Language Testing in Asia

RESEARCH Open Access

# Application of nonparametric item response theory in determining the one-dimensionality and adaptability of TOEFL iBT listening test

Hamed Ghaemi*

*Correspondence: ghaemiacademy@gmail.com

Bahar Institute of Higher Education, Mashhad, Iran

**Abstract**

Listening comprehension in English, as one of the most fundamental skills, has an essential role in the process of learning English. Non-parametric item Response Theory (NIRT) is a probabilistic-nonparametric approach to item response theory (IRT) which determines the one-dimensionality and adaptability of test. NIRT techniques are a useful tool for researchers who wish to construct one-dimensional tests. The current study utilized the NIRT to examine adaptability of TEOFL iBT listening test, administered on 400 EFL university students in the Iranian context. The results illustrated no main concerns in terms of item adaptability. It was concluded that the ordering of items according to their mean is invariant across examinees. Dimensionality analysis results depicted that the test is one dimensional -confirming evidence of the validity of the test in measuring a single ability dimension. The empirical application depicted a potential and feasible approach, whereby NIRT could be used as a valuable method for exploring the behavior of scaled items in response to varying levels of a latent trait in education research.

**Keywords:** Double monotonicity model, Nonparametric item response theory, Monoton homogeneity model, Listening comprehension, TOEFL iBT

## Introduction

### Listening comprehension

Tests as the key instructive exploration instruments are usually used for the students' evaluations of "courses, projects, and clerkships just as for understudy self-appraisals" (Mokken & Lewis, 1982, p. 2). A reliable test is viewed as a reasonable instrument for the estimation of examinees' inactive attributes or capacities. Estimation and appraisal of assorted factors are unavoidable in instruction and brain research and other sociologies. It is for the explanation that estimation licenses researchers to describe people's disparities as far as the wonder being referred to (Wind, 2017).

Listening includes the impression of sounds made by the speaker, of pitch designs that show the focal point of the data, and of the pertinence of the current subject examined. As per Roland Barthes, listening can be perceived on three levels: alarming, unraveling,

and a comprehension of what the sound is created and how the sound means for the audience. Individuals tune in for 45% of their time imparting.

Perhaps, the main test utilized in instructive settings all through the world is trial of listening and understanding capacity. An individual who gets and gets data or a guidance, and afterward decides not to follow it or not to consent to it, has paid attention to the speaker, despite the fact that the outcome is not what the speaker needed. Listening is a term wherein the audience pays attention to the person who delivered the sound to be tuned in. As Semiotician, Roland Barthes portrayed the qualification among tuning in and listening. "Listening is a physiological wonder; listening is a mental demonstration." We are continually listening, more often than not subliminally. Listening is finished by decision. It is the interpretative activity taken by somebody to comprehend and possibly make significance of something they hear (Oppenheim, 2020, p. 87).

Alarming, being the main level, is the identification of natural sound signals. This implies that specific spots have certain sounds related with them. This is best clarified utilizing the case of somebody's home. Their home has certain sounds related with it that make it natural and agreeable. An interruption, a sound that is not comfortable (e.g., a squeaking entryway or section of flooring, a breaking window), makes the occupant of the home aware of the likely risk. Unraveling, the subsequent level depicts identifying designs when deciphering sounds (Sijtsma & Molenaar, 2019).

Understanding, the third degree of tuning in implies knowing how and what one says will influence another. This kind of listening is significant in therapy, the investigation of the oblivious psyche. Barthes states that the psychoanalyst should wind down their judgment while paying attention to their patient to speak with their patient's oblivious in a fair-minded design (Oppenheim, 2020). This is the same way that audience members should wind down their judgment when paying attention to other.

The entirety of the three degrees of listening capacity is inside a similar plane and at times at the same time. In particular, the second and third levels, which cross over boundlessly, can be entwined in that getting comprehension and inferring importance are essential for a similar cycle. In that, the kid, after hearing the door handle turn getting, can therefore expect that somebody is at the entry (determining meaning).

