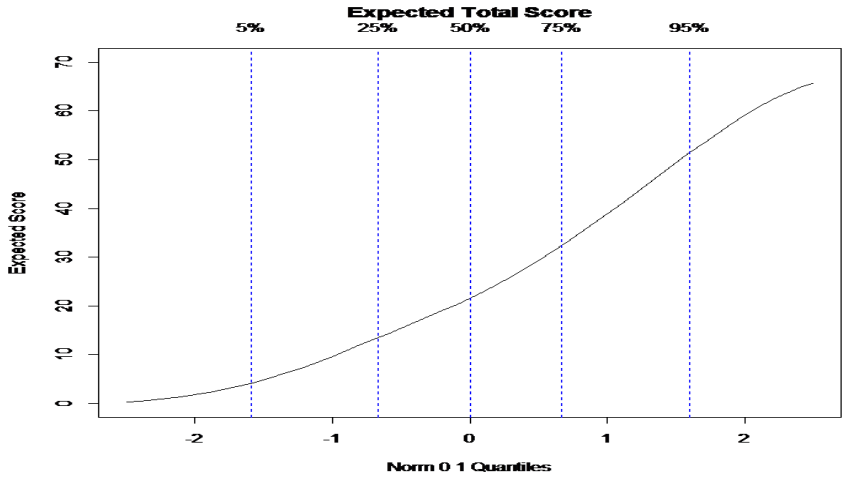
To create a new object with the addition of the groups argument based on "gender" variable, run the following code:

```
> Moddif <- ksIRT(data1, key=key,
format = 2, miss = c("option"),
NAweight = 0, groups=gender)
```

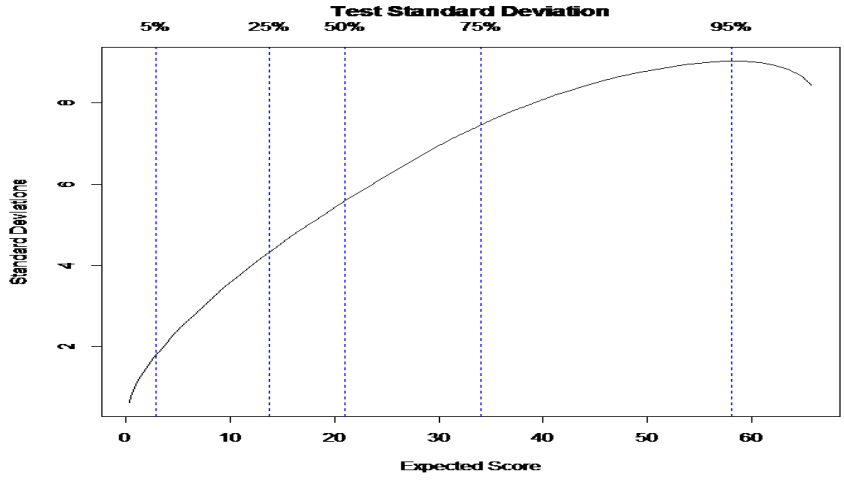
Figure 7. Test Summary Plots for the Children's Test Anxiety Scale



(a) Test Density



(b) Expected Test Score



(c) Standard Deviation

Finally, the following commands can be run to produce different plots for DIF analysis:

```
> plot(Moddif, plottype =
"expectedDIF", lwd = 2)

> plot(Moddif, plottype =
"densityDIF", lwd = 2)

> plot(Moddif, plottype =
"OCCDIF", item = 4)

> plot(Moddif, plottype =
"EISDIF", item = 4)
```

Figure 8a shows the pairwise expected scores or QQ-plot between the distributions of the scores for females (on the *x*-axis) and males (on the *y*-axis). In the QQ-plot, the expected number correct or the total score values for any pair of subgroups corresponding to the various standard normal quantiles are plotted against each other, which summarizes differences in performance between the groups. The horizontal and vertical blue dashed lines indicate the 5%, 25%, 50%, 75%, and 95% quantiles for the two groups. When the two groups have almost the same performance, the relationship will appear as a nearly diagonal line; a truly diagonal line is plotted as a reference (Ramsay, 2000). However, if the groups have different performance, the solid line will deviate from the diagonal line. For the Children's Test Anxiety scale, as can be seen in Figure 8a, there is a subtle difference between the two groups in terms of the distribution of their expected scores. Females have higher scores in the ranges of 2 to 8, and 31 to 52, while males have higher scores in the range of 10 to 21. By reading off the plot, we can find that females with 45 total scores (on the *x*-axis) have higher scores by about 4 or 5 over males with 41 total scores at the same quantile position (on the *y*-axis), indicating that these discrepancies are not considerable.

