SKASE Journal of Translation and Interpretation, 2025; 18(1): 50–77
doi: 10.33542/JTI2025-1-05

The Rasch Analysis of Item Response Theory: An Untouched Area in Evaluating Student Academic Translations

Alireza Akbari, University of Isfahan

Abstract
Classical test theory (CTT) has been widely applied to analyze objective test data. CTT utilizes aggregated data and descriptive approaches to contribute to the detection and elimination of measurement error sources and a test's unreliability. On the other hand, the Rasch analysis of item response theory (IRT) employs complex methods and scales to generate outputs that identify measurement error sources and a test's unreliability. Additionally, it is used to examine the interaction between an item's difficulty and students' ability (in our case, students' translation competence). In this regard, eighteen multiple-choice translation questions were thoroughly analyzed through the Rasch analysis to substantiate the beneficial outputs that can be maintained. The application of IRT provides a deeper insight considering "a range of information on the behavior of individual test items and individual students as well as the underlying construct being examined" (Tavakol & Dennick 2013: 838). To evaluate item facility/difficulty across students' competence, visual displays such as item characteristic curves (ICCs) and item difficulty stability (item parameter invariance) were used to identify how well an item's difficulty corresponded to students' competence. In addition, other Rasch premises and characteristics such as a test's unidimensionality, local independence, and Rasch fit statistics (infit and outfit) were analyzed to identify how well a translation test only measured one underlying construct, how well items were independent of one another, and how well items fit the model, respectively. The limitations and implications were also discussed.

Keywords*: Classical Test Theory; Rasch analysis; Item Response Theory, item difficulty; competence; translation evaluation*

1. Introduction

In an educational setting, the quality of assessment methods is as consequential as the quality of teaching and learning methods. Undergraduate, graduate, and postgraduate examination data must be assessed and evaluated based on psychometric parameters and methods to acknowledge, surveil, and improve the quality of assessment processes. Instructors and tutors must maintain a clear understanding of student performance to identify and mitigate sources of variation in examination data. Tavakol and Dennick (2013: 838) have noted that, in addition to enhancing the validity and reliability of assessments, post-examination analysis of objective test data can provide diagnostic feedback that improves teaching methods and curricula.

The post-examination analysis identifies questions deviating from normative or control boundaries. These atypical questions can diminish the overall quality of assessment questions (Wright & Stone 1979). Therefore, the primary objective of this research article is to use Rasch analysis in the context of translation evaluation products to examine assessment scores. Several factors, including unidimensionality, Rasch item fit statistics, local independence, etc., may influence these scores.

In addition to affecting scoring, psychometric evaluations like Rasch analysis provide in-depth understanding of how test items relate to the underlying traits they aim to assess. For example, confirming unidimensionality ensures the test measures one specific construct, enhancing the clarity of score interpretation. When items show poor fit statistics, it may reveal problems such as unclear wording, cultural bias, or misalignment with educational objectives, potentially compromising the assessment's fairness and effectiveness. By pinpointing and resolving these issues, educators can improve test design, enabling more accurate measurement of students' capabilities.

Furthermore, local independence—the premise that responses to individual items do not directly affect one another—is essential for ensuring the validity of test scores. This assumption can be violated by item redundancy or shared stimuli, leading to inflated reliability estimates and misrepresentation of student knowledge. Utilizing Rasch modeling helps identify these dependencies, enabling corrective measures like revising or eliminating related items. As a result, this approach contributes to the development of assessment tools that produce valid, unbiased, and generalizable outcomes.

Ultimately, employing rigorous psychometric analysis to evaluate translation evaluation products enhances translation assessment accuracy and enriches the wider realm of language education. Through careful examination of test item quality and structure, both researchers and practitioners are able to improve the congruence between educational goals and assessment methods. This congruence leads to a more transparent and impactful assessment experience for learners, which, in turn, promotes improved educational results and more precise teaching interventions.

2. State of the art

2.1. Skirmish between Classical Test Theory (CTT) and Item Response Theory (IRT)

Items in an examination can be investigated based on different theories. Two widely known theories in educational and psychological practices are classical test theory (CTT hereafter) and item response theory (IRT hereafter). The former is rooted in the traditional approach to psychometric analysis, while the latter is rooted in psychological contexts.

CTT is easy to use and offers beneficial results, while IRT is more complex; nevertheless, it provides a detailed analysis of an assessment by considering both item behavior and individual performance. In many universities, post-examination data are linked to CTT models and

parameters, including item analysis (*p-value*), item discrimination (*d-index*), descriptive statistics, standard deviations (SD), measures of peakedness and skewness, standard error of measurement (SEm), measures of item characteristics, and intra-class correlation. CTT primarily addresses sources of measurement error and the reliability of total scores (Fu & Feng 2018). In CTT, a student's competence is assessed through the number of questions answered correctly, while the performance of a group is represented by aggregate statistics. CTT identifies the relationship between total scores and relevant variables using both statistical parametric and non-parametric tests (e.g. the chi-square test, paired sample t-test). Despite its focus on test and error measurement, CTT provides limited insights into the interactions between students and items. This results in inadequate guidance on how a participant should perform on specific test items, rendering CTT incapable of assessing a participant's proficiency level (Hambleton et al. 1991; Akbari 2020; Akbari & Shahnazari 2025). CTT evaluates data quality and the extent of missing item data (Petrillo et al. 2015). Moreover, CTT is associated with sample-size-dependent aggregate values such as item difficulty and item discrimination. This implies that, regardless of the question's quality, the correlation between individual questions and the overall score varies based on sample size (Tavakol & Dennick 2013). Consequently, CTT provides minimal insight into the quality of questions compared to item-based theories, such as IRT. According to CTT, evaluators relate observed scores to true scores by addressing sources of error and factors that impact a test's reliability (Novick 1966). An observed score is defined as a combination of the estimated true scores of the test-takers plus/minus some unobservable errors (Awopeju & Afolabi 2016: 264). A true score represents what a test-taker knows, but it is influenced by various sources of error (Akbari 2022). Although CTT results prompt initial data analysis, educators and researchers must investigate the link between individual abilities (in this case, translation competence) and item difficulty or ease. To establish this link, educators should implement another theory known as IRT, which overcomes the limitations of CTT and is utilized to assess a latent trait (Baker & Kim 2004). The latent trait refers to individual ability or competence. IRT is designed to measure the degree of participants' competence (hypothetical latent construct) (Akbari 2020). As Lasnier (2000: 58) highlighted, competence is a complex ability to act that results from the integration, mobilization, and organization of various knowledge forms (declarative knowledge) and skills (cognitive, affective, psychomotor, or social) effectively applied in standard situations.

In IRT, item analysis and discrimination are utilized to investigate the correlation between test items and the competences of respondents (Reckase 2009). IRT finds application across various academic fields, including nursing, psychology, and medicine. While test items can differ in ease and complexity, and individuals possess varying competency levels, opting for IRT over CTT tends to provide more accurate results. This indicates that respondents' competences might not be the same, despite receiving identical test scores. Consequently, a respondent is more likely to answer a question correctly if there is a strong or positively skewed relationship between their competence level and the item's difficulty level.

On the other hand, the responder has a limited probability of successfully answering an item if there is a negative or negatively skewed connection between their competence and the item's difficulty. The estimate method is used to analyze and examine the qualities of the items and the competence of the responders. Fox (2010: 6) has maintained that measurement error is the sole factor causing differences in latent variable estimates across sets of items measuring the same underlying construct. Estimates of item characteristics derived from sample respondents within the same population are equivalent and only vary because of sampling error.

Three parameters are available in IRT: the 1-parameter logistic model (1-PL), also known as the multifaceted Rasch model or Rasch analysis; the 2-parameter logistic model (2-PL); and the 3-parameter logistic model (3-PL). 1-PL concentrates chiefly on an item's difficulty or easiness; 2-PL and 3-PL cover item difficulty, item discrimination, and respondents' guessing behavior.

Table 1: The distinction between CTT and IRT (Hambleton & Jones 1993: 43)

Area	Classical Test Theory	Item Response Theory
Model	Linear	Nonlinear
Level	Test	Item
Assumptions	Weak (i.e. easy to meet with test data)	Strong (i.e. more difficult to meet with test data)
Item-ability relationship	Not specified	Item characteristic functions
Ability	Test scores estimated true scores are reported on the test score scale (or a transformed test score scale)	Ability scores are reported on the scale $-\infty$ to $+\infty$ (or a transformed scale)
Invariance of the item and person statistics	No-item and person parameters are sample-dependent	Yes-item and person parameters are sample-independent; if the model fits the data
Item statistics	p (probability), r (reliability)	b (the two-parameter model), a (the one-parameter model), and c (the three-parameter model), plus corresponding item information functions

Thanks to the constraints regarding CTT (see Table 1), this research paper aims to investigate how and to what extent the Rasch model/analysis can be employed to design multiple-choice tests for evaluating the knowledge of translation students. This article aims to demonstrate that the Rasch analysis/model can be effectively applied to enhance the quality of a translation test.

2.2. Rasch assumptions and features

Within the Rasch model, an item's difficulty and individual competence must be evaluated/measured in identical logits (units). This implies that individual competence is the

genuine logarithm of the ratio of the probability of odds (success) (Statistical Consulting 2021) and the probability of failure. The odds equation is given in (1), where p denotes the probability.

(1) Odds (success) = $Ln\frac{p}{1-p}$

If $Ln\frac{p}{1-p} \geq 0$ (positive values), this will illustrate greater levels of competence. In contrast, if $Ln\frac{p}{1-p} \leq 0$ (negative values), this will demonstrate lower levels of competence. For instance, if an individual responds to 85% of the questions correctly within a test, the success ratio for the test is $Ln\frac{0.85}{0.15} \geq +1.73$ logits, demonstrating individual competence. To calculate logits regarding an item's difficulty, the numerator and the denominator must be reversed. This means that the formula for an item's difficulty is as in (2) where p denotes the probability.

