International Electronic Journal of Mathematics Education
2024, 19(1), em0761
e-ISSN: 1306-3030
https://www.iejme.com

MODESTUM

Research Article

OPEN ACCESS

Measuring students' conceptual understanding of real functions: A Rasch model analysis

Anela Hrnjičić [1]*, Adis Alihodžić [1]

[1] Department of Mathematics, Faculty of Science, University of Sarajevo, Sarajevo, BOSNIA & HERZEGOVINA
***Corresponding Author:** anela.hrnjicic@gmail.com

Citation: Hrnjičić, A., & Alihodžić, A. (2024). Measuring students' conceptual understanding of real functions: A Rasch model analysis. *International Electronic Journal of Mathematics Education, 19*(1), em0761. https://doi.org/10.29333/iejme/13942

ARTICLE INFO

Received: 21 Aug. 2023

Accepted: 21 Nov. 2023

ABSTRACT

Understanding the concepts related to real function is essential in learning mathematics. To determine how students understand these concepts, it is necessary to have an appropriate measurement tool. In this paper, we have created a web application using 32 items from conceptual understanding of real functions (CURF) item bank. We conducted a psychometric analysis using Rasch model on 207 first-year students. The analysis showed that CURF is a dependable and valid instrument for measuring students' CURF. The test is uni-dimensional; all items are consistent with the construct and have excellent item fit statistics. The results indicate that the items are independent of each other and unbiased towards the gender and high school background of the students.

Keywords: conceptual understanding, real function, mathematics assessment, Rasch model, multiple representations

INTRODUCTION

Mathematics teaching requires frequent assessment of students' conceptual understanding in order to offer them a timely response to overcome revealed misconceptions and provide the opportunity to understand new concepts related to those previously adopted. This is very important when students encounter the concept of a real function, which cannot be considered as a separate concept (connected to sub-concepts such as domain, extremum, flow, etc.), as it connects algebra and geometry and can be represented in various ways (Dreyfus & Eisenberg, 1982). Understanding one sub-concept of a function helps students to understand a new concept. Students' ability to represent function concepts using multiple contexts is of great importance for the cognitive development of a student's conceptual knowledge of functions (Bell, 2001). It is helpful to ask questions that require transitioning from one representation to another to assess students' understanding of a concept. A solid understanding of a concept, like function, requires recognizing and smoothly transitioning between different representations. According to Hitt (1998), this is an effective way to evaluate students' conceptual understanding.

The creation of an item bank that encompasses a multitude of items functioning together requires several steps. In a recent study, Hrnjičić et al. (2022) detailed the development of a real function item bank known as conceptual understanding of real functions (CURF). CURF comprises different kinds of functions, concepts of function zero, sign of a function, function evenness/oddness, the limit and asymptotes of a function, the maximum/minimum, and increasing and decreasing of a function. The item bank measures conceptual understandings with items that require the capability to shift from one representation of the same concept to another, including verbal, algebraic, and graphical representations of content.

After creating an item bank, we must regularly assess if the items still effectively measure the given construct (Glamočić et al., 2021). Since the content and cognitive validity of the items were already established in a previous study by Hrnjičić et al. (2022), our goal for this study is to use Rasch model to determine construct validity, evaluate other psychometric properties of the item bank, and ensure its ability to measure the comprehension of real functions among first-year university students. In order to examine a larger sample of students in a more effective way, we created a computer-assisted assessment (CAA) as a web application that contains all items from CURF item bank and applied it on a similar sample of respondents as in the research by Hrnjičić et al. (2022). We have decided to use Rasch model because it helps us ensure the quality of our item bank (Boone et al., 2013; Wolfe, 2000). It also enables us to evaluate whether the items in our bank work well together to measure a single construct and to inspect the three basic assumptions of Rasch model: item fit, uni-dimensionality, and local independence. Additionally, we

This study is a continuation of the recently published research by the authors, which is available at: https://doi.org/10.30935/scimath/12222 can check for any bias in the items towards the gender of the students or the type of high school completed. This study and paper by Hrnjičić et al. (2022) will provide a systematic overview of developing an item bank using Rasch framework.

THEORETICAL BACKGROUND

Real functions of one real variable are an indispensable part of mathematics curriculum for primary school, high school and university. Students encounter difficulties in understanding concepts covered by CURF item bank (Hrnjičić et al., 2022). In order to have an opportunity to check possible misconception at any moment it is necessary to have a quality assessment that would offer us reliable and trustworthy evaluations. Objectivity, reliability, and validity of measurement are crucial in determining the quality of assessments (Glamočić et al., 2021). When we talk about objectivity, we are referring to how much the results of measurements depend on the person who administers the test and scores the answers given by students (Marcus & Bühner, 2009). In order to fulfil the requirement of objectivity, we developed an online CAA. CAA is suitable for computer marking because the correct answers are defined in advance so there is no subjectivity in marking. Bull and McKenna (2004) defined a CAA as "the use of computers to assess students". The purpose of CAA is to deliver, mark and analyze assignments or examinations (Bull & McKenna, 2004). Previous studies (Ashton et al., 2003; Nugent, 2003; Sealey et al., 2003) have demonstrated that CAA is an effective method for evaluating students and delivering timely feedback on individual and class progress. Test validation involves gathering evidence to determine if the test items accurately represent the concept domain and if the test measures the proposed properties (Day & Bonn, 2011). Reliability directs to the viscosity of a test within itself and over time, and it is determined through statistical calculations that examine both individual items and the test as a whole (Day & Bonn, 2011).