Figure 8b depicts the total score distribution plot (e.g., kernel density functions) for the two groups. The observed scores for females are shown by the solid blue line and males by the red dashed line. A slight difference in the distribution of the observed scores between the groups is observed for the Children's Test Anxiety Scale. Females have higher scores in the range of 7 to 22, while males have relatively higher scores in the range of 39 to 63. Overall, the two plots, Figures 8a and 8b, suggest an agreement in behavior of the two groups based on their observed scores, showing the lack of a substantial difference in the distribution of the scores between the two groups.

Figure 9 further demonstrates the OCCs for different options of Item 4 of the Children's Test Anxiety Scale across the two groups. On the OCC graphs, the blue curves represent the score distributions for female respondents, the red curves for male respondents, and the black curves, as the overall curve, for all respondents. Lack of DIF is evidenced by overlapping OCCs for the two groups. With regard to Item 4 of the scale, there is a lack of DIF for Options 2 and 3, as the curves almost overlap, but females have a higher probability than males to mark Option 0, while males a have higher probability than females to mark Option 1, indicating a difference in the functioning of these options in the two groups.

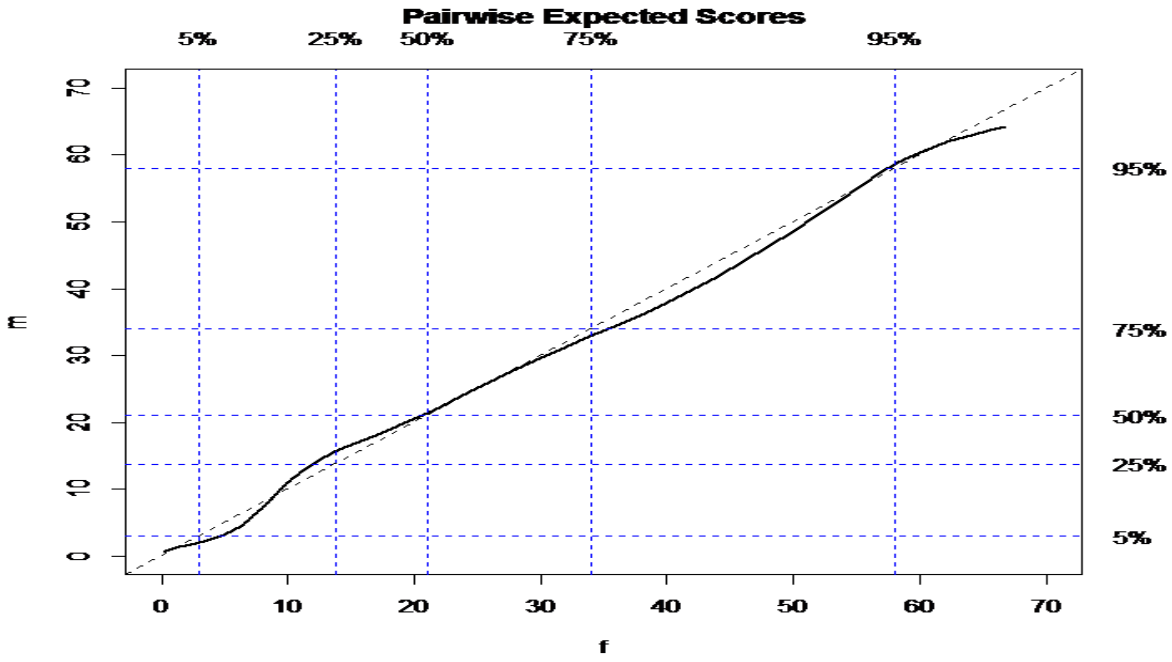
Finally, Figure 10 shows the expected item scores plot for Item 4 of the Children's Test Anxiety Scale across the two groups. On the graph, the blue curve denotes the expected score for female respondents, the red curve for male respondents, and the black curves, the overall curve, for all respondents. The vertical dashed lines display the points below which 5%, 25%, 50%, 75%, and 95% of respondents fall based on their total scores, and that different color points on the plot indicate how respondents from the groups actually scored on the item. As presented in Figure 10, male respondents have relatively greater expected scores compared to female respondents.