#### Nonparametric item response theory

All IRT models should have three essential standards in common. Adaptability means that the test should only measure one trait, local independence means that responses to one item should not be influenced by responses to other items, and monotonicity means that the latent trait and the likelihood of a precise response should have a constant connection (Wind, 2017). Mokken's (1971) DMM requires another hypothesis called invariant item ordering (IIO) or nonintersecting item response functions (IRF), which states that individual item response functions should not interfere with any other item response functions, so as long as this assumption holds, the items are ordered the same for all examinees (Wind, 2017).

When it comes to creating hierarchical scales, the IIO attribute is crucial. "If these four assumptions are not disproportionately broken," according to Van der Ark, 2021, "greater sum scores are realized as corresponding to higher values on the latent trait, signaling respondents can be reliably ordered on the latent trait by their sum scores." Mokken

(1971) proposed the MHM and DMM formulations of polytomous data to operate with rating scales data. Despite the fact that these models follow the same essential principles as dichotomous formulations, the polytomous model assumptions are assessed for each item at the overall level as well as within rating scale categories (Wind, 2017). As a result, when student progress levels within a rating scale item increase, the collective likelihood of a rating in or above each rating scale group should similarly increase. For polytomous models, local independence indicates that learners' responses to the items should be statistically autonomous after taking into consideration their ability.

In educational and psychological testing, raw total scores are used to rank examinees from the most gifted to the least gifted or from the most anxious to the least worried. According to the fundamental assumptions of item response theory (IRT), raw scores are treated as ordinal scale data, allowing only the testees' order to be determined rather than their differences in strengths and weaknesses. However, in order for raw scores to be considered ordinal, response patterns must adhere to an axiom known as transitivity (Sijtsma & Molenaar, 2019), which states that if an examinee answers a more difficult question correctly, then an easier item should have been correctly answered as well. Nonparametric item response theory (Mokken, 1971) is a nonparametric item response theory (NIRT) model that measures dichotomous or polytomous unidimensional scales.

Nonparametric IRT models, in contrast to parametric IRT models, place a greater emphasis on deep "model fit examination and data inquiry to understand the test, the items, and the respondents" (Meijer & Baneke, 2004, p. 54). NIRT offers a variety of statistical methods, such as a psychometric data reduction method, for developing scales to evaluate people and items related to personality and cognitive attributes in order to establish the association between items and latent traits. These methods determine whether the data satisfy the ordinal scale's axiom. If the NIRT models suit the data, the examinees should be able to be well ordered based on their overall scores (Junker & Sijtsma, 2001, p. 66). As a result, a greater item score indicates a higher level of characteristic. Many researchers (e.g., Birnbaum, 1968; Wind, 2017) have stressed that NIRT is an effective tool in situations when primary response procedures fail, such as emotional factors. NIRT arose from the Guttman scaling model, which states that items on a scale are hierarchically ordered, implying that test items are well-organized by degree of difficulty (Sijtsma & van der Ark, 2017). As a result, the Guttman scaling model is unavoidable because it does not allow for any arbitrary errors (Weesie, 1999). NIRT can be used in both a confirmatory and exploratory strategy, and it uses the same criteria as the monotone homogeneity model (MHM) and the double monotonicity model (DMM) models (Guttman, 2020).

## Method

### TOEFL iBT® listening section

The TOEFL iBT® listening section is designed to measure your ability to understand conversations and lectures in English. It includes listening for the following:

- Basic comprehension
- Pragmatic understanding (speaker's attitude and degree of certainty) and connecting and synthesizing information

There are 2 types of listening items in the Speaking section — lectures and conversations. Both use campus-based language.

- Three to 4 lectures, each 3–5 min long, with 6 questions per lecture
- Two to 3 conversations with 2 speakers, each 3 min long, with 5 questions per conversation

You can take notes on any audio item throughout the test to help you answer questions. You have 41 to 57 min to complete the section.