CURF item bank consisted of 32 items and was invented by Hrnjičić et al. (2022). The content validity of the item bank was verified through an expert survey, while the cognitive validity, wording, and clarity of items were determined using student surveys. Descriptive statistics within the framework of classical test theory (CTT) were utilized to evaluate both the field-tested items and the test as a whole. The item and test analyses showed that all 32 items possess good psychometric characteristics and can be integrated into a scale that reliably measures students' CURF in university introductory mathematics courses (Hrnjičić et al., 2022). We conducted a second pilot test on the 32 items using Rasch model to validate this further.

A Rasch Model Approach

Tests can be designed using either CTT or probabilistic test theory (Mešić et al., 2019). CTT assumes that a person's test score comprises their "true" score and a measurement error. In contrast, probabilistic test theory allows statements about the outcome as probabilities for observable variables (Schmid, 1992). In CTT theory, we measure the knowledge dimension by adding up the test scores on all items. In item response theory (IRT), the unobservable trait is known as the latent trait (θ), and we cannot measure it directly (Bond & Fox, 2015). Rasch model is considered the simplest probabilistic test model (Bond & Fox, 2015). Rasch (1960) devised a probabilistic model for specific intelligence and achievement tests. According to this model, a person with a higher capacity than another person should have a greater possibility of solving any item of the same type. Also, if one item is more complex than another, the probability of solving the second item is more significant for any person. The assumptions of Rasch model include representing each person through its ability (B_n), each item through its difficulty (D_i), which in the end can be demonstrated by numbers along one line (Bond & Fox, 2015). The likelihood of someone succeeding on a test item is established by their ability (the number of correct answers) compared to the difficulty level of the item (the number of people who have succeeded). Different models for testing probabilities exist, and they vary based on the shape of the item's characteristic function. This function models the relationship between a person's responses to test items and their ability level (Bond & Fox, 2015; Hambleton et al., 1991.) According to Hambleton et al. (1991), Rasch model's item characteristic function can be expressed, as in Eq. (1):

$$P_{ni} = \frac{e^{(B_n - D_i)}}{1 + e^{(B_n - D_i)}}, \tag{1}$$

where P_{ni} represents the probability that an examinee with ability B_n will answer the item i correctly when selected randomly. The computer yields a log-transformed display of estimates for a person's ability and item difficulty, presented on a logit scale (logarithmic transformation of the odds or probability of success). Every estimate comes with an associated margin of error, as explained by Bond and Fox (2015).

Rasch model is a construct validation instrument (Fisher Jr, 1994). Cronbach and Meehl (1955) introduced the term "construct validity", and according to Overton (1999), it represents the extent to which a test (or score) can be used to estimate a theoretical construct or trait. Rasch model implies that observed behaviors are indicators of an underlying construct (Bond & Fox, 2015). The model also allows for estimating reliability for both individuals and items. The item reliability index assesses the reproducibility of item placement along a given path, assuming the same items are given to another group of individuals with similar characteristics. The item separation index indicates the distribution or separation of items on the measured variable. Likewise, the person reliability index measures the repeatability of a person ordering on a parallel set of items measuring the same construct. The person separation reliability estimates how well individuals can be distinguished based on the measured variable. In this article, Rasch model's three core assumptions - item fit, uni-dimensionality, and local independence - are involved in analyzing the construct validity and reliability of CURF item bank. Uni-dimensionality means that only one latent trait is present in item responses, while local independence requires that test items should not be related. Local dependency may be evident when a participant's item responses depend on their answers to other test items. Item fit statistics are divided into outfit statistics, which focus on unexpected responses away from the item's or person's measurement, and infit statistics, which focus on unexpected responses close to the measurement (Bond & Fox, 2015).

METHODOLOGY

Research Design

Our research design follows the steps outlined by Liu (2010). It includes testing a representative sample of the target population, conducting Rasch modelling, reviewing item fit statistics and the Wright map, and establishing validity and reliability claims for the measurement instrument. For the purpose of this research, we created an online accessible CAA, which contains all 32 items, presented in the same order as in the original CURF bank. The reason why we decided on the online CAA available via the link is because of its advantages such as: rapid scoring and feedback, originality, flexibility, interactivity, innovation, cost reduction and operational efficiency in terms of paper use and scoring services, item-banking, and consistency with item response theory (Brown, 2013).

Participants

At the beginning of the study, 207 first-year college students enrolled in introductory mathematics courses participated in the research. Convenience sampling was used for obtaining the student sample, based on the availability of access to state universities. We obtained informed consent from all participants who were tested at the beginning of the first semester, before they attended any classes on the concepts covered in CURF. The only inclusion criterion was that the students studied these concepts in the final year of high school, and the similarity of the mathematics curriculum in high school. At the beginning of the analysis, we excluded 40 students whose answer to the question: "*have you studied about real functions in high school?*" was no, as well as seven students who did not study these concepts according to the same curriculum as the other students. Thus, our final sample consisted of 160 freshmen, namely: 92 students from the Faculty of Traffic and Communication of the University of Sarajevo, 39 from the Faculty of Science of the University of Sarajevo and 29 students from the Faculty of Mechanical Engineering of the University of Zenica. 64 students were female and 96 were male. 70 students graduated from grammar school and 90 from technical high school.