The package **KernSmoothIRT** (Mazza et al., 2014, 2022) can also provide more arguments for class 'ksIRT' for subjects. Table 2 gives further codes along with their descriptions for analyzing subjects.

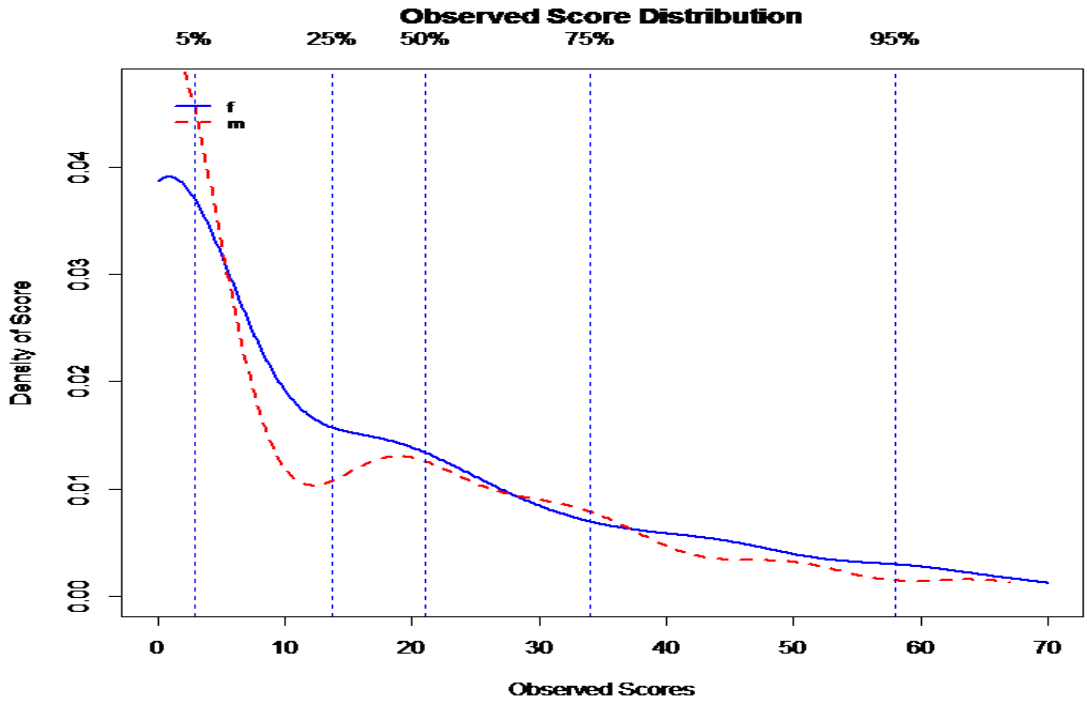
Conclusion

In this article, the basics of kernel smoothing IRT (KS-IRT; Ramsay, 1991) were introduced to applied researchers not acquainted with this approach, and R functions were provided to demonstrate how the **KernSmoothIRT** package (Mazza et al., 2014, 2022) in R can be conveniently used. To empirically illustrate the functions of the approach, the data of 160 respondents to the Persian Translation of the Children's Test Anxiety Scale (Shoahosseini & Baghaei, 2020; Wren & Benson, 2004) were analyzed,

Figure 8. The Pairwise Expected Scores (QQ-Plot) and Kernel Density Functions for Females and Males on the Children's Test Anxiety Scale