Consequently, this study aimed to provide a pragmatic approach to explore TOEFL iBT listening comprehension test from IRT perspectives, using nonparametric item response theory (NIRT; Mokken, 1971).

### Participants and setting

Four-hundred BA students (75% females, 25% male) majoring in English as a foreign language (EFL) and translation studies from different universities in Iran participated in the study. The TOEFL iBT listening test adapted from TPO software was administered in the academic year of 2019. All participants were native speakers of Farsi with the age range of 20–27.

### Instrumentation

Participants' listening comprehension was measured using a TOEFL iBT listening test adapted from TPO software consisting of 28 four-option multiple-choice items. The test consisted of three lectures each one having six questions and two conversations each one including five questions. The Cronbach's alpha reliability of the TOEFL iBT listening test was 0.84. The mean and standard deviation of the sample on the test were 19 and 17.79, respectively.

## Results

The Mokken adaptability coefficients comprising item-pair adaptability (Hij), item adaptability (Hj), and total adaptability (H) were calculated. According to Mokken (1971), a scale is considered weak as long as $0.40 \leq H < 0.50$, an average scale if $0.50 \leq H < 0.60$, and robust if $H \geq 0.60$; the values of Hij must be larger than zero or nonnegative representing a nonnegative relation between an item and the latent trait.

As a consequence, all Hij coefficients must be positively correlated, and items must be appropriately uniform with other items (Jong & Molenaar, 1987). These assumptions can lead to the development of instruments that coincide to more accurate values of reliability and homogeneousness than tools related to conventional classical test theory (CTT) reliability analysis (Van Schuur, 2011). However, if it is not possible to revise the test, systematic removal or replacement of the misfitting items should be conducted.

According to Table 1, the results displayed that Hij and Hj were positive for all the items. Moreover, $Hj > 0.40$ was attained for all the items of the TOEFL iBT listening comprehension test (Table 1). Besides, H was 0.621, which showed the scale is robust. These outcomes propose that each of the items contributes to a meaningful overall ordering of participants in terms of TOEFL iBT listening comprehension ability. There- fore, the

**Table 1** Item adaptability

| Items | Hj[a] | Monotonicity | | | |
|---|---|---|---|---|---|
| | | Active comparison (ac) | Violations (vi) | Significant violations (zsig) | Crit. |
| 1 | 0.78 | 7 | 0 | 0 | 0 |
| 2 | 0.90 | 7 | 0 | 0 | 0 |
| 3 | 0.62 | 7 | 0 | 0 | 0 |
| 4 | 0.81 | 4 | 0 | 0 | 0 |
| 5 | 0.89 | 7 | 0 | 0 | 0 |
| 6 | 0.71 | 4 | 0 | 0 | 0 |
| 7 | 0.66 | 4 | 0 | 0 | 0 |
| 8 | 0.69 | 7 | 0 | 0 | 0 |
| 9 | 0.70 | 4 | 0 | 0 | 0 |
| 10 | 0.83 | 4 | 0 | 0 | 0 |
| 11 | 0.44 | 7 | 0 | 0 | 0 |
| 12 | 0.92 | 4 | 0 | 0 | 0 |
| 13 | 0.85 | 7 | 0 | 0 | 0 |
| 14 | 0.83 | 2 | 0 | 0 | 0 |
| 15 | 0.97 | 7 | 0 | 0 | 0 |
| 16 | 0.65 | 5 | 0 | 0 | 0 |
| 17 | 0.94 | 7 | 0 | 0 | 0 |
| 18 | 0.80 | 5 | 0 | 0 | 0 |
| 19 | 0.92 | 5 | 0 | 0 | 0 |
| 20 | 0.78 | 5 | 0 | 0 | 0 |
| 21 | 0.65 | 5 | 0 | 0 | 0 |
| 22 | 0.55 | 7 | 0 | 0 | 0 |
| 23 | 0.88 | 7 | 0 | 0 | 0 |
| 24 | 0.57 | 5 | 0 | 0 | 0 |
| 25 | 0.59 | 5 | 0 | 0 | 0 |
| 26 | 0.87 | 5 | 0 | 0 | 0 |
| 27 | 0.89 | 7 | 0 | 0 | 0 |
| 28 | 0.77 | 4 | 0 | 0 | 0 |
| 29 | 0.74 | 5 | 0 | 0 | 0 |
| 30 | 0.78 | 7 | 0 | 0 | 0 |
| 31 | 0.98 | 4 | 0 | 0 | 0 |
| 32 | 0.56 | 5 | 0 | 0 | 0 |
| 33 | 0.54 | 7 | 0 | 0 | 0 |
| 34 | 0.29 | 4 | 0 | 0 | 0 |
| 35 | 0.58 | 4 | 0 | 0 | 0 |
| 36 | 0.99 | 4 | 0 | 0 | 0 |
| 37 | 0.88 | 5 | 0 | 0 | 0 |
| 38 | 0.63 | 5 | 0 | 0 | 0 |
| 39 | 0.85 | 7 | 0 | 0 | 0 |
| 40 | 0.69 | 5 | 0 | 0 | 0 |