Instrument

CURF is a multiple-choice test with a dichotomous scoring system: one point for a valid answer and zero point for a wrong one. Each question has four possible answers, one of which is correct, and the other three are distractors. CURF CAA test can be carried out online without a time limit. The test consists of 32 questions spread across four cards. Card A contains questions that require a transition from graphic to verbal representation, card B contains questions that demand a transition from graphic to algebraic representation, card C contains questions that need a transition from algebraic to graphic representation, and card D contains questions that direct a transition from algebraic to verbal representation. Each concept mentioned has four questions that measure the ability to move from one representation to another of the same concept. Items A1, B1, C1, and D1 require the ability to recognize the kind of a function (e.g., linear, quadratic, etc.). Items A2, B2, C2, and D2 refer to the concept of zero of a function. Items A3, B3, C3, and D3 relate to even functions, while A4, B4, C4, and D4 relate to odd functions. Items A5, B5, C5, and D5 examine the ability to determine in which intervals the function is positive and in which it is negative, while items A6, B6, C6, and D6 examine understanding of the limit and asymptotes of a function. Items A7, B7, C7, and D7 require the ability to recognize in which intervals the function increases/decreases, and items A8, B8, C8, and D8 refer to the concept of maximum/minimum of a function. At the end of the test, students are shown a report on how many items they have answered correctly, as well as the percentage of correct answers. You can find question cards in English by visiting the link: https://sites.google.com/view/supplementalmaterialaphd1/.

Procedure

Before testing the students, we organized meetings with faculty administration and professors who give lectures on introductory mathematics courses. On this occasion we asked them for their cooperation and introduced them to the details of conducting the test. Each faculty had a specific date and time for the test. At the beginning of the test, the subject professor introduced the students to the test procedure and forwarded the link for accessing the test. In order to further motivate them, the professors decided to evaluate the test work as a student activity during the semester.

Data Analysis

We utilized software WINSTEPS version 3.72.3 for our Rasch analysis, following Szabó's (2008) suggestion to check for person misfit by examining negative point-biserial score correlations and person infit/outfit. Positive point-biserial correlation statistics indicated that all items were functioning as predicted, while negative or close-to-zero values suggested inconsistencies with the construct. We also used mean square residual (MNSQ) and standardized mean square residual (ZSTD) statistics to assess item fit. After addressing these factors, we checked for multidimensionality using principal component analysis (PCA) of the residuals. We calculated reliability and separation indices for items and persons and verified local item independence by examining standardized residual correlations. We also utilized the Wright map to explore the relationship between item difficulty and a person's ability and investigated differential item functioning (DIF) concerning gender and high school type. Overall, our quality measure research aimed to ensure our analysis's fairness, accuracy, and consistency.

Table 1. Estimations of parameters for CURF item bank

Item	Measure	Standard error	Infit		Outfit		PT-Measure		Exact match	
			MNSQ	ZSTD	MNSQ	ZSTD	Corr.	Exp.	Obs (%)	Exp (%)
C6	2.03	0.22	1.18	1.4	1.61	2.6	0.29	0.47	83.1	81.3
B6	1.76	0.21	1.03	0.3	1.05	0.3	0.44	0.47	82.4	78.2
B7	1.55	0.20	1.04	0.4	1.09	0.6	0.43	0.46	77.7	75.8
B4	1.35	0.20	1.11	1.2	1.13	1.0	0.39	0.46	67.6	74.0
A4	1.20	0.19	1.18	2.1	1.11	0.9	0.34	0.46	65.5	72.7
C7	1.02	0.19	0.99	-0.1	0.94	-0.5	0.46	0.45	69.6	71.1
A3	0.91	0.19	1.16	2.1	1.15	1.4	0.34	0.45	60.8	70.3
D7	0.78	0.18	1.01	0.2	0.99	-0.1	0.43	0.44	73.0	69.4
C4	0.58	0.18	1.03	0.5	0.96	-0.3	0.42	0.43	62.8	68.3
C2	0.45	0.18	1.11	1.6	1.18	1.7	0.34	0.43	62.8	67.9
D8	0.38	0.18	0.98	-0.2	1.15	1.4	0.42	0.42	68.2	67.8
D4	0.18	0.18	0.82	-2.9	0.76	-2.4	0.55	0.41	77.7	67.6
A2	0.15	0.18	1.04	0.7	1.01	0.1	0.38	0.41	66.2	67.6
A6	0.12	0.18	1.15	2.1	1.34	2.8	0.28	0.41	65.5	67.7
B8	0.12	0.18	1.06	1.0	1.03	0.3	0.37	0.41	64.2	67.7
C3	0.08	0.18	1.21	3.0	1.20	1.7	0.26	0.41	54.1	67.8
D5	-0.19	0.19	0.86	-2.1	0.77	-1.9	0.50	0.39	74.3	69.2
D6	-0.22	0.19	0.92	-1.1	0.85	-1.2	0.45	0.39	64.2	69.4
D2	-0.29	0.19	0.83	-2.5	0.72	-2.2	0.52	0.38	76.4	69.9
D3	-0.29	0.19	0.85	-2.2	0.92	-0.5	0.49	0.38	75.0	69.9
B3	-0.36	0.19	1.01	0.1	0.92	-0.5	0.38	0.38	68.9	70.5
C5	-0.40	0.19	0.93	-1.0	0.85	-1.0	0.43	0.37	73.6	70.8
B1	-0.47	0.19	0.90	-1.3	0.82	-1.2	0.45	0.37	74.3	71.4
C8	-0.47	0.19	0.94	-0.8	0.87	-0.8	0.42	0.37	73.0	71.4
B5	-0.51	0.19	0.90	-1.2	0.79	-1.4	0.45	0.36	75.0	71.8
B2	-0.73	0.20	1.04	0.5	0.91	-0.5	0.34	0.35	73.0	74.4
A7	-0.89	0.20	1.00	0.1	0.94	-0.2	0.33	0.33	77.0	76.4
A8	-1.11	0.21	1.11	1.0	1.05	0.3	0.24	0.31	75.7	79.1
C1	-1.30	0.22	0.85	-1.2	0.73	-1.1	0.42	0.30	83.1	81.4
D1	-1.57	0.24	1.08	0.6	0.90	-0.2	0.24	0.27	83.8	84.5
A1	-1.75	0.25	0.94	-0.3	0.67	-1.0	0.33	0.26	86.5	86.5
A5	-2.10	0.28	0.85	-0.7	0.55	-1.3	0.36	0.23	89.9	89.8
Mean	0.00	0.20	1.00	0.0	0.97	-0.1			72.7	73.2
SD	0.99	0.02	0.11	1.4	0.21	1.3			8.0	5.9