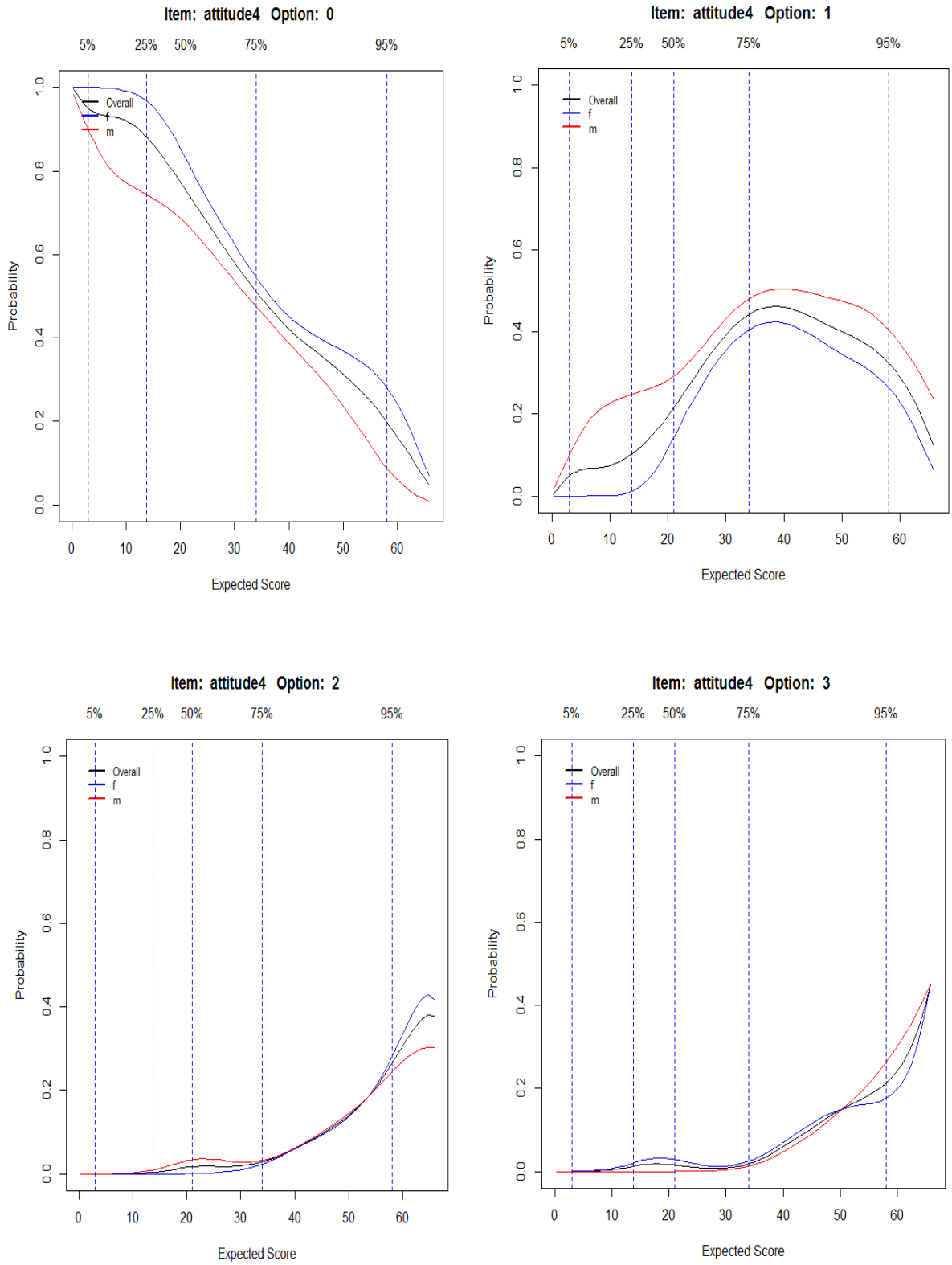


(a)



(b)

Figure 9. Option Characteristic Curves (OCCs) for Females and Males related to Item 4 of the Children's Test Anxiety Scale



and the resultant outputs or plots were interpreted at test- and item-level as well as DIF across genders.