[a] Scale H = 0.621

scale can be safely treated as unidimensional and can order the respondents using the total scale on the listening ability scale.

Table 1 shows that there is no violation of monotonicity for any of the items as illustrated by the vi column, meaning the collation of participants with their total scores

is defensible. Crit value is used to measure the violation of monotonicity if there is any. According to Molenaar and Sijtsma (2020), crit statistic < 50 is considered as not extremely violating monotonicity; therefore, the items can be safely united in any Mokken scale. $H^T < 0.4$ shows that the item ordering is erroneous, $0.40 \leq H^T < 0.50$ means low accuracy, $0.50 \leq H^T < 0.60$ indicates medium accuracy, and $H^T \geq 0.60$ suggests high accuracy (Van der Ark, 2001). The Crit value (Sijtsma & Molenaar, 2019; Van Schuur, 2011) is used as a measure of the effect size of the violation of IIO.

The $H^T$ value in this study is shown to be 0.21, indicating the item ordering is not much true. The values of the (zsig) column in Table 2 illustrate five items (items 3, 8, 19, and 21) violated the IIO assumption, even though according to the crit values and the backward item selection method only two items of 3 and 8 should be removed. According to (zsig) column, one significant violation was observed for items 8, 3, 19, and 21, and 2 significant violations were observed for item pairs involving item 3. These viola tions indicate that the difficulty of these five items is not invariant across the range of the ability, meaning that they do not retain a constant level of difficulty across the ability continuum. Therefore, these items can be considered as potential items for removal.

In practice, AISP identifies and excludes non- or low-discriminating items, allowing all questions with sufficient psychometric quality to be incorporated in the final scales. What's more, unprompted or sub-charac ters that vary from the intended characteristic may be distinguished (Sijtsma & van der Ark, 2017).

## Discussion and conclusion

Researchers that create tests or construct multi-item surveys to measure constructs might use Mokken scaling approaches as a scaling tool (Sijtsma et al., 2021). Mokken scaling can also be used as a secondary analysis technique to assess the applicability and performance of more well-known parametric item response theory (IRT) methods, such as the Rasch family of models (Rasch, 1960), which rely on more robust statistical assumptions. It can also be used to test whether existing items are consistent with these assumptions when established items are applied to new respondent samples. Mokken models incorporate the Guttman concept into a probabilistic framework, allowing researchers to describe data while accounting for measurement error (Van Schuur, 2020).

The main advantage of NIRT over more widely used item response models, such as the Rasch model, is that it relaxes some of the rigid (logistic ogive or sigmoid shape) assumptions regarding the nonlinear behavior of response probabilities that the parametric IRT models impose (Sijtsma & Molenaar, 2019). Sijtsma and Molenaar (2019) carried out a study on the use of Non-parametric item response theory on a reading comprehension test and found out that efficient way to explore the behavior of items in scales in an attempt to order persons and items by addressing the underpinning assumptions of IRT through two nonparametric models, namely, MHM and DMM. The findings of their studies are in the same line with results of the present study in which NIRT was used for TOEFL iBT listening items.