(2) An item's difficulty = $Ln\frac{1-p}{p}$

For instance, if an item in a test is correctly responded to 65% of the time, its difficulty will be $Ln\frac{0.35}{0.65} = -0.61$. With that in mind, both the individual competence and an item's difficulty can be illustrated on the identical scale of units/logits. In both equations, zero shows the hub of both the competence range and difficulty range. Based on the Rasch principles, the variation between the individual competence and the difficulty of an item represents the likelihood of a correct response. Therefore, if the individual's competence is higher than the difficulty of an item, this illustrates the likelihood of a correct response. Based on the Rasch analysis, the individual competence and an item's difficulty are independently measured (Andrich 2004). This implies that disseminating items within a test cannot impact individual competence estimates. Likewise, disseminating the individual competence does not affect an item's difficulty estimates.

2.2.1. Unidimensionality

The Rasch model's central tenet is the assumption of unidimensionality. The statistical assessment of unidimensionality, however, has received little attention. The term *unidimensionality* refers to the need for a test to assess a single underlying measurement construct that accounts for variations in respondents' answers. According to Yu et al. (2007: 9), the use of unidimensionality serves to "support the validity of interpretations based on a total score, particularly when assessing development and analysis is conducted within the IRT framework." Violations of unidimensionality can significantly impact the evaluation of items and the assessment of competence. As a result, a researcher's responsibility is to ensure that a test's unidimensionality is maintained even in the presence of several aberrant items.

To state it clearly, if an item cannot support the underlying measurement construct, it must be removed from the test. The exclusion of items within a test ensures the test's construct validity. Unidimensionality in a test can be measured through the principal component analysis of residuals. The primary factor in the principal component analysis of residuals is referred to as the Rasch

factor. If an item or datum supporting the Rasch factor is excluded, the analysis of residual data will reveal additional factors, referred to as the 1st, 2nd, 3rd, 4th, and 5th contrasts. For example, suppose the eigenvalue of the first contrast is less than two. In that case, this implies that the data supporting the Rasch factor does not guarantee a further underlying measurement construct, and the unidimensionality of a test is approved. Conversely, if the eigenvalue of the first contrast is more than 2, this implies that data contributing to the Rasch factors are multidimensional and associated with different content. Besides, if the content has no significant difference, multidimensionality may not be approved, and the difference may occur by chance.

Unidimensionality can be viewed in several ways. Two comprehensive method reviews stand out. The first, by Hattie (1985: 151), assessed various conventional methods and revealed that "many lacked empirical support for the adequate assessment of unidimensionality". The second, Tate's (2003: 170) review, showed that most existing methods functioned effectively "within the limits of their associated perspectives and assumptions" when evaluating unidimensionality. Tools for measuring unidimensionality include the Dimensionality Evaluation to Enumerate Contributing Traits (DETECT) (Zhang & Stout 1999), the Test of Essential Dimensionality (DIMTEST) (Stout 1990), and Hierarchical Cluster Analysis with Proximity and Matrix (HCA/CCPROX) (Roussos et al. 1998).

2.2.2. Local independence

Another aspect of the Rasch model is local independence, which states that test items should not be related to one another. According to statistics, the probability of answering an item correctly must be independent of the responses to the other items (Tavakol & Dennick 2013). According to Baghaei (2008: 1105), there cannot be any association between two items "after the effect of the underlying trait is conditioned." As a result, $r = 0$ indicates that there is no correlation between the residuals. According to Lord & Novick (1968), a correlation is necessary since a test assesses a latent trait. In agreement with Lord & Novick, Lee (2004: 78) has stated that test items are locally dependent if substantial relationships remain after controlling for the latent trait.

In simple terms, the independence assumption is breached when the value of one item is predicted by the value of another. Consequently, items that exhibit positive correlations indicate that one of the two questions may be redundant. According to Baghaei et al. (2013: 841), a correlation coefficient (inter-correlations) exceeding 0.50% signals local item dependency, thus warranting further investigation of these items. For example, if an item has a correlation coefficient of 68% with item 2, this indicates a local dependency between items 1 and 2, suggesting that both items are essential for the test.

2.2.3. Rasch fit statistics (infit and outfit)

Rasch item fit statistics indicate the degree to which items align with the model. This means how well an item analysis/difficulty or the individual competence supports the underlying measurement construct (Smith et al. 2008). Rasch item fit statistics are applied to detect misfitting items in a test

and to gauge the test's dimensionality. The Rasch model employs chi-squared and t-test techniques using the Wilson-Hilferty transformation to calculate an expected score from an observed score. Even though data fitting the Rasch model must demonstrate minimal deviations, large deviations between observed and expected scores reveal a slant test sketch. As reported by Smith et al. (2008: 3),

> the mean square fit statistics have a chi-square distribution, and an expected value of 1, where fit statistics greater than 1 can be interpreted as demonstrating more variation between the model and the observed scores, e.g., a fit statistics of 1.25 for an item would indicate 25% more variation (or 'noise') than predicted by the Rasch model, in other words, there is an underfit with the model.

The Rasch model provides two kinds of item fit data: outfit statistics and infit statistics. The former is shown by infit-*t* (infit ZSTD) and infit mean-square residual (infit MNSQ). Outfit mean-square residual (outfit MNSQ) and outfit-*t* (outfit ZSTD) serve as instances of the latter. Mean-square values determine if the observed data are compatible with the Rasch model (Bond & Fox 2007). Specifically, values in the range of 0.70 to 1.30 are regarded as a satisfactory match (Wright & Linacre 1994). According to Wright & Linacre (1994: 371), values over 1.30 and less than 0.70 are over-fitting and misfitting, respectively. When the mean-square is 0.70, it shows a 30% deviation in Rasch-model-predicted-randomness; when it is 1.30, it shows that the data has 30% more randomness than the model predicted. Each test has its suitable item mean-square ranges for outfit and infit (see Table 2).

Table 2: Item mean-square ranges (Wright & Linacre 1994)

Reasonable Item Mean-Square Ranges for Infit and Outfit	
Type of Test	Range
MCQ* (High Stakes)	0.80-1.20
MCQ (run-of-the-mill)	0.70-1.30
Rating scale (survey)	0.60-1.40
Clinical observation	0.50-1.70
Judged (agreement encouraged)	0.40-1.20

* Multiple-Choice Questions

Infit-*t* shows to what extent a question/item fits the Rasch model. The values regarding infit-*t* are located between ±2 (see Table 3). Linacre (2002: 878) has stated that observed data higher and lower than 2 are noticeably unpredictable and are too predictable, respectively.

Table 3: Standardized values regarding infit-t (Linacre 2002)

Standardized Value	Implications for Measurement
≥ 3	Data very unexpected if they fit the model (perfectly), so they probably do not. But, with a large sample size, substantive misfit may be small.
2.00 - 2.90	Data noticeably unpredictable.
-1.90 – 1.90	Data have reasonable predictability.
≤ -2.00	Data are too predictable. Other dimensions may be constraining the response pattern.

Both infit and outfit statistics are used to assess the degree to which an individual's responses align with the Rasch model. The infit statistic, also known as inlier-sensitive or information-weighted-fit, measures the fit of an individual's responses within the expected range. On the other hand, the outfit statistic, sometimes referred to as outlier-sensitive fit, evaluates the fit of an individual's responses that fall outside the expected range. From this perspective, infit statistics provide a higher degree of sensitivity and diagnostic ability when it comes to unforeseen data related to items situated close to the student's level of competence. This is due to their ability to provide valuable insights into the relationship between the student's competence and the item's difficulty (Linacre 2011). In contrast, outfit statistics provide diagnostic characteristics and demonstrate sensitivity to atypical observed data, such as when students perceive items as either excessively simple or difficult, or when answers to items deviate significantly from their competence (Tavakol & Dennick 2013).

2.2.4. Item difficulty stability (Item parameter invariance)

Item difficulty stability is a characteristic of the Rasch model that offers valuable insights into the consistency of item values in assessments. Stability signifies that a student's skill level does not affect an item's characteristics. This is illustrated through a scatterplot of item difficulties, which reveals a correlation between high- and low-performing students. With a 95% confidence interval (CI), no items show instability related to competence on these scatterplots. In establishing calibrated items for item banks, the stability of item difficulty serves as an essential tool for identifying items that perform well across the varying competence range. This ensures that evaluators and graders have efficient access to a wide range of tested items, categorized according to their difficulty and students' abilities.

2.2.5. Item characteristics curves (ICCs)

Item characteristic curves (ICCs) are the primary building block of the Rasch model. The ICC is used to show an item's properties based on difficulty index (1-PL), discriminating index (2-PL), and guessing behavior (3-PL). The ICC provides valuable information about the test item. The ICC is used to respond to the following queries (Philip & Odunayo 2017):

- Does the item discriminate well between the good and the poor testees?
- How do the low/high ability testees respond to the item?

The ICCs possess two distinct technical qualities: item difficulty and item discrimination. The concept of item difficulty refers to the position of an item along the ability scale (Baker 2001). For example, an easy item is suitable for students with lower levels of competence, whereas a complicated item is suitable for students with higher levels of competence. Thus, the complexity of an item may be seen as an indicator of the student's proficiency level. Conversely, item discrimination refers to the degree to which an item distinguishes between students who do well (above the location index) and those who perform poorly (below the location index). The discrimination of an item is reflected in the steepness of its item characteristic curve (ICC).

In other words, the steeper the ICC, the better an item discriminates. On the other hand, "the flatter the curve, the less the item can discriminate since the probability of correct response at low ability levels is nearly the same as it is at high ability levels" (Baker 2001: 34). Using both an item, difficulty, and discrimination can form the ICC.