Unlike PIRT models which prescribe a specific shape for the relation between the latent trait and the probability of giving a correct answer or endorsing a test item (e.g., normal ogive or logistic), NIRT models relax this assumption and allow models to estimate ICCs without imposing a specific form. As an exploratory IRT approach, KS-IRT has the potential to offer visual information about the functioning of items in a measure. The graphical representations give initial feedback about the functioning of items at item-level. By analyzing visual displays of items, practitioners can identify poorly functioning items, check model fit, and find the appropriate parametric model for further data analysis (Lee et al., 2009; Ramsay, 2000). The inspection of plots also allows practitioners to check whether the monotonicity assumption is satisfied, whether items have adequate discrimination across the latent dimension, and if all items of the measure function similarly across different subgroups. Therefore, the use of KS-IRT can be a supplemental tool for researchers within the framework of CTT and IRT (Douglas & Cohen, 2001; Junker & Sijtsma, 2001; Sijtsma & Molenaar, 2002; Stout, 2001).

Although the KS-IRT approach proved useful in analyzing the functioning of items, it includes some limitations which should be taken into consideration. The major drawback of the KS-IRT is that it only gives visual illustrations for the evaluation of items and do not provide any numerical values. This makes a challenge for researchers to thoroughly check the psychometric characteristics of a particular measure. More specifically, as no specific boundaries or criteria for evaluating graphs is available, the interpretation and analysis of graphs, especially plots or graphs on DIF, are liable to be arbitrarily or subjectively explained. Consequently, it is highly suggested to use the KS-IRT as a supplementary tool to traditional CTT and PIRT models. Another limitation of the KS-IRT is that it fails to parameterize item difficulties. Wind (2019) also recognized the shortcomings of NIRT models, including MSA, which can be extended to different research contexts in which the KS-IRT is intended to be used. She argued that

Figure 10. Overall Expected Item Score (EIS) and EIS of Females and Males for Item 4 of the Children's Test Anxiety Scale

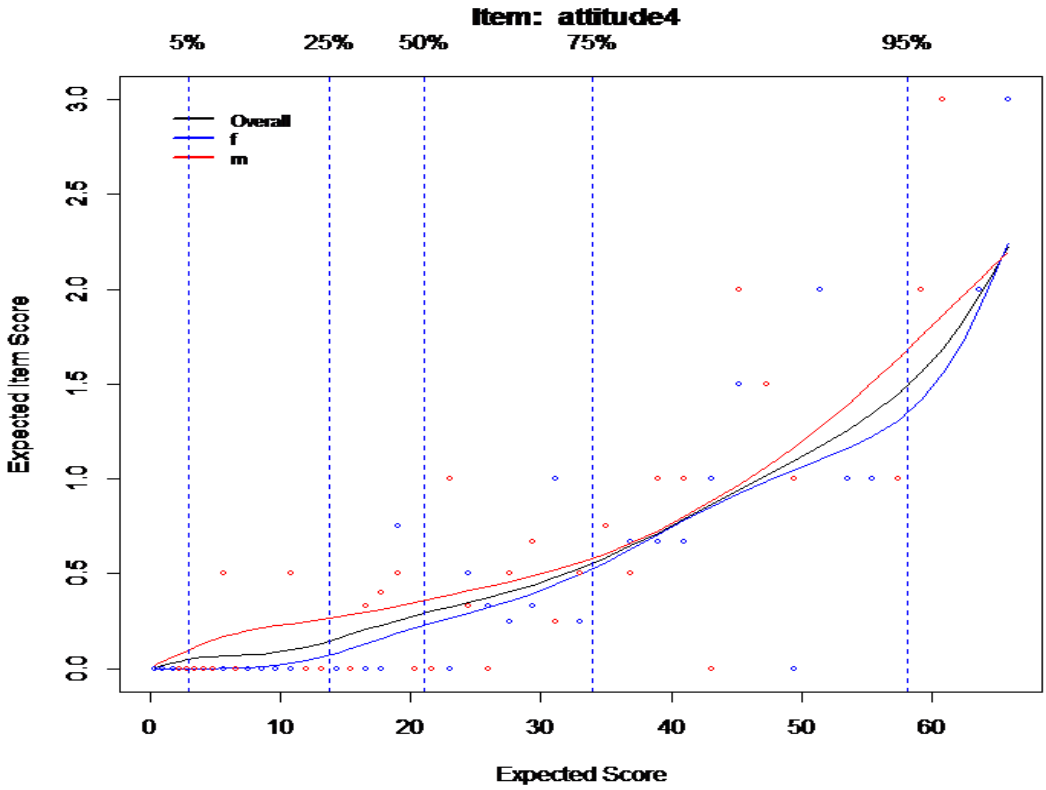


Table 2. Further Arguments for Class 'ksIRT'