In the typical parametric approach, the item characteristic curve is expected to follow a smooth and symmetric S-shaped function based on the family of logistic or probit cumulative distribution functions with single, 2-parameter, or more complex (3- or

**Table 2** items violating IIO

| Items | Hj[a] | IIO | | | | Backward selection | | |
|---|---|---|---|---|---|---|---|---|
| | | Active comparison (ac) | Violations (vi) | Significant violations (zsig) | Crit | Step 1 | Step 2 | Step 3 |
| **21** | **0.74** | **59** | **0** | **0** | **0** | **0** | **0** | **0** |
| 23 | 0.65 | 59 | 4 | 0 | 12 | 0 | 0 | 0 |
| 25 | 0.67 | 59 | 1 | 0 | 0 | 0 | 0 | 0 |
| 28 | 0.69 | 58 | 2 | 1 | **39** | 2 | 2 | 0 |
| 27 | 0.64 | 59 | 4 | 1 | 48 | 2 | 2 | NA |
| 22 | 0.70 | 56 | 5 | 0 | 12 | 0 | 0 | 0 |
| 24 | 0.88 | 58 | 3 | 0 | 3 | 0 | 0 | 0 |
| 20 | 0.68 | 59 | 6 | 1 | 28 | 2 | 0 | 0 |
| 16 | 0.65 | 59 | 3 | 1 | 22 | 2 | 0 | 0 |
| **3** | **0.92** | **58** | **5** | **2** | **48** | **3** | **NA** | **NA** |
| 18 | 0.80 | 59 | 3 | 0 | 12 | 0 | 0 | 0 |
| 5 | 0.89 | 57 | 3 | 0 | 8 | 0 | 0 | 0 |
| 9 | 0.90 | 58 | 0 | 0 | 0 | 0 | 0 | 0 |
| 6 | 0.81 | 59 | 2 | 0 | 3 | 0 | 0 | 0 |
| 7 | 0.86 | 59 | 4 | 0 | 16 | 0 | 0 | 0 |
| 26 | 0.81 | 59 | 2 | 0 | 4 | 0 | 0 | 0 |
| 29 | 0.72 | 58 | 2 | 0 | 0 | 0 | 0 | 0 |
| 30 | 0.83 | 58 | 3 | 0 | 5 | 0 | 0 | 0 |
| 31 | 0.82 | 58 | 3 | 0 | 5 | 0 | 0 | 0 |
| 32 | 0.83 | 51 | 0 | 0 | 0 | 0 | 0 | 0 |
| 17 | 0.74 | 59 | 0 | 0 | 0 | 0 | 0 | 0 |
| 13 | 0.65 | 59 | 4 | 0 | 12 | 0 | 0 | 0 |
| 15 | 0.67 | 59 | 1 | 0 | 0 | 0 | 0 | 0 |
| **8** | **0.69** | **58** | **2** | **1** | **39** | **2** | **2** | **0** |
| 11 | 0.64 | 59 | 4 | 1 | 48 | 2 | 2 | NA |
| **2** | **0.70** | **56** | **5** | **0** | **12** | **0** | **0** | **0** |
| 1 | 0.88 | 58 | 3 | 0 | 3 | 0 | 0 | 0 |
| 4 | 0.81 | 59 | 2 | 0 | 4 | 0 | 0 | 0 |
| **19** | **0.72** | **58** | **2** | **0** | **0** | **0** | **0** | **0** |
| 10 | 0.83 | 58 | 3 | 0 | 5 | 0 | 0 | 0 |
| 12 | 0.82 | 58 | 3 | 0 | 5 | 0 | 0 | 0 |
| 14 | 0.83 | 52 | 0 | 0 | 0 | 0 | 0 | 0 |
| 33 | 0.72 | 58 | 1 | 1 | 48 | **0** | **2** | 0 |
| 34 | 0.83 | 58 | **2** | **0** | **12** | 0 | 2 | 0 |
| 35 | 0.82 | 52 | 4 | 0 | 3 | **0** | **0** | 0 |
| 36 | 0.83 | 59 | **5** | 0 | 4 | 0 | 0 | 0 |
| 37 | 0.74 | 59 | 3 | **0** | **0** | 0 | 0 | 0 |
| 38 | 0.65 | 59 | 2 | 0 | 5 | **0** | **0** | 0 |
| 39 | 0.67 | **58** | **2** | 0 | 5 | 0 | 0 | 0 |
| 40 | **0.69** | 59 | 3 | 0 | 0 | 0 | 0 | 0 |