Based on the Rasch model, the standard model of the ICC is "the cumulative form of the logistic function" (Philip & Odunayo 2017: 25). The range of ability/competence in the ICC is from -∞ to +∞. Following Baker (2001), the range of the ICC is from −3.00 to +3.00. The ICC equation is as in (3).

(3) $$P(\theta) = \frac{1}{(1+\exp(-a(\theta-b)))}$$

In (3), "p" denotes probability, "exp (e)" is the constant 2.718, "b" represents the difficulty parameter, "a" denotes the discrimination parameter, and "θ" is the competence level ranging from -3.00 to +3.00.

3. Method

3.1. Participants' profiles and research setting

This research involved thirty-five translation students enrolled in the Bachelor of Arts program, who signed a consent form prior to their participation. The individuals were enrolled in the English Translation program at Islamic Azad University, Shahreza branch. All participants involved in the translation process were proficient in Persian (L1) and English (L2). Despite variations in

participants' English language proficiency levels, it was generally assumed that their proficiency was satisfactory. This assumption was based on the requirement for participants to pass relevant courses, such as Advanced Translation I and II, and Journalistic Translation as part of their study programs. To evaluate the participants' cognitive abilities, eighteen multiple-choice translation questions related to economics were administered. The participants were instructed to provide their responses within a specified 90-minute time frame, corresponding to the class hour. The examination was sourced from the database containing economic translation banks of questions of the Islamic Azad University. Furthermore, the inquiries addressed the topics covered in the classroom. The marking of each question followed a dichotomous system, where a score of 1 was given for a correct reply, and a score of 0 was assigned for an erroneous response. The potential maximum score for this examination was 18. It is essential to acknowledge that no penalty was imposed for wrong answers.

In light of the COVID-19 pandemic, respondents were asked to provide their answers through the Islamic Azad University, Shahreza branch's online Learning Management System (LMS). To minimize cheating among participants, an approach was used in which the questions were designed in a non-consecutive manner. Additionally, using a secure LMS for the assessment allowed for controlled access and timestamped submissions, improving the integrity of data collection by minimizing the chance of unauthorized collaboration or outside help. This method not only maintained the validity of the outcomes but also enabled effective monitoring and analysis of participant engagement throughout the remote assessment.

3.2. Psychometric packages

This paper employed Winsteps and STATA to interpret the results of the Rasch model. Winsteps is a psychometric software that converts dichotomous items into linear measures. It is also used for evaluating multiple-choice questions (MCQs), rating scales (RSs), and partial credit (PC). The Winsteps software connects qualitative analysis to quantitative analysis.

Besides, this research paper used STATA to interpret item characteristic curves (ICCs). STATA is used to measure the following analyses, namely (to name a few) Bayesian econometrics, interval-censored Cox model, multivariate meta-analysis, panel-data multinomial logit, Leave-one-out meta-analysis, Galbraith plots, zero-inflated ordered logit model, item-response theory, and treatment-effect lasso.

4. Data Analysis and Results

4.1. Unidimensionality assumption

As noted, a test should assess a single measurement construct that accounts for variations in examinee responses. Thus, this research utilized principal component analysis of residuals to

ensure the test's unidimensionality. To verify this unidimensionality, it is essential to check the eigenvalue for our test.

Table 4: Unidimensionality premise of the whole test

Standardized Residual Variance in Eigenvalue Units = Questions Information Units	
Rasch factor eigenvalue of the whole test (items)	14.0729
1st contrast eigenvalue	**1.6453**
Total No of MCQs	18

Table 4 displays essential indicators pertaining to the unidimensionality premise of the entire test as derived from Rasch analysis. The eigenvalue of the Rasch factor for all items measures 14.0729, indicating the total variance explained by the primary measurement dimension. An eigenvalue lower than 2.0 in the first contrast indicates that the residual variance not explained by the primary Rasch dimension is minimal and does not establish a coherent secondary dimension. In simpler terms, the test items collectively signify a single dominant latent trait or construct, confirming the unidimensionality assumption. This observation is vital as it substantiates the interpretation that the test scores reflect one underlying ability or characteristic, thereby validating the practice of summing item scores to create a meaningful measure. Furthermore, with 18 multiple-choice questions contributing to this structure, the results reveal that the test items work together effectively, free from significant multidimensional noise that could interfere with the measurement.

4.2. Local independence assumption

Local independence is assumed when the order in which questions are asked does not affect question difficulty. Typically, the breach of this assumption is examined using item pairs (Debelak & Koller 2019). A correlation coefficient (inter-correlations) greater than 0.50% indicates local item dependency, warranting further investigation of the items.

Table 5 displays the results regarding the local independence assumption among test items, derived from Pearson's correlations and disattenuated correlations among question clusters identified through principal component analysis contrasts. The correlations among item clusters generally fall below 0.50, suggesting that responses to these items are not strongly correlated beyond what is expected from the underlying trait being measured. This low correlation level reinforces the idea that each item operates independently and does not have significant shared variance with other items.

Table 5: Local independence confirmation

Approximate Relationships between the Student's Measures			
Principal Component Analysis Contrast	Questions Clusters	Pearson's Correlations	Disattenuated Correlation
1	1-3	-0.2186	-1.0000
1	1-2	0.3638	1.0000
1	2-3	0.3307	1.0000
2	1-3	-0.3167	-1.0000
2	1-2	0.2527	1.0000
2	2-3	0.1472	0.3306
3	1-3	-0.1528	-1.0000
3	1-3	-0.1801	-1.0000
3	2-3	0.3692	0.9718
4	1-3	0.1059	1.0000
4	1-2	0.3784	1.0000
4	2-3	0.1426	1.0000
5	1-3	0.0495	1.0000
5	1-2	0.1888	1.0000
5	2-3	0.4432	1.0000

In addition, the disattenuated correlations, which account for the measurement error, remain within acceptable limits and do not indicate a significant dependency between item clusters. This further confirms that there is no substantial violation of local independence. In other words, the probability of a correct response on one item is not directly influenced by responses to another item, ensuring that the test measures the intended construct reliably without item overlap or redundancy.

4.3. Rasch infit and outfit statistics

The degree to which the test items fit the model and the identification of misfitting items are done using Rasch infit and outfit statistics. The dimensionality of the test is measured using Rasch fit statistics. Rasch fit statistics are shown in Table 6, which includes outfit mean-square residual (infit ZSTD) and outfit-*t* (infit MNSQ) (or standardized as a z-score) and infit mean-square residual (infit MNSQ) and infit-*t* (infit ZSTD). It is necessary to include mean-square values in tests within the permissible range of 0.70 to 1.30 (0.70 MNSQ 1.30). Furthermore, t-test results falling between the range of -2 and +2 (i.e. $-2 < \text{ZSTD} < +2$) are legitimate ranges that must be included in a test.

Table 6: Rasch infit and outfit statistics

					Infit		Outfit					
Entry Number	Total Score	Total Count	Measure	Model S. E.	MNSQ	ZSTD	MNSQ	ZSTS	Exact OBS%	Match EXP%	Questions	
16	2	35	3.08	0.75	0.93	0.07	5.07	2.35	94.3	94.2	1-4-2-3-1-4	INVS
14	4	35	2.26	0.56	1.03	0.20	2.63	1.86	88.6	88.4	1-4-2-3-4-1	INVS
7	24	35	-1.08	0.40	1.30	1.75	2.03	2.85	60.0	74.2	1-4-3-2	INVS
4	30	35	-2.27	0.52	1.02	0.17	2.00	1.51	88.6	86.2	1-3-4	INVS
6	26	35	-1.42	0.42	1.19	0.92	1.61	1.57	77.1	77.5	3-4-1	INVS
2	34	35	-4.17	1.03	1.08	0.39	0.94	0.42	97.1	97.1	2-3	
5	28	35	-1.80	0.46	1.05	0.28	1.08	0.32	80.0	81.5	2-1-4	
15	1	35	3.84	1.02	1.01	0.32	0.59	0.13	97.1	97.1	1-3-2-4-1-3	INVS
12	6	35	1.73	0.48	1.00	0.08	0.77	-0.30	80.0	83.3	1-3-2-4-3	
10	23	35	-0.92	0.39	0.95	-0.21	0.88	-0.37	74.3	72.7	2-4-3-1	
11	27	35	-1.60	0.44	0.91	-0.30	0.76	-0.51	80.0	79.1	1-3-1-2-4	
8	22	35	-0.77	0.39	0.87	-0.82	0.83	-0.68	82.9	71.1	1-4-2-3	
1	18	35	-0.20	0.37	0.84	-1.25	0.79	-1.07	77.1	66.3	1-4	
17	1	35	3.84	1.02	0.79	0.05	0.18	-0.45	97.1	97.1	1-4-3-1-2-4	INVS
13	7	35	1.51	0.45	0.71	-1.20	0.50	-1.21	82.9	80.7	1-4-3-2-4	INVS
9	29	35	-2.02	0.48	0.70	-1.06	0.50	-1.05	88.6	83.7	1-3-2-4	INVS

According to Table 6, items 16, 14, 7, 4, 15, 13, 17, 9, and 6 must be investigated because they are located outside the acceptable ranges ([0.70 < MNSQ < 1.30] and [-2 < ZSTD < +2]). These items do not contribute to the underlying measurement construct and students' competence.

4.4. Item difficulty stability

Item difficulty stability provides helpful information about the features of items' values. This feature demonstrates a correlation between high and low-performing students. Using item difficulty stability, one can substantiate the extent to which an item's difficulty differs between high/and low-performing students. Figure 1 illustrates a scatterplot of items' difficulty from high-performing versus low-performing participants. Considering a 95% confidence interval (CI), no unstable items regarding competence can be detected on such scatterplots.