Codes	Descriptions
subjthetaML(Mod1)	Returns the maximum likelihood estimate for each subject.
subjETS(Mod1)	Returns a vector with each subjects expected test score.
subjEIS(Mod1)	Returns a matrix containing each subject's expected item score. The rows represent items and the columns, subjects.
subjOCC(Mod1, stype="ObsScore")	Returns a list containing a matrix for each item. Each matrix in the list contains a row for each option with each column representing a subject with the probability of selecting that option for each subject. The scale on which to evaluate each subject. `stype = "ObsScore"` uses the subject's observed test score. `stype = "ExpectedScore"` uses the subject's expected test score. `stype = "MLScore"` uses the maximum likelihood estimate for the subject's overall score. `stype = "Theta"` uses the subject's rank on the thetadist scale. `stype = "MLTheta"` uses the maximum likelihood estimate for the subject on the thetadist scale.
subjEISDIF(Moddif)	It returns a matrix containing each subject's expected item score. The rows represent items and the columns, subjects.
subjETSDIF(Moddif)	Returns a vector with each subjects expected test score.
subjOCCDIF(Moddif)	It returns a list containing a matrix for each item for each of the different groups. Each matrix in the list contains a row for each option with each column representing a subject with the probability of selecting that option for each subject. The scale on which to evaluate each subject. `stype = "ObsScore"` uses the subject's total score. `stype = "Theta"` uses the subject's rank on the thetadist scale. `stype = "ThetaML"` uses the maximum likelihood estimate for the subject on the thetadist scale. `codestype = "ScoreML"` uses the maximum likelihood estimate for the subject on the overall test score scale.

> the lack of a parametric form prevents PIRT 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. (pp.18–19)

Irrespective of these drawbacks, the KS-IRT has the potential to provide preliminary feedback about the performance of test items.

References

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

Baghaei, P., & Effatpanah, F. (2022). *Elements of psychometrics* (2nd Ed.). Sokhan Gostar Publishing.

Baker, F. B., & Kim, S. H. (2004). *Item response theory: Parameter estimation techniques* (2nd edition). Marcel Dekker.

Beevers, C. G., Strong, D. R., Meyer, B., Pilkonis, P. A., & Miller, I. R. (2007). Efficiently assessing negative cognition in depression: An item response theory analysis of the Dysfunctional Attitude Scale. *Psychological Assessment*, *19*(2), 199–

209. https://doi.org/10.1037/1040-3590.19.2.199

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* (pp. 397–479). Addison-Wesley.

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

Effatpanah, F., & Baghaei, P. (2022a, 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.

Effatpanah, F., & Baghaei, P. (2022b). Exploring rater quality in rater-mediated assessment using the non-parametric item characteristic curve estimation. *Psychological Test and Assessment Modeling, 64*(3), 216–252. https://www.psychologie-aktuell.com/fileadmin/Redaktion/Journale/ptam_2022-3/PTAM__3-2022_2_kor.pdf

Effatpanah, F., & Baghaei, P. (submitted). Nonparametric kernel smoothing item response theory analysis of Likert items.

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

Firoozi, F. (2021). Mokken scale analysis of the reading comprehension section of the International English Language Testing System (IELTS). *International Journal of Language Testing*, *11*(2), 91–108. https://www.ijlt.ir/article_138059.html

Fischer, G. H. (1995). Derivations of the Rasch model. In G. H. Fischer & I. W. Molenaar (Eds.), *Rasch models: Foundations, recent developments, and applications* (pp. 15–38). Springer New York.

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

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

Junker, B. W., & Sijtsma, K. (2001). Nonparametric item response theory in action: An overview of the special issue. *Applied Psychological Measurement*, *25*(3), 211–220. https://doi.org/10.1177/01466210122032028

Khan, A., Lewis, C., & Lindenmayer, J. P. (2011). Use of non-parametric item response theory to develop a shortened version of the Positive and Negative Syndrome Scale (PANSS). *BMC Psychiatry*, *11*(178), 1–23. https://doi.org/10.1186/1471-244X-11-178

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

Lord, F. M. (1952). *A theory of test scores*. Psychometric Society.