[a] $H^T = 0.24$

even 4-parameter models). The model becomes increasingly restrictive as the number of parameters calculated for each item decreases. The number of parameters required to characterize each item's shape and location grows, so does the number of features

in the data that the final scale can handle. The present study applied NIRT to assess the psycho- metric properties and to determine the adaptability of a TOEFL listening comprehension test. In particular, this study used the NIRT to investigate whether the measurement quality of the TOEFL listening comprehension test is satisfactory to order learners and make decisions according to the measurement results.

The work of Meijer et al. (2020) contains a full comparison between the Mokken and Rasch techniques. The decision to keep or delete an item from the final scale under a parametric IRT model approach is based in part on item fit, whether informally or formally graphically assessed, or testing using, for example, a multiple degree of freedom likelihood ratio (chi-square) test (LRT). Some features of item misfit can be traced back to a divergence from the item regression model's anticipated functional form, which is commonly logistic or normal ogive (probit). Similarly, the results of NIRT analysis to explore the TOEFL listening comprehension test showed that MHM fitted all items of the test very well as measured by the adaptability coefficient which specifies that the test's items can order students in terms of their listening comprehension well on the latent trait, meaning that students with a higher level of listening comprehension ability score higher on the test. Accordingly, it allows the researchers to interpret the participants' total scores as an indicator of student ordering on the latent variable.

It should be noted, however, that using this nonparametric approach for scale building means that the researcher only has ordinal information about the location of items and people on the latent measurement continuum, not ability assessments or item locations (Molenaar & Sijtsma, 2020). The monotone homogeneity model (MHM) and the double monotonicity model (DMM) are two NIRT models for dichotomous items that have been detailed in the literature: both were introduced in a work by Mokken (1971) over 40 years ago for dichotomous (binary) questions. Molenaar (Sijtsma & Molenaar, 2019) offered an extension of both models for polytomous (ordinal) objects a decade later.

If the data are suitably fit to the MHM for dichotomous item replies, the ordering of respondents with regard to their latent value/"ability" on the basis of the simple sum score of their right responses (denoted as X+) is a highly important quality that these results ensure. Apart from ties, another aspect of the MHM is that for each selection of items, the expected order of the persons on the latent measurement continuum is the same. Person ordering with items with monotonely nondecreasing IRFs can be thought of as "item-free," at least in theory, a property that might be useful for applied researchers who face challenges in their research that can only be overcome by exposing different individuals to different items, such as to avoid repeated exposure in repeated measure studies or panel designs (Junker & Sijtsma, 2001). If the dichotomous items fit the more restrictive DMM, it means that the items are ordered in the same order (in terms of the likelihood of the indicated response) at all points along the latent measurement continuum. Invariant item ordering (IIO) is the name for this functionality (Wind, 2017).

IIO allows researchers to order items based on their difficulty (facility) or commonality/prevalence, a feature that aids researchers in communicating valuable qualities of scale item hierarchical ordering to users. This quality is particularly valued and has been widely used, for example, in IQ or developmental assessments, as well as in rehabilitation outcome evaluation, where recovery is defined as the reacquisition of skills or functions of varying degrees of difficulty in a predictable order. On a unidimensional

continuum, this emphasizes the formation of cumulative hierarchical scales. Scales that fit the DMM have a number of other advantages; for example, if a respondent is known to have answered 6 (out of 10) dichotomous items correctly, the 6 items answered correctly were most likely the easiest items in the set (because the DMM is a probabilistic model).