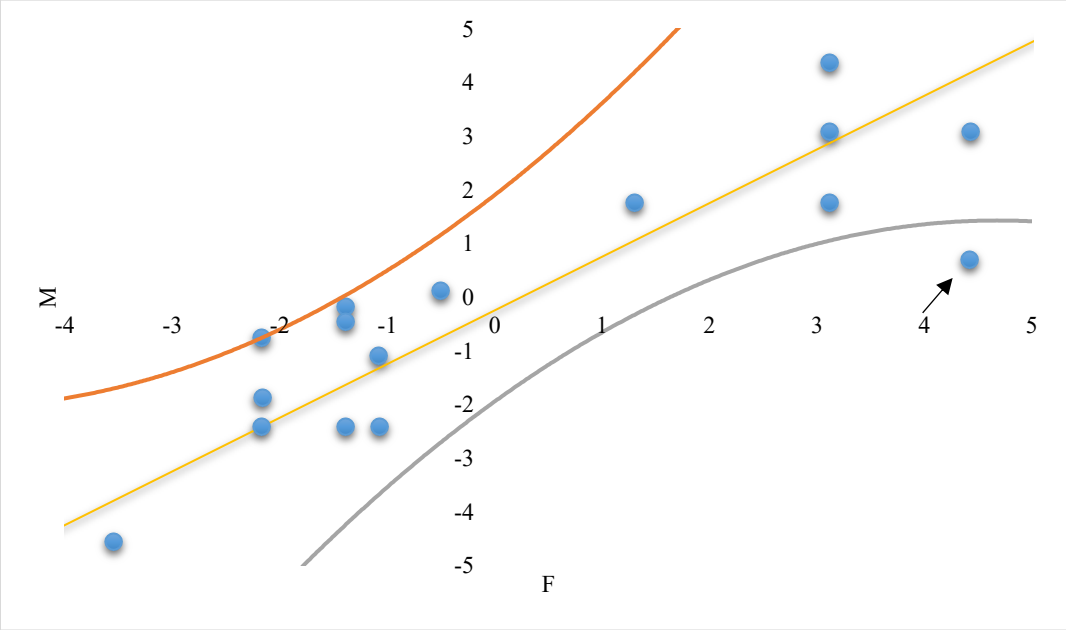
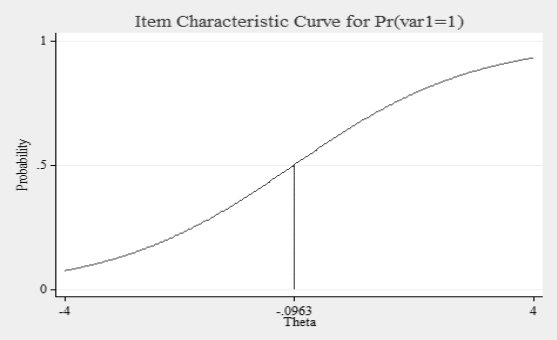


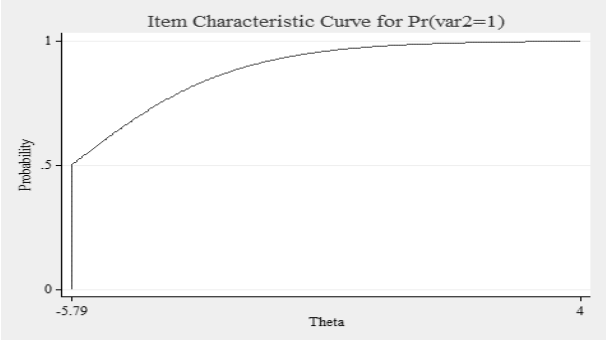
Figure 1: Item Difficulty Stability

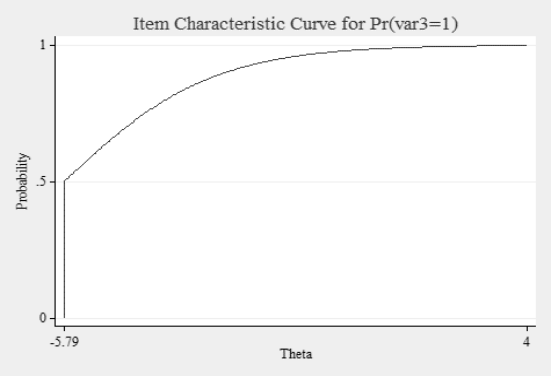
As shown in Figure 1, all plotted items except one lie within the diagonal areas (the control limits). It is worth noting that the straight line between two diagonal lines is not a regression line, but rather a Rasch-modelled relationship. According to Figure 1, we can conclude that the difficulty of all items, except for Item 12 (Appendix A), is stable (invariant). An item outside the diagonal areas must be investigated to facilitate the psychometric fashion among high and low-performing participants.

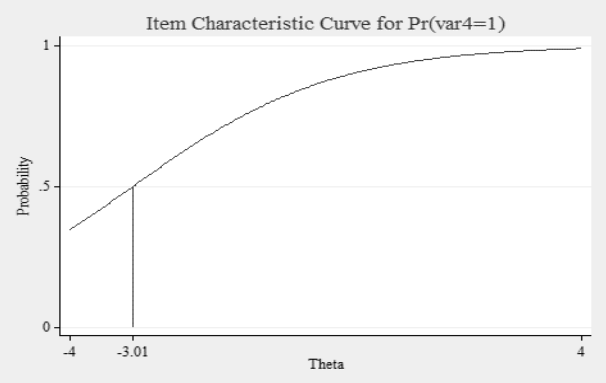
4.5. Item Characteristic Curves

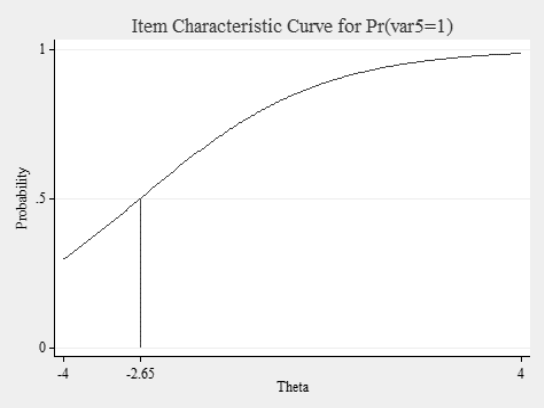
ICCs are used to show the difficulty of each item. The application of the ICCs provides beneficial information about the test item. ICCs have two technical properties: item difficulty, which is indicated by a location index, and item discrimination, which assesses the steepness of the curves.