Note. Measure: Parameter estimation; PT-Measure (Corr. & Exp.): Point-biserial correlation (actual observations & predictions by model); & Exact match: Percentage of points that conform to prediction & percentage that model predicts

RESULTS

We conducted a Winsteps analysis on the responses of 160 students to 32 items, first checking for person misfits. Six individuals with negative point-biserial correlations were identified and removed from our dataset. We re-ran Winsteps on a sample of 154 students and 32 items. We removed another person due to a negative correlation. After that we ran Winsteps on a sample of 153 students and 32 items and the correlations for all items and all persons were positive. According to Linacre (2022), MNSQ of fit statistics is considered productive for measurement if its values are in the interval 0.50-1.50. The acceptable range for infit MNSQ values for individuals is between 0.50 and 1.50, with a more specific range of 0.58 to 1.38. However, two individuals had an outfit MNSQ value greater than 1.50, with minimal deviation from the ideal range at 1.66 and 1.62. Their standardized infit or outfit statistics, measured as ZSTD, were outside the acceptable range of -2.00 to 2.00 (Bond & Fox, 2015), with values greater than 2.00. We deleted these two persons from the database. We ran Winsteps again for 151 students and 32 items (**Table 1**) and concluded that all persons had infit MNSQ below 1.50, and only one person had outfit indices slightly above 1.50, more precisely an outfit MNSQ of 1.52. Since this person had good fit indices and positive correlations, we kept it in the database.

Our analysis discovered that none of the items had negative correlations with point measurements. The values for point-measure correlation statistics are ranging from 0.24 to 0.55. It is necessary to examine item fit statistics because misfitting items could display additional dimensions (Liu, 2010). According to Linacre (2019), a good fit between the model and data is indicated when infit and outfit MNSQ values fall within the range of 0.50 to 1.50. In this study, all 32 items had infit statistics within this range, ranging from 0.82 to 1.18. **Figure 1** visually displays MNSQ infit index for item difficulty measures, with the size of each item's bubble reflecting the standard error of its difficulty estimate. Outfit statistics were also within the ideal range for 31 items, except for item C6, whose outfit MNSQ value of 1.61 indicated a slight deviation.

The estimates of item difficulty in logits range from -2.10 to 2.03. Person abilities are in the range from -1.74 to 3.87, while three persons have a value of 5.12 and they answered everything correctly. The mean measure for items in logit is 0.00, and standard deviation (SD) is 0.99. The value of the estimated mean of person ability of 0.54 in comparison of the values of the estimated mean of item 0.00 shows that the test was easier for this sample. The person reliability is at 0.82, and the item reliability is at 0.96.