This would apply to common or rare symptoms in health applications, with rare symptoms usually indicating the presence of common problems (if a cumulative hierarchical scale holds). Furthermore, if the DMM fits the item response data, the IIO property is predicted to hold in any subgroup of the population and is thus deemed to be "person-free." Because this failure can result from the presence of differential item functioning (Junker & Sijtsma, 2001), where the issue of IRF shape possibly differing across groups is considered, if a researcher finds that the DMM does not fit the data, it may be an indication that measurement invariance needs to be considered. Despite their importance, broader questions of measurement invariance and DIF in NIRT are outside the focus of this paper. Fitting the MHM does not (theoretically) imply that respondents can be sorted based on their sum score X+ for items scored in more than two categories (i.e., polytomous items) (Wind, 2017).

However, according to one simulation research (Jong & Molenaar, 1987), X+ can be used in practice as a proxy for sorting people according to their latent values without causing many major problems. Importantly, the IIO feature is not implied by fitting the DMM to polychromous item response data. If this feature is desired, different ways for evaluating this aspect have been presented (Oppenheim, 2020). These approaches have yet to be incorporated in the commercial software MSPW in, but they are useful because they have lately become freely available within the freeware statistical computing environment R's package "mokken" (Van der Ark, 2021).

The Nonparametric item response theory is a beneficial instrument for inspecting the performance of items on scales in an effort to order people and things by using two nonparametric models, MHM and DMM, to address the IRT's supporting assumptions.

The psychometric traits, as well as the adaptability, of a TOEFL iBT listening comprehension exam developed and delivered in the Iranian context were measured using NIRT. The NIRT was utilized in this study to realize if the measurement quality of the TOEFL iBT listening comprehension test is good enough to rank learners and make judgments based on the results.

Many NIRT researchers (Sijtsma & Van der Ark, 2017; Van der Ark, 2021) claim that NIRT can provide scholars with a complete examination of item adaptability structure. The findings of NIRT analysis of the TOEFL iBT listening comprehension exam revealed that MHM fit all test items very well, as indicated by the adaptability coefficient, which specifies that the test's items may order the students appropriately on the latent trait in terms of their TEOFL iBT listening comprehension and denoting that students with a higher level of listening comprehension ability score higher on the test.

As a result, the researchers can use the total scores of the participants to determine student ordering on the latent variable. Nevertheless, the $H^T$ value of 0.24 specified the fact that the test items are not precisely ordered. In addition, based on the IIO assumption, two items (items 3 and 11) should be deleted to attain IIO. Considering the IIO, we can draw the

conclusion that the order- ing of items according to their means is invariant in subgroups (Ligtvoet et al., 2010).

To summarize, this research applies Non-parametric item response theory to educational assessments and assesses the NIRT models' fit. The MHM and DMM both fit the test, demonstrating its validity and usefulness.

**Abbreviations**

| | |
|---|---|
| TEOFL | Test of English as a Foreign Language |
| MHM | Monoton homogeneity model |
| IRT | Item response theory |
| NIRT | Nonparametric item response theory |
| DMM | Double monotonicity model |
| IIO | Invariant item ordering |
| IRF | Item response functions |
| EFL | English as a foreign language |
| TPO | TOEFL preparation online |
| CTT | Classical test theory |
| ASIP | Automated item selection procedure |

**Acknowledgements**
We acknowledge that the paper is not submitted to any other journals and is not under review anywhere.