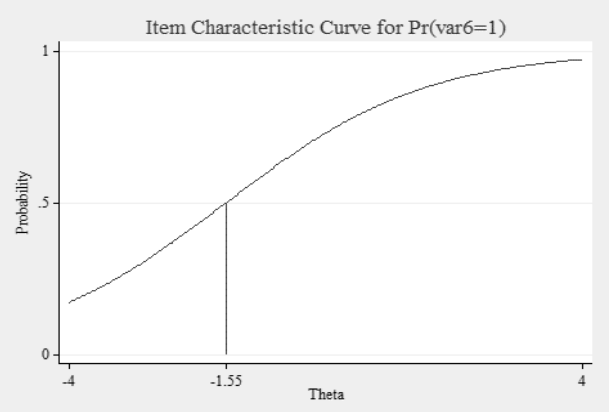

Item Characteristic Curve for Pr(var1=1)
Probability
1
.5
0
-4
-.0963
4
Theta

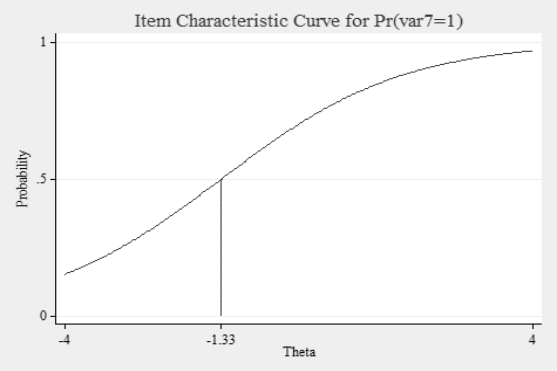

Item Characteristic Curve for Pr(var2=1)
Probability
1
.5
0
-5.79
4
Theta

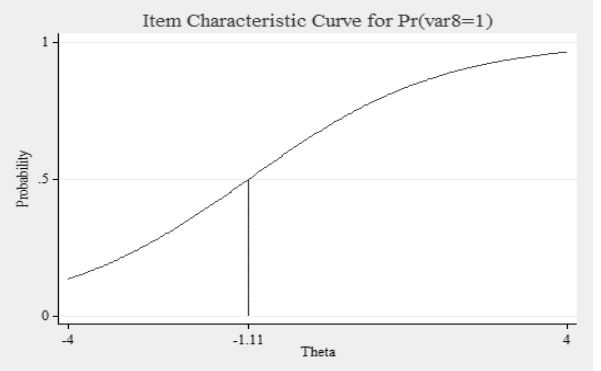

Item Characteristic Curve for Pr(var3=1)
Probability
1
.5
0
-5.79
4
Theta

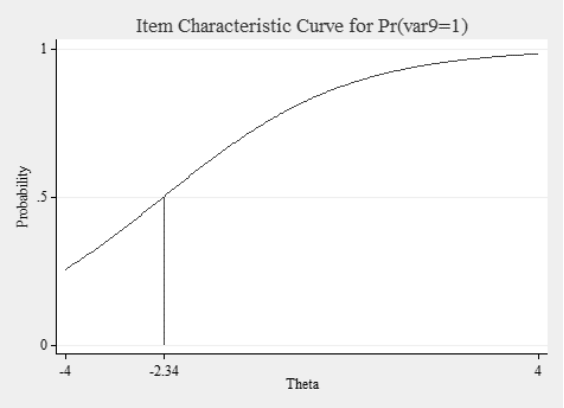

Item Characteristic Curve for Pr(var4=1)
Probability
1
.5
0
-4
-3.01
4
Theta

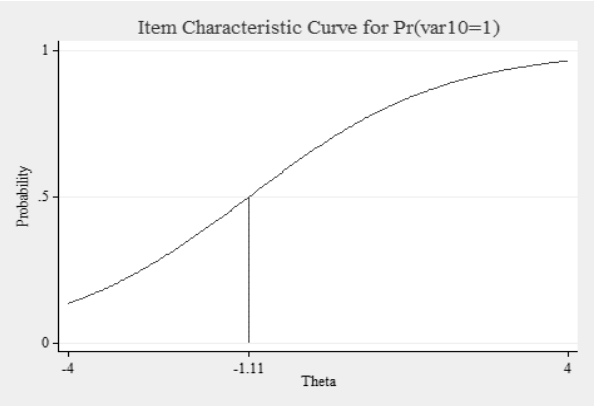

Item Characteristic Curve for Pr(var5=1)
Probability
1
.5
0
-4
-2.65
4
Theta

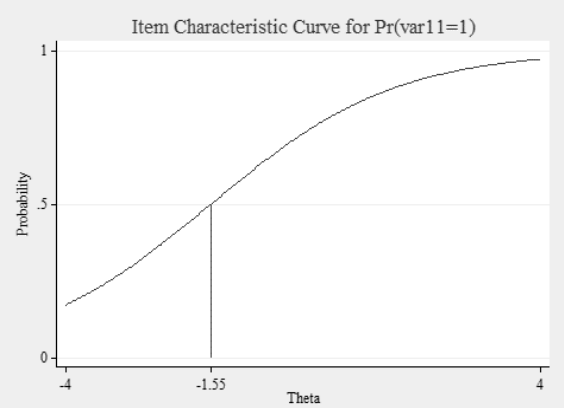

Item Characteristic Curve for Pr(var6=1)
Probability
1
.5
0
-4
-1.55
4
Theta

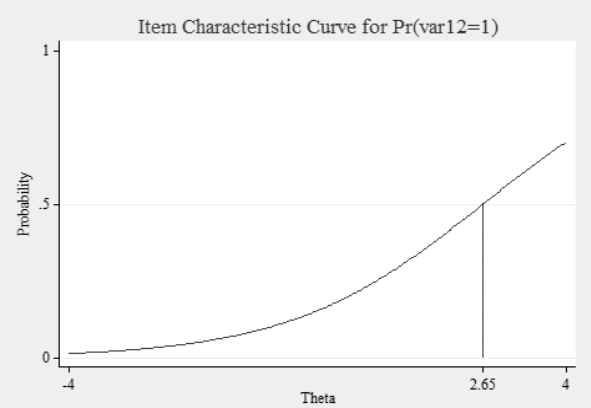

Item Characteristic Curve for Pr(var7=1)
Probability
1
.5
0
-4
-1.33
4
Theta

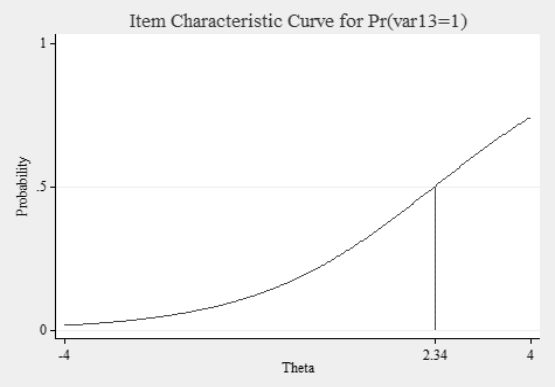

Item Characteristic Curve for Pr(var8=1)
Probability
1
.5
0
-4
-1.11
4
Theta

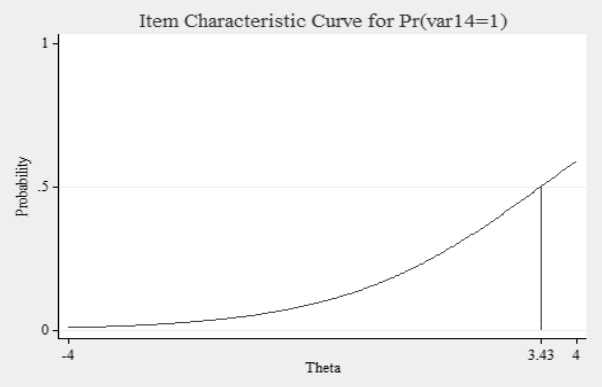

Item Characteristic Curve for Pr(var9=1)
Probability
1
.5
0
-4
-2.34
4
Theta


Item Characteristic Curve for Pr(var10=1)
Probability
1
.5
0
-4
-1.11
4
Theta


Item Characteristic Curve for Pr(var11=1)
Probability
1
.5
0
-4
-1.55
4
Theta


Item Characteristic Curve for Pr(var12=1)
Probability
1
.5
0
-4
2.65
4
Theta


Item Characteristic Curve for Pr(var13=1)
Probability
1
.5
0
-4
2.34
4
Theta


Item Characteristic Curve for Pr(var14=1)
Probability
1
.5
0
-4
3.43
4
Theta

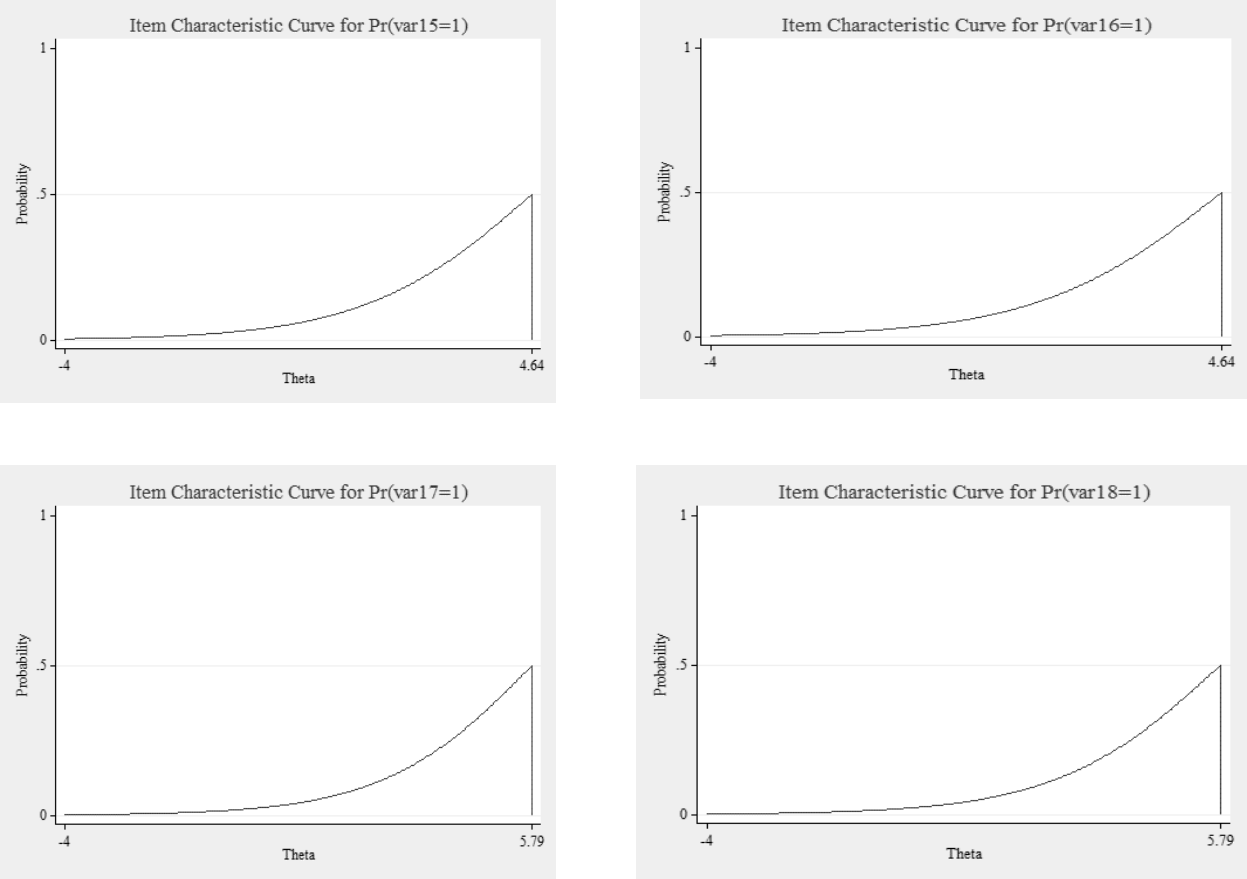


Figure 2: ICCs for each item

According to Figure 2, the X-axis (θ) demonstrates participants' competence, and the Y-axis is associated with the results concerning S-shaped curves. Akbari (2020: 487) noted that "the estimate of the difficulty degree of items corresponds to the ability scale at which the degree of the probability corresponds to 0.05". In this light, the degree of difficulty of each item is as follows:

ITEM 2 (an easy item) (Correlation Coefficient = -5.794838) < ITEM 3 < ITEM 4 < ITEM 5 < ITEM 9 < ITEM 11 < ITEM 6 < ITEM 7 < ITEM 10 < ITEM 8 < ITEM 1 < ITEM 13 < ITEM 12 < ITEM 14 < ITEM 16 < ITEM 15 < **ITEM 18 (a complex item)** (Correlation Coefficient = 5.790611).

5. Discussion

CTT, or the true score theory, presupposes systematic effects among students' responses associated with variation in the ability of interest (Magno 2009: 6). The central focus of CTT is on the observed test scores, where an observed score (TO) equals a true score (T) plus an error score (E). An error source and a true score are independent. With that in mind, CTT assumes every individual has a true score maintained if no errors are found in measurement. CTT concentrates on the following factors: (1) total test scores, (2) response frequency, (3) the test's reliability, and (4) and r_{it} values (item-total correlation). Although the mentioned factors are widely applied, their

application yields three bottlenecks. The first bottleneck is that item difficulty estimates are group-dependent. This suggests that "a test item functions be easy or difficult given a sample of examinees and these indices change when a different sample takes the test" (Magno 2009: 4). The second one is that item (*p*-value) and person parameters are sample-dependent. The third limitation is that examinees' competence scores are test-dependent. This implies that an examinee's competence is subject to various factors, resulting in the test's weak consistency (or replicability).