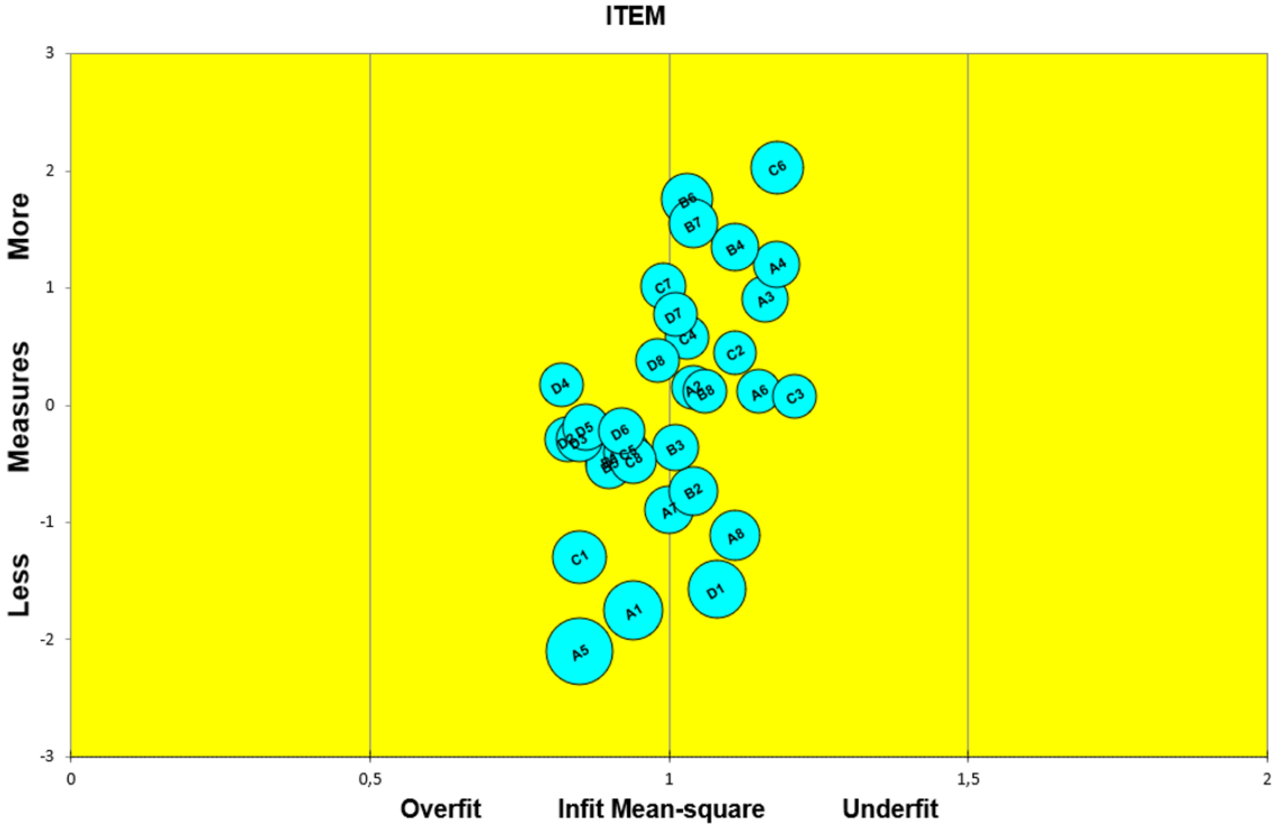


Figure 1. Dispersal of items due to difficulty & infit MNSQ (Source: Authors' own elaboration)

According to Oon et al. (2016), if the reliability index is close to 1 it shows internal consistency. The item reliability value of 0.96 suggests that the construct is estimated consistently.

In order to check the representative aspect of validity, the item strata separation is given–for the test to be accepted as representative, there should be at least two levels (Smith Jr, 2001). For persons, the separation index is 2.15 (satisfactory), and for items it is 4.83, showing a wider range of item difficulty. Based on Rasch model algorithm, the unstandardized fit estimates (also known as mean squares) are designed to have an average of one (Bond & Fox, 2015).

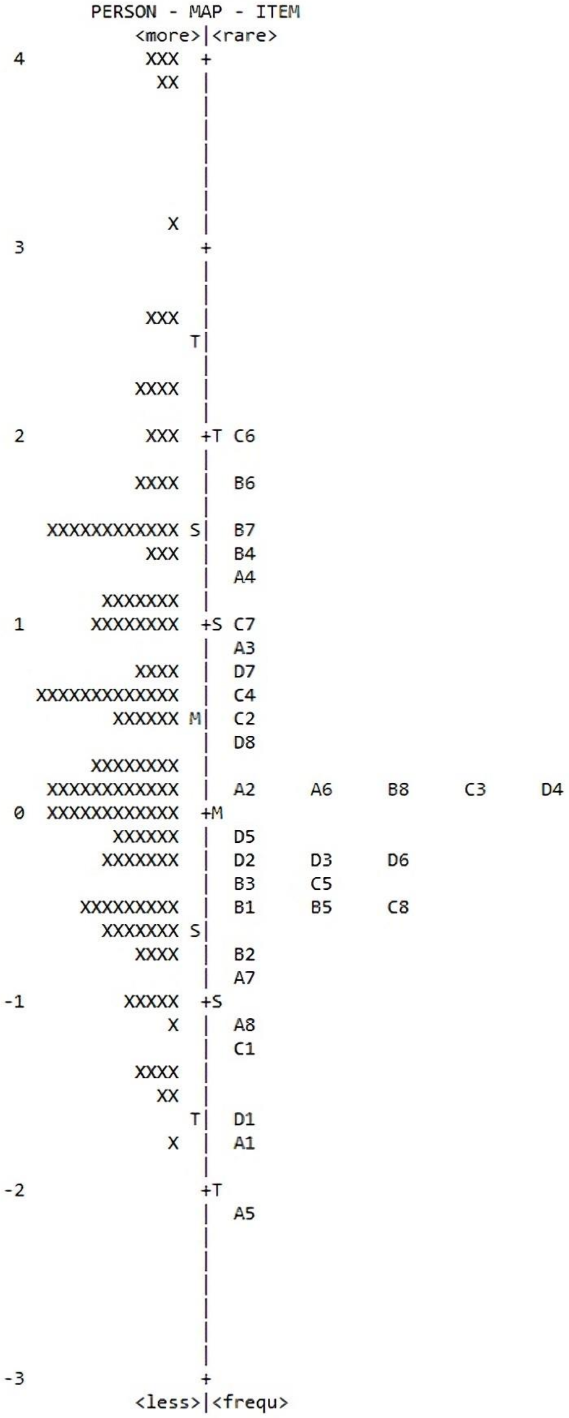
In our study, the unstandardized item fit statistics for CURF item bank have means that are very close to the anticipated one (infit mean squares=1.0 & outfit mean squares=0.97), with little deviation from the ideal for both infit and outfit mean squares (infit SD=0.11 & outfit SD=0.21). Additionally, the values for the average fit and SDs of individuals are acceptable (infit=1.00, SD=0.18, outfit=0.97, & SD=0.28).