**Author's contributions**
The author read and approved the final manuscript.

**Funding**
This study did not use any source of funding, and it was done using self-funding.

**Availability of data and materials**
All the data collected for this study are available.

## Declarations

**Competing interests**
The author declares no competing interests.

Received: 27 February 2022 Accepted: 8 June 2022
Published online: 01 September 2022

**References**

Birnbaum, A. (1968). Some latent trait models. In F. M. Lord, & M. R. Novick (Eds.), *Statistical theories of mental test scores*, (pp. 397–424). Reading: Addison-Wesley.
Guttman, L. (2020). The basis for scalogram analysis. In S. A. Stouffer, L. Guttman, E. A. Suchman, P. F. Lazarsfield, S. A. Star, & J. A. Clausen (Eds.), *Measurement & prediction, The American soldier. Vol IVth edition*. New York: Wiley.
Jong, A. D., & Molenaar, I. W. (1987). An application of Mokken's model for stochastic, cumulative scaling in psychiatric research. *Journal of Psychiatry Research*, *21*, 137–149.
Junker, B. W., & Sijtsma, K. (2001). Nonparametric item response theory in action: An overview of the special issue. *Applied Psychological Measurement*, *25*, 211–220. https://doi.org/10.1177/01466210122032028.
Ligtvoet, R., Van der Ark, L. A., te Marvelde, J. M., & Sijtsma, K. (2010). Investigating an invariant item ordering for polytomously scored items. *Educational and Psychological Measurement*, *70*, 578–595. https://doi.org/10.1177/0013164409355697.
Meijer, R. R., & Baneke, J. J. (2004). Analyzing psychopathology items: A case for nonparametric item response theory modeling. *Psychological Methods*, *9*, 354–368. https://doi.org/10.1037/1082-989X.9.3.354.
Meijer, R. R., Sijtsma, K., & Smid, N. G. (2020). Theoretical and empirical comparison of the Mokken and the Rasch approach to IRT. *Applied Psychological Measurement*, *14*(3), 283–298.
Mokken, R. J. (1971). *A theory and procedure of scale analysis*. Berlin: De Gruyter.
Mokken, R. J., & Lewis, C. (1982). A nonparametric approach to the analysis of dichotomous responses. *Applied Psychological Measurement*, *6*, 417–430.
Molenaar, I. W., & Sijtsma, K. (2020). *MSP5 for Windows: A program for Nonparametric item response theory for polytomous items (version 5.0) [Software manual]*. Groningen: University of Groningen, ProGAMMA.
Oppenheim, A. N. (2020). *Questionnaire design, interviewing and attitude measurement*, (pp. 201–205). London: Pinter.
Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen: Danmarks Paedogogiske Institut.

Sijtsma, K., Emons, W. H., Bouwmeester, S., Nyklicek, I., & Roorda, L. D. (2021). 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*, *17*(2), 275–290.
Sijtsma, K., & Molenaar, I. W. (2019). *Introduction to Nonparametric item response theory*, (1st ed., ). Thousand Oaks: SAGE.
Sijtsma, K., & van der Ark, L. A. (2017). A tutorial on how to do a Nonparametric item response theory on your test and questionnaire data. *British Journal of Mathematical and Statistical Psychology*, *70*(1), 137–158.
Van der Ark, L. A. (2001). A tutorial on how to do a Nonparametric item response theory on your test and questionnaire data. *British Journal of Mathematical and Statistical Psychology*, *70*(1), 137–158.
Van der Ark, L. A. (2021). R package Mokken V 3.0.6. Retrieved from https://cran.r-project.org/web/packages/mokken.
Van Schuur, W. H. (2011). *Ordinal item response theory: Nonparametric item response theory*. Thousand Oaks: SAGE.
Van Schuur, W. H. (2020). Nonparametric item response theory: Between the Guttman scale and parametric item response theory. *Politi- cal Analysis*, *11*(2), 139–163.
Weesie, J. (1999). *MOKKEN: Stata module: Nonparametric item response theory. Software Components RePEc:boc:bocode:sjw31, RePEc Econ- Papers*. http://econpapers.repec.org/software/bocbocode/sjw31.htm.
Wind, S. A. (2017). An instructional module on Nonparametric item response theory. *Educational Measurement: Issues and Practice*, *36*(2), 50–66.

## Publisher's Note