In the context above, although the evaluation of psychometric properties is primarily conducted through the application of CTT, scholars' surveillance has increased towards utilizing more complex approaches and models. To achieve this, various scholars have advocated for the replacement of classical theories with updated methods and theories across disciplines. In this direction, introducing the Rasch model paves the way for a practical and complex method. This model is associated with IRT and is used to evaluate the psychometric properties of measurement tools. Compared to CTT, applying the Rasch model procures precise results. According to Dabaghi et al. (2020: 174), several issues arise when using classic methods that treat raw scores, linear combinations of these scores, and responses with ordinal scales as interval scale data.

The Rasch model establishes a probabilistic continuum on which two factors are located: item difficulty and students' competence. To build the Rasch continuum, three significant principles must be observed.

(1) The probability of an item's difficulty is the result of the juxtaposition of that item with all other items.
(2) Difficulties are ascertained using item comparison, resulting in establishing a relative scale, hence giving rise to an unlimited array of scale points (OECD 2009). This implies that the most common scale point locates items' difficulties at zero.
(3) The Rasch continuum allows for probabilistic calculations of item difficulty, which can vary for different subpopulations (ibid.).

To maintain the integrity and objectivity of assessments in translation evaluation, it is necessary to establish valid and reliable assessments that adhere to the norms of fairness and objectivity outlined in the curriculum (Tavakol & Dennick 2013). This research employed the Rasch model as an alternative to CTT to preserve diagnostic information about "perfect" or objective translation tests. By using the Rasch model, translation examiners can develop fair examinations.

When examining test data using Rasch analysis, it is essential to examine the concatenation of stages. The first step is to consider the test's dimensionality. Unidimensionality ensures that the test measures only one underlying measurement construct (see Table 7). For instance, in a translation test, all items within a test must measure only students' competence in translation alone. Practically, unidimensionality cannot be applied due to several factors, including cognitive, personality, and test performance factors (e.g. motivation, guessing, and anxiety). Thus, the test's unidimensionality is met when the test only measures one dominant component: the subjects' achievement (Parmaningsih & Saputro 2021). Smith (1996: 27) stated that unidimensionality is met when highly correlated factors dominate the exam data; therefore, if "one factor dominates, use Rasch". To measure the test's dimensionality, six types of datasets are available:

Table 7: Dimensionality datasets (Tennant et al. 1996)

Details of Datasets		
Dataset	**Structure**	**Contents**
1	Unidimensional	20 items.
2	Two orthogonal dimensions ($r < 0.05$)	10 items in each dimension. Items generated in difficulty order (1=easiest, 20=hardest). Interlaced items with item 1 assigned to dimension 1 and item 2 assigned to dimension 2 … to ensure equal difficulty for each dimension.
3	Two orthogonal dimensions ($r < 0.05$)	10 items in each dimension. Items generated in difficulty order (1=easiest, 20=hardest). Dimensions stacked with easy items 1-10 in Dimension 1 and hardest items 11-20 in Dimension 2.
4	Two orthogonal dimensions ($r < 0.05$)	16 items in Dimension 1 and 4 items in Dimension 2. (Items 5,10,15,20). Items generated in difficulty order (1=easiest, 20=hardest).
5	Two correlated dimensions ($r = 0.70$)	10 items in each dimension. Items generated in difficulty order (1=easiest, 20=hardest). Interlaced items with item 1 assigned to dimension 1 and item 2 assigned to dimension 2 … to ensure equal difficulty for each dimension.
6	Two correlated dimensions ($r = 0.70$)	10 items in each dimension. Items generated in difficulty order (1=easiest, 20=hardest). Dimensions stacked with easy items 1-10 in Dimension 1 and hardest items 11-20 in Dimension 2.

As noted earlier, principal component analysis of residuals evaluates unidimensionality. This analysis contests the assertion that residuals can be seen as mere random noise. It achieves this by pinpointing the component that explains the largest share of residual variation (Arrindell & van der Ende 1985). This component represents the first contrast in the correlation matrix of residuals (ibid.). Statistically, it can be concluded that if the eigenvalue for this first contrast falls below 2.0, the idea of residuals being random noise is not refuted. The first comparison ($1.6453 \leq 2$) showed apparent variations in the residuals, suggesting that the significance of the loadings for the first contrast supports the concept of unidimensionality.

Determining the probability of each response incurs a cost. This forms the basis for the assumption of independence that scholars and researchers use to assess the likelihood of responses. Statistically, Item X is considered independent of Item Z if the occurrence of Item X does not influence the likelihood of Item Z occurring. When we measure the likelihood of a response, we measure the likelihood that "the first is right under the condition that the second is wrong for a person at a given ability and items at given difficulty levels" (Lee 2004: 84). This is the premise of the local independence of the Rasch model. Items that are put into the Rasch model must be independent. This implies that no correlations should exist between two item clusters. As shown in Table 5, the item clusters were not correlated through the latent trait that the test measured ($r < 0.50$). In the case of significant correlation coefficients ($r > 0.50$), items are locally dependent, and "there is a subsidiary dimension in the measurement which is not accounted for by the main Rasch dimension" (Baghaei 2008: 1106). The findings in Table 5 supported the local item

independence premise, prevented reliability inflation, and provided an accurate assessment of the translation test's quality. According to Wang & Wilson (2005: 6), if the assumption of local item independence is not maintained, any statistical study relying on this assumption will provide inaccurate results. Inaccurate estimation of latent variables and item parameters is a common consequence of model misspecification. This, in turn, may lead to erroneous conclusions in later statistical analyses, including assessments of group disparities and correlations among latent variables.

Rasch fit statistics are used to check for overfitting and misfitting items. Misfitting and overfitting items must be excluded from further analysis due to "the violation of the model assumption or redundancy" (Shun-Chin et al. 2011: 127). Two types of Rasch fit statistics were applied in this research: (1) the mean-square residual (MNSQ) and (2) standardized z-score (standardized fit statistics) (ZSTD). Infit MNSQ is used to diminish the influence of outlying residuals. Infit MNSQ weights "each squared standardized residual" through its statistical information (Waterbury 2020: 61). Compared to the outfit, infit is less influenced by "extreme outlying residuals" (Waterbury 2020: 61). Variance-weighted is the version to calculate both MNSQ and ZSTD. Wright & Linacre (1994) and Linacre (2002) proffered cut-off values from 0.70 to 1.30 (MNSQ) and ±2 (ZSTD), respectively. As stated in Table 6, the items (2, 5, 15, 12, 10, 11, 8, 1, 17, 13, 9, 18) were within the acceptable range and fit the model. However, items 16, 14, 7, 4, and 6 were excluded from further analysis due to their misfitting and overfitting values.

Item difficulty stability or item parameter invariance shows an "invariance up to linearity across a non-equivalent group or measurement conditions" notwithstanding sample characteristics (high and low-performing students) (Paek et al. 2021: 50). Item difficulty stability holds when parameter values are stable (invariant) across various examinee groups. To assess item difficulty stability (parameter invariance) using samples, the approximate values of the item parameters are applied to determine whether the Rasch model is appropriate across various conditions. With that in mind, as shown in Figure 1, only one item (Item 12) lay outside the diagonal areas. This may be because item 12 was formulated in a biased or non-representative way. Therefore, this item must be investigated to fine-tune the test's psychometric properties among high and low-performing participants.

Item characteristic curves (ICCs) are used to determine the statistical association between item discrimination and item difficulty. ICCs are used to assess the association between participants' competence and their likelihood of responding to an item correctly. Item discrimination echoes the steepness of curves; likewise, maximum steepness occurs when the possibility of correctly responding to an item is 0.50. In this direction, item difficulty is the "value of *theta* where this maximum steepness occurs, in other words, where the probability of getting the item correct is 0.50" (Park et al. 2020). According to Figures 2 and 3, for any given item, the likelihood of a correct response approximates 0 for low-performing and 1 for high-performing participants. The values of item difficulties are -4, -0.09, and +4 for items 2 (the easiest item), 1 (the neutral item), and 18 (the most difficult item), respectively.

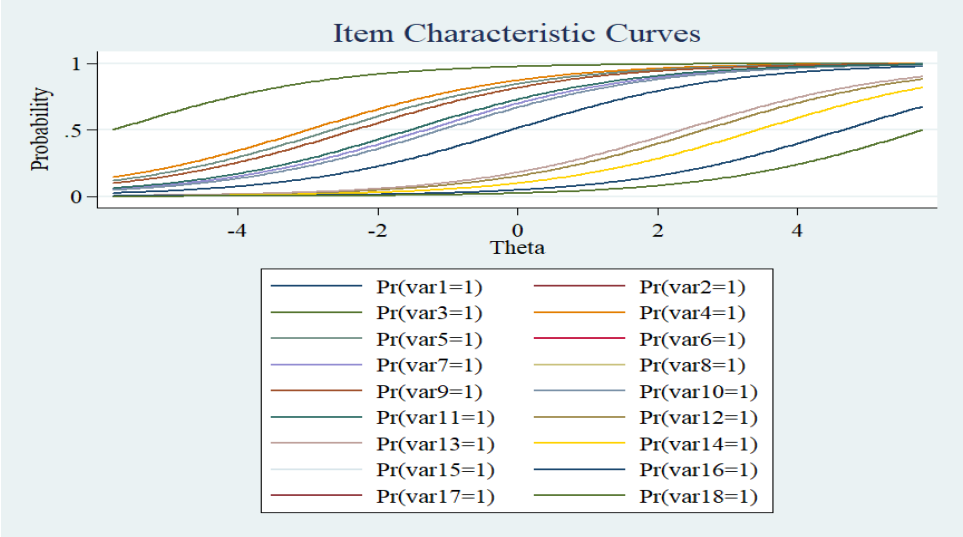


Figure 3: ICCs for the whole items

6. Conclusion

This research paper attempted to apply the Rasch model of item response theory as a method of post-examination analysis in translation evaluation products. This case has been untouched in translation testing and translation quality assessment. The Rasch model examines the relationship between an item's difficulty and a participant's ability. Therefore, the application of the Rasch analysis is to show an interaction between students' different competences and exam items. The Rasch analysis provides a foundation for diagnostic and quality feedback on test items and students' competence, enabling educators and university professors to design effective tests. Besides, using the Rasch model in a classroom context provides beneficial information regarding a test's quality independent of students' competence. Moreover, the Rasch analysis empowers test constructors to evaluate the effectiveness of their assessment.