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

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. https://doi.org/10.18637/jss.v058.i06

Mazza, A., Punzo, A., & McGuire, B. (2022). *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

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd Ed., pp. 13-104). American Council on Education and Macmillan.

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

Molenaar, I. W. (1997). Nonparametric models for polytomous responses. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 367–380). Springer.

Olsson, U., Drasgow, F., & Dorans, N. J. (1982). The polyserial correlation coefficient. *Psychometrika, 47*(3), 337–347. https://doi.org/10.1007/BF02294164

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

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*, 611–630. https://doi.org/10.1007/BF02294494

Ramsay, J. O. (1997). A functional approach to modeling test data. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 381–394). Springer-Verlag, New York.

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

Samejima, F. (1969). *Estimation of latent ability using a response pattern of graded scores (psychometric monograph no. 17)*. Psychometric Society. Retrieved from http://www.psychometrika.org/journal/online/MN17.pdf

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

Shoahosseini, R., & Baghaei, P. (2019). Validation of the Persian translation of the Children's Test Anxiety Scale: A multidimensional Rasch model analysis. *European Journal of Investigation in Health, Psychology, and Education*, *10*(1), 59–69. https://doi.org/10.3390/ejihpe10010006

Sijtsma, K., & Meijer, R. R. (2007). Nonparametric item response theory and related topics. In C. R. Rao & S. Sinharay (Eds.), *Handbook of statistics, vol. 26: Psychometrics* (pp. 719–746). Elsevier, North Holland.

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

Sijtsma, K., Emons, W. H., Bouwmeester, S., Nyklícek, 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: An International Journal of Quality of Life Aspects of Treatment, Care and Rehabilitation*, *17*(2), 275–290. https://doi.org/10.1007/s11136-007-9281-6

Silverman, B. W. (1986). *Density estimation for statistics and data analysis, volume 26 of monographs on Statistics & Applied Probability*. Chapman & Hall, London.

Stout, W. (2001). Nonparametric item response theory: A maturing and applicable measurement modeling approach. *Applied Psychological Measurement*, *25*(3), 300-306. https://doi.org/10.1177/01466210122032109

Sueiro, M. J., & Abad, F. J. (2011). Assessing goodness of fit in item response theory with nonparametric models: A comparison of posterior probabilities and kernel-smoothing approaches. *Educational and Psychological Measurement*, *71*(5), 834–848. https://doi.org/10.1177/0013164410393238

Tabatabaee-Yazdi, M., Motallebzadeh, K., & Baghaei, P. (2021). A Mokken scale analysis of an English reading comprehension test. *International Journal of Language Testing*, *11*(1), 132–143. https://www.ijlt.ir/article_130373.html

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

Wind, S. A. (2019). 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

Wren, D. G., & Benson, J. (2004). Measuring test anxiety in children: Scale development and internal construct validation. *Anxiety, Stress, & Coping: An International Journal, 17*(3), 227–240. https://doi.org/10.1080/10615800412331292606

Zumbo, B. D. (2007). Three generations of DIF analyses: Considering where it has been, where it is now, and where it is going. *Language Assessment Quarterly, 4*(2), 223–233. https://doi.org/10.1080/15434300701375832

Citation:
Effatpanah, F., & Baghaei, P. (2023). Kernel smoothing item response theory in R: A didactic. *Practical Assessment, Research, & Evaluation*, *28*(7). Available online: https://scholarworks.umass.edu/pare/vol28/iss1/7/

Corresponding Author:
Farshad Effatpanah
Islamic Azad University, Mashhad Branch, Mashhad, Iran
Ostad Yusofi St., Mashhad 918714757, Iran.
Email: farshadefp [at] gmail.com

Author Note
Farshad Effatpanah: farshadefp@gmail.com
ORCID ID: https://orcid.org/0000-0003-3970-5588
Purya Baghaei: puryabaghaei@gmail.com
ORCID ID: https://orcid.org/0000-0002-5765-0413

Acknowledgments

Funding
The author(s) received no specific funding for this work from any funding agencies.

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

Competing Interests
The authors declare that they have no competing interests.

Research Data Policy and Data Availability Statements
The datasets generated during and/or analyzed during the current study are available from the corresponding author on reasonable request.