After examining negative point-biserial and item fit statistics, Linacre (1998) suggests checking for multidimensionality. We used PCA of residuals in Winsteps to check for this. The measure's raw variance explained is 26.9%, with 10.6% explained by persons and 16.3% by items. If the data fit the model perfectly, 27.3% would have been explained. The difference between the observed variance and that expected by the model was minimal at only 0.4%. The variance of the residuals represents the unexplained variance in a data set, while the variance of the persons and items defines the explained variance, as per Bond and Fox (2015). In our paper SD value of person ability is 1.23, while SD value of item difficulty is 0.99.

According to the nomogram (Linacre, 2008), the expected "variance explained by Rasch measures" is consistent with our data and results. Uni-dimension can be judged by using the eigenvalue in Winsteps. The critical factor to consider in the process is the amount of variance, also known as eigenvalue, present in the first contrast, as stated by Bond and Fox (2015). If the first contrast has less than three units or eigenvalue, we can conclude that the test is unidimensional, as suggested by Linacre (2022, p. 647). Our paper shows that the first contrast's units or eigenvalue is less than three, specifically 2.8 items strong.

We investigated local dependency by examining the most significant standardized residual correlations. As per Linacre (2017), a cause for concern may be the positive between-item correlations greater than 0.7. In our case, all standardized residual correlations are less than 0.70 (the highest correlation is 0.33 for A3 and A4 items). In order to find out more about the item person relationships in relation to the underlying variable, we examine the Wright map (**Figure 2**), where we can see the person distribution against the item distribution.

The map variable contains items on the right side, and the left is occupied by *X*s representing students. Each *X* indicates two students on the map. The bottom portion of the map comprises the most straightforward items and the less capable students, while the top portion contains the most challenging items and the most capable students (Linacre, 2022). According to Bond and Fox (2015), Rasch model typically sets a 50.0% chance of success for individuals on an item positioned at the same point on the item-person logit scale.



```
        PERSON - MAP - ITEM
            <more>|<rare>
 4           XXX  +
              XX  |
                  |
                  |
                  |
                  |
                  |
               X  |
 3                +
                  |
                  |
             XXX  |
                 T|
                  |
            XXXX  |
                  |
 2           XXX  +T C6
                  |
            XXXX  |  B6
                  |
    XXXXXXXXXXXX S|  B7
             XXX  |  B4
                  |  A4
         XXXXXXX  |
 1      XXXXXXXX  +S C7
                  |  A3
            XXXX  |  D7
   XXXXXXXXXXXXX  |  C4
          XXXXXX M|  C2
                  |  D8
        XXXXXXXX  |
    XXXXXXXXXXXX  |  A2     A6     B8     C3     D4
 0  XXXXXXXXXXXX  +M
          XXXXXX  |  D5
         XXXXXXX  |  D2     D3     D6
                  |  B3     C5
       XXXXXXXXX  |  B1     B5     C8
         XXXXXXX S|
            XXXX  |  B2
                  |  A7
-1         XXXXX  +S
               X  |  A8
                  |  C1
            XXXX  |
              XX  |
                 T|  D1
               X  |  A1
                  |
-2                +T
                  |  A5
                  |
                  |
                  |
                  |
                  |
                  |
-3                +
            <less>|<frequ>
```

Figure 2. Item-person map for CURF test (Winsteps) (Source: Authors' own elaboration)

Test information function is a technique used in a Rasch analysis to evaluate how a test is functioning (Boone & Staver, 2020). It provides us with additional insight into person-test targeting (Mešić et al., 2019). Test information is a concept that shows the amount of information each item provides as it relates to any person parameter (Bond & Fox, 2015). Well-targeted persons provide more information (and less error) as opposed to poorly targeted persons (Bond & Fox, 2015) (**Figure 3**).

To ensure fairness, we conducted a DIF analysis to check for any bias in our items. Specifically, we tested whether our items were similar for boys and girls using Mantel-Haenszel procedure–a widely-used method for analyzing dichotomous items (**Figure 4**).

Our results showed that, on average, boys had a measure of 0.52 logits while girls had 0.56 logits, indicating that girls were slightly more successful. It is important to note that extreme scores from both persons and items were excluded from our analysis as they do not exhibit differential ability across items. Then we inspected whether some of $|DIF\ Contrast| \geq 0.64$ or statistically significant ($Mentel - Haenszel\ Prob\ \leq 0.05$) (Linacre, 2017). Items with a $|DIF\ Contrast| \geq 0.64$, which are statistically significant are: A8 and D1 and they are harder for boys and easier for girls. This is in line with the conclusion that girls find the test to be slightly easier than the boys.

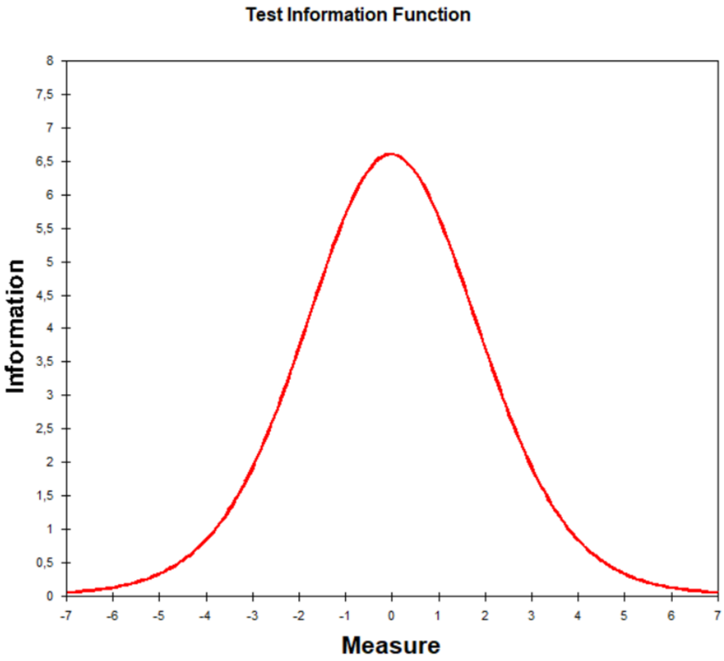


Figure 3. Test information function (Source: Authors' own elaboration)

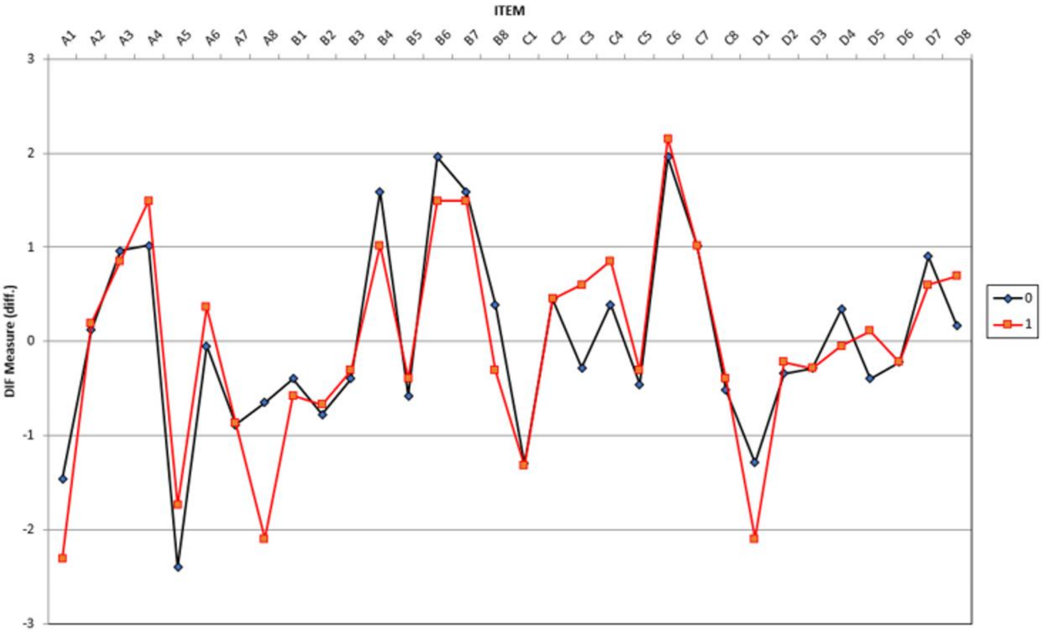


Figure 4. DIF values by gender (0: Male & 1: Female) (Source: Authors' own elaboration)

In **Figure 5**, we conducted a uniform DIF analysis based on the completed high school type. The majority of the items did not demonstrate DIF. However, A4 and B8 show statistically significant DIF, $p \leq 0.05$. A4 is more difficult for grammar school, while B8 is more difficult for technical school.

DISCUSSION

According to an investigation by Hrnjičić et al. (2022), CURF item bank underwent pilot testing and was analyzed using descriptive statistics within the framework of CTT. The study marked the first time Rasch dichotomous model was used to calibrate CURF item bank. Rasch measurement model is considered superior for construct validation of measures based on CTT, as per Crocker and Algina (1986). All items showed positive point-measure correlations, confirming that they operate in the planned direction and exist in the same direction as the latent variable. Upon analyzing the overall fit of the data to the proposed model, we found that all 32 conceptual items had excellent item fit statistics, so there was no need to remove any from the analysis. The only item slightly out of fit was item C6, based on the outfit MNSQ statistic. However, since it is a minimal outfit that would not distort or degrade the measure or the construct, it would not be necessary to eliminate this item from the test.

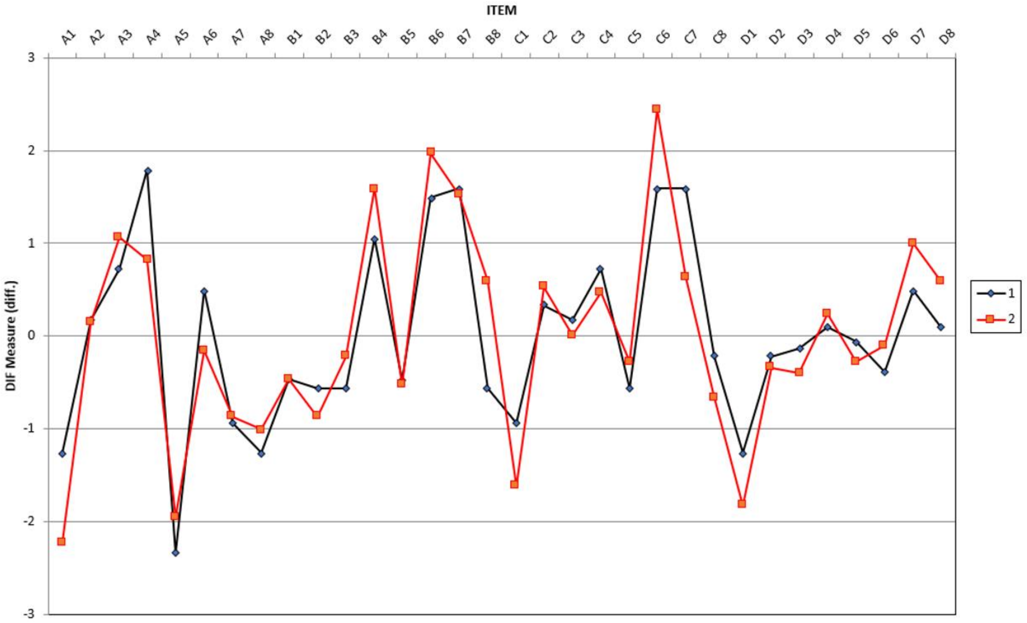


Figure 5. DIF values according to completed high school (1: Grammar school & 2: Technical school) (Source: Authors' own elaboration)

In order to obtain information about how much was measured and how well it was measured, we created the pathway map (**Figure 1**), which provides the following conclusion: most of the 32 items are found on the pathway, so the quality-control aspect of item performance may be considered satisfactory (Bond & Fox, 2015). The item reliability index of 0.96 indicates that we can rely on the order of item estimates to be replicated in other suitable samples using CURF. Person reliability is at satisfactory level. Five levels of item performance are identified by item strata of 4.76, while person strata of 2.15 indicates that we can identify two different groups of respondents. All of the values met the requirements for accurate measurement. The unstandardized item and person infit statistics mean value is one, while the unstandardized item and person outfit statistics mean value is very close to the expected one (more precisely 0.97).

Our uni-dimensionality examination found that the measure accounted for 26.9% of the raw variance. As per Sumintono and Widhiarso's (2015) guidelines, any value over 20.0% is acceptable. They also recommend that the 'unexplained variance' be at most 15.0% for an ideal result. In each of the five contrasts the unexplained variance was less than 10.0% and we can consider it as an ideal value (Sumintono & Widhiarso, 2015). As the first eigenvalue in raw variance unexplained by Rasch model was smaller than three this instrument is unidimensional.

When analyzing the positions of items on the Wright map's latent variable continuum (**Figure 2**), it becomes apparent that most individuals are positioned opposite the items. It implies that the items are well-suited for these individuals. If there are gaps of more than 0.50 logits in the distribution of item difficulties, new items are necessary, according to Linacre (2022). The biggest gap is evident between A5 and A1 and has a value of 0.35. However, there is a group of person locations (26 persons) above the hardest item (C6), in the region above 2.03 logits, which indicates that our test cannot give us accurate information about students with extraordinary ability. Our outcomes show that our test is more effective for students with lower abilities than those who excel on the Wright map. The questions located in the lower section of the Wright map focus primarily on comprehension of the sign of a function concept (A5: determining where the function is positive/negative based on a given image), identification of functions (A1, C1, & D1: exponential and quadratic functions), and recognition of local minimum and maximum points (A8). The questions located at the top of the map assess understanding of odd functions (A4 & B4), function limits (B6 & C6: connecting function values with appropriate graphs), and function flow (B7: describing the algebraic growth/decrease of a function based on a given graph).

Established on the test information function, we can determine that the most accurate measurements are obtained at approximately zero logits. The least precise measurements are found at the low and high ends of the latent trait continuum. It is vital to note that the test information function is symmetrical when the ability level is at 0.00. The highest amount of test information is approximately 6.60, with a standard error of estimate of 0.39. The most minor standard error of measurement is located at the peak of the test information function. Persons whose ability corresponds to the peak of CURF are measured most precisely.

We discovered that our dataset met the assumption of local item independence. None of the pairs of items had a standardized residual correlation greater than 0.70. The highest correlation was found between items "A3" and "A4", with a value of 0.33. It is worth noting that these items both relate to similar concerns. In item A3, the task was to recognize the graph of an even function on the basis of four offered graphs, and in item A4, the task was to recognize the graph of the odd function on the set of offered graphs. A Rasch calibration was accomplished