6.1. Limitations of the research

This paper examined the potential for several bottlenecks. One of the bottlenecks is the small number of translation participants who partook in this research. This research was conducted during the COVID-19 pandemic, making it impossible to access many participants. Therefore, the data obtained may need to be more precise when compared to data from many participants, as it may be affected by incorrectly ordered parameters (Chen et al. 2013). Besides, the translation MCQ task was virtually done (not in real life). The researchers could not control the participants' behaviors (such as cheating or consulting) to remove any moderator variables, including offline and online dictionaries. Another bottleneck is that a researcher must have adequate knowledge of statistical software packages such as STATA, Winsteps, and R to provide and interpret Rasch's data precisely.

6.2. Implications of the research

Unlike translation scoring methods/rubrics following the principles and regulations of CTT, such as holistic translation scoring (an intuitive-impressionistic method), analytic method (based on a matrix of error levels and types), and preselected items evaluation (PIE) method (based on traditional p-value and d-index), the application of the Rasch analysis in translation scoring, in particular, and IRT, in general, will provide accurate scoring results. For instance, the Rasch model has the potential to be applied in the optimized version of the preselected items evaluation (OPIE) method. The general principles of the OPIE method follow IRT (Akbari et al. 2025). OPIE calculates the degree of an item's difficulty based on Feldt's p_G-value ($0.55 \leq p_G\text{-value} \leq 0.67$), taking the guessing parameter into account. However, the application of Rasch fit statistics facilitates the achievement of accurate results in the OPIE method. Additionally, unidimensionality and item local independence are two robust factors that are often overlooked in holistic, analytic, PIE, and OPIE methods. It is hoped that applying the Rasch assumptions and features takes one step towards objectifying translation evaluation products and designing high-quality translation tests.

Appendix A: Item difficulty stability

Entry	F	S.E.	M	S.E.	Upper x - Identity	Upper y - Identity	Lower x - Identity	Lower y – Identity	Upper x - Empirical	Upper y - Empirical	Lower x - Empirical	Lower y - Empirical	t-statistic
1	-0.5026	0.5239	0.114	0.5319	-0.80299	0.414391	0.660316	-1.04892	-0.82809	0.412893	0.702701	-0.9928	-1.15529
2	-3.5478	1.0725	-4.5456	1.8636	-6.03091	-2.06249	-1.81656	-6.27684	-6.37196	-1.95224	-1.87293	-6.08358	0.349681
4	-2.1657	0.6841	-2.4119	0.7969	-3.19509	-1.38251	-1.13658	-3.44102	-3.35254	-1.32206	-1.18666	-3.31093	0.000262
5	-1.3888	0.577	-2.4174	0.7982	-2.74535	-1.06085	-0.81492	-2.99128	-2.90659	-1.02365	-0.8617	-2.90143	0.794665
6	-2.1672	0.6844	-0.7597	0.5583	-2.20606	-0.72084	-0.47492	-2.45198	-2.26787	-0.66726	-0.47358	-2.3149	-1.87201
7	-1.08	0.5499	-1.08	0.5841	-1.74322	-0.41678	-0.17086	-1.98914	-1.82554	-0.39539	-0.17752	-1.90873	-0.30655
8	-1.3893	0.577	-0.1694	0.5337	-1.42665	-0.13205	0.113873	-1.67257	-1.46295	-0.10176	0.142238	-1.57577	-1.86496
9	-2.1628	0.6836	-1.8734	0.6859	-2.84415	-1.19205	-0.94612	-3.09008	-2.96722	-1.13472	-0.98265	-2.9571	-0.5528
10	-1.3892	0.577	-0.4581	0.5422	-1.57663	-0.27067	-0.02475	-1.82255	-1.62656	-0.24013	-0.00845	-1.72601	-1.48656
11	-1.0722	0.55	-2.4193	0.7987	-2.57315	-0.91835	-0.67243	-2.81907	-2.73539	-0.89204	-0.71837	-2.74421	1.135521
12	4.4252	1.8652	0.6934	0.5496	0.776665	4.341935	4.58786	0.53074	0.796989	4.02509	4.636384	0.499476	1.792696
13	1.3019	0.6228	1.76	0.67	0.757451	2.304449	2.550374	0.511526	0.795695	2.224835	2.675906	0.498287	-0.76963
14	3.1226	1.0749	1.7587	0.6698	1.322435	3.558865	3.80479	1.07651	1.372403	3.36586	3.918482	1.027863	0.882723
15	4.4316	1.8707	3.0837	1.0535	1.776605	5.738695	5.98462	1.53068	1.870818	5.435199	6.171992	1.485544	0.513275
16	3.1149	1.0717	3.0815	1.0525	1.749107	4.447293	4.693218	1.503182	1.824083	4.266823	4.89963	1.442629	-0.14149
17	3.1215	1.0744	4.3608	1.8555	1.762883	5.719417	5.965342	1.516958	1.818987	5.556863	6.304485	1.437949	-0.6927

Appendix B: Economic multiple-choice translation questions

1- as accumulation proceeds, there is a steady mean level of profits in the growing country.
a. ثابت b. موثر c. مشخص d. موقتی

2- marshal made many pronouncements on current problems.
a. اعتراف تکان دهنده b. مسیولیت عمده c. اظهارات زیاد d. نتیجه گیری های زیاد

3- In the slump, conventions had broken down and expectations had nothing to go on.
a. در شرایط عدم اطمینان b. با وجود ناامیدی c. در هنگام مباحثه d. در هنگام رکود

4- There are some difficulties about the marginal productivity of capital let alone the coordinating function.
a. چه برسد به b. با احتمال آنکه c. در ضمن d. مبادا که

5- Indeed, it can develop into runaway inflation.
a. مارپیچ تورمی b. شکاف تورمی c. تورم متوقف شده d. تورم افسارگسیخته

6- The ponderance of declining industries, does not provide an adequate explanation of the persistence of high unemployment.
a. سلطه صنایع رو به زوال b. شکست صنایع کوچک c. ظهور صنایع کم بازده d. افزایش صنایع کوچک

7- high demand for these commodities bids up their price and encourages their productions.
a. fixes b. changes c. decreases d. increases

8- economists of widely differing political persuasions.
a. اقتصاددانان مختلف با عقاید سیاسی بسیار متفاوت
b. اقتصاددانان با اعتقادات سیاسی کاملا گوناگون
c. اقتصاددانانی که وابستگی های سیاسی بسیار متفاوتی دارند
d. اقتصاددانان با جهت گیری سیاسی متنوع

9- high prices choke off the demand for oil-related commodities.
a. قیمت های بالا موجب افزایش تقاضا برای کالاهای مرتبط به نفت می شود
b. قیمت های بالا موجب توقف تقاضا برای کالاهای وابسته به نفت می شود
c. قیمت های بالا عرضه کالاهای مشتق شده از نفت را با مشکل مواجه می کند
d. قیمت های بالا عرضه کالاهای وابسته به نفت را رونق می بخشد

10- when labor is the only cost, commodities ought to exchange at prices corresponding to the labor time embodied in them.
a. زمان اختصاص یافته به کار
b. مقدار زمان صرف شده برای کار
c. زمان کار متباور در آنها
d. محصول یک ساعت کار

11- The higgling and haggling of the market establishes an equilibrium.
a. بالا و پایین رفتن b. کم و زیاد کردن c. مناسبات برابر d. چک و چانه زدن

12- in Wicksell's theory of distribution, workers and means production are separate factor but all on the same footing.
a. در مرحله رشد مساوی b. دارای وضع یکسان c. با سود برابر d. دارای ابعاد مشابه

13- at the time of high inflation, saving is discouraged.
a. مردم از نگهداری پولشان در بانک دلسرد می شوند
b. پس انداز باعث دلسردی میشود
c. ذخیره پول مایوس کننده می شود
d. مردم تشویق به عدم پس انداز میشوند.

14- the new decision proved not to be in the public interest.
a. ثابت b. موثر c. مشخص d. موقتی

15- they plan to produce fuel-efficient aircraft.
a. هواپیماهایی با احتراق سوخت کامل
b. هواپیماهایی کم مصرف
c. هواپیماهایی کم سوخت
d. هواپیماهایی دارای ظرفیت زیاد سوخت

16- the deficit is an influence depressing effective demand.
a. فشار آوردن b. برابر کردن c. کم کردن d. کساد کردن

17- next product is the value of the increment of product expected from employing a man minus the additional expenses.

a. بهره تولید b. مازاد محصول c. افزایش محصول d. باقیمانده تولید

18- a tariff policy or import quotas, may effectively prevent certain commodities from entering the country.

a. سهمیه وارداتی b. تعرفه وارداتی c. سیاست واردات d. ضرورت واردات

References

Akbari, Alireza. 2020. Measuring the degree of difficulty and translation competence: The two intertwining variables in modern translation multiple-choice item testing. *Journal of Applied Research in Higher Education* 12(3): 475–494.

Akbari, Alireza. 2022. *Charted and uncharted territories in translation and interpreting research methods*. California, USA: Nova Science Publishers.

Akbari, Alireza & Shahnazari, Mohammadtaghi. 2025. Challenging the illusion of objectivity: An in-depth analysis of the Preselected Items Evaluation (PIE) method in translation evaluation. *Journal of Applied Research in Higher Education* 17(3): 1109–1124.

Akbari, Alireza & Shahnazari, Mohammadtaghi & Afrouz, Mahmoud. 2025. Accurate evaluation and consistent results: The case of the optimized version of the Preselected Items Evaluation method. *Onomázein* (forthcoming).

Andrich, David. 2004. Controversy and the Rasch model: A characteristic of incompatible paradigms? *Med Care* 42(1 Suppl): I7–16. doi: 10.1097/01.mlr.0000103528.48582.7c.

Arrindell, Willem A. & van der Ende, Jan. 1985. An empirical test of the utility of the observations-to-variables ratio in factor and components analysis. *Applied Psychological Measurement* 9:165–178.

Awopeju, O. & Afolabi, Eyitayo. 2016. Comparative Analysis of Classical Test Theory and Item Response Theory-Based Item Parameter Estimates of Senior School Certificate Mathematics Examination. *European Scientific Journal* 12 (8): 263–284. doi: 10.19044/esj.2016.v12n28p263.

Baghaei, Purya. 2008. Local dependency and Rasch measures. *Rasch Measurement Transactions* 21: 1105–1106.

Baker, Frank. B. 2001. *The basics of Item Response Theory*. New York: College Park.

Baker, Frank B. & Kim, Seok. Ho. 2004. *Item Response Theory: Parameter estimation techniques.* 2nd Edition. Boca Raton: CRC Press.

Bond, Trevor. G. & Fox, Christine M. 2007. *Applying the Rasch Model*. New York: Routledge.

Chen, Wen-Hung & Lenderking, William & Jin, Ying & Wyrwich, Kathleen. W. & Gelhorn, Heather & Revickim, Dennis. A. 2013. Is Rasch model analysis applicable in small sample size pilot studies for assessing item characteristics? An example using PROMIS pain behavior item bank data. *Qual Life Res* 23(2): 485–93. doi: 10.1007/s11136-013-0487-5.

Dabaghi, Sahar & Esmaielzadeh, Fatemeh & Rohani, Camelia. 2020. Application of Rasch analysis for development and psychometric properties of adolescents' quality of life instruments: A systematic review. *Adolescent Health, Medicine and Therapeutics* 11: 173–197.

Debelak, Rudolf & Koller, Ingrid. 2019. Testing the local independence assumption of the Rasch model with Q3-based nonparametric model tests. *Applied Psychological Measurement* 44(2): 103–117. doi: 10.1177/0146621619835501.

Fox, Jean-Paul. 2010. *Bayesian Item Response Modeling: Theory and applications*. Amsterdam: Springer.

Fu, Jianbin & Feng, Yuling. 2018. A comparison of score aggregation methods for unidimensional tests on different dimensions. *ETS Research Report Series* 2018 (1): 1–16. doi: 10.1002/ets2.12194.

Hambleton, Ronald K. & Russell, Jones, W. 1993. Comparison of classical test theory and item response theory and their applications to test development. *Educational Measurement: Issues and Practice* 12(3): 38–47. doi: 10.1111/j.1745-3992.1993.tb00543.x.

Hambleton, Ronald K. & Swaminathan, H. & Rogers, Jane. 1991. *Fundamentals of Item Response Theory*. New York: SAGE Publications.

Hattie, John A. 1985. Methodology review: Assessing unidimensionality of tests and items. *Applied Psychological Measurement* 9(2): 139–164. doi: 10.1177/014662168500900204.

Lasnier, François. 2000. *Réussir la Formation par Compétences*. Montreal, Canada: Guérin.

Lee, Yong-Won. 2004. Examining passage-related Local Item Dependence (LID) and measurement construct using Q3 statistics in an EFL reading comprehension test. *Language Testing* 21(1): 74–100.

Linacre, John Michael. 2002. What do infit and outfit, mean-square and standardized mean? *Rasch Measurement Transactions* 16: 878–879.

Linacre, John Michael. 2011. *A user's guide to Winsteps Rasch model computer programs*. Chicago, USA: Winsteps.

Lord, Frederic M. & Novick, Melvin R. 1968. *Statistical theories of mental test scores*. Reading, Mass: Addison-Wesley.

Magno, Carlo. 2009. Demonstrating the difference between Classical Test Theory and Item Response Theory using derived test data. *The International Journal of Educational and Psychological Assessment* 1(1): 1–11.

Novick, Melvin R. 1966. The axioms and principal results of classical test theory. *Journal of Mathematical Psychology* 3(1): 1–18. doi: https://doi.org/10.1016/0022-2496(66)90002-2.

OECD. 2009. *PISA data analysis manual: SPSS.* 2nd Edition: Paris: OECD Publishing.

Paek, Insu & Liang Xinya & Lin, Zhongtian. 2021. Regarding item parameter invariance for the Rasch and the 2-parameter logistic models: An investigation under finite non-representative sample calibrations. *Measurement: Interdisciplinary Research and Perspectives* 19(1): 39–54. doi: 10.1080/15366367.2020.1754703.

Park, Seohee & Reeger, Adam & Aloe, Ariel M. 2020. *Technically speaking: Determining test effectiveness with Item Response Theory*. (Available from https://iowareadingresearch.org/blog/technically-speaking-item-response-theory) (Accessed 2020-09-22).

Parmaningsih, Triwik Jatu & Sari Saputro, Dewi Rento. 2021. Rasch analysis on Item Response Theory: Review of model suitability. *AIP Conference Proceedings* 2326(1): 020017. doi: 10.1063/5.0040305.

Petrillo, Jennifer & Cano, Stefan J. & McLeod, Lori D. & Coon, Cheryl D. 2015. Using Classical Test Theory, Item Response Theory, and Rasch measurement theory to evaluate patient-reported outcome measures: A comparison of worked examples. *Value Health* 18(1): 25–34. doi: 10.1016/j.jval.2014.10.005.

Philip, Adewole & Odunayo, Ojo B. 2017. Application of Item Characteristic Curve (ICC) in the selection of test items. *British Journal of Education* 5(2): 21–41.

Reckase, M. D. 2009. *Multidimensional Item Response Theory*. New York: Springer.

Roussos, Louis A. & Stout, William F. & Marden, John I. 1998. Using new proximity measures with hierarchical cluster analysis to detect multidimensionality. *Journal of Educational Measurement* 35(1): 1–30.

Shun-Chin, Yang & Tsou, Mei-Yung & Chen, En-Tzu & Chan, Kwok-Hon & Chang, Kuang-Yi. 2011. Statistical item analysis of the examination in anesthesiology for medical students using the Rasch model. *Journal of the Chinese Medical Association* 74(3): 125–129.

Smith, Adam B. & Rush, Robert & Fallowfield, Lesley J. & Velikova, Galina & Sharpe, Michael. 2008. Rasch fit statistics and sample size considerations for polytomous data. *BMC Medical Research Methodology* 8(1): 33. doi: 10.1186/1471-2288-8-33.

Smith, Richard M. 1996. A comparison of methods for determining dimensionality in Rasch measurement. *Structural Equation Modeling: A Multidisciplinary Journal* 3(1): 25–40. doi: 10.1080/10705519609540027.

Statistical Consulting. 2021. *How do I interpret odds ratios in logistic regression?* (Available from https://stats.idre.ucla.edu/stata/faq/how-do-i-interpret-odds-ratios-in-logistic-regression/) (Accessed 2021-08-20)

Tate, Richard. 2003. A comparison of selected empirical methods for assessing the structure of responses to test items. *Applied Psychological Measurement* 27(3): 159–203. doi: 10.1177/0146621603027003001.

Tavakol, Mohsen & Dennick, Reg. 2013. Psychometric evaluation of a knowledge-based examination using Rasch analysis: An illustrative guide: AMEE Guide 72. *Medical Teacher* 35(1): e838–e848. doi: 10.3109/0142159X.2012.737488.

Tennant, A. & Hillman, M. & Fear, J. & Pickering, A. & Chamberlain, M. A. 1996. Are we making the most of the Stanford health assessment questionnaire? *Rheumatology* 35(6): 574–578. doi: 10.1093/rheumatology/35.6.574.

Wang, Wen-Chung & Wilson, Mark. 2005. Exploring local item dependence using a random-effects facet model. *Applied Psychological Measurement* 29(4): 296–318. doi: 10.1177/0146621605276281.

Waterbury, Glenn T. 2020. *RISE and shine: A comparison of item fit statistics for the Rasch model*. Virginia, USA: James Madison University. (Doctoral dissertation).

Wright, Benjamin D. & Linacre, John. M. 1994. Reasonable mean-square fit values. *Rasch Measurement Transactions* 8: 370–371.

Wright, Benjamin D. & Stone, Mark. H. 1979. *Best test design*. Chicago, IL: MESA Press.

Yu, Chong Ho & Osborn Popp, Sharon & DiGangi, Samuel & Jannasch-Pennell, Angel. 2007. Assessing unidimensionality: A comparison of Rasch Modeling, Parallel Analysis, and TETRAD. *Practical Assessment, Research, and Evaluation* 12(14): 1–19.

Zhang, Jinming & Stout, William. 1999. The theoretical detect index of dimensionality and its application to approximate simple structure. *Psychometrika* 64(2): 213–249. doi: 10.1007/BF02294536.

Alireza Akbari, Ph.D.
Faculty of Foreign Languages
University of Isfahan, Iran
E-mail: dictogloss@gmail.com/ alireza.akbari@fgn.ui.ac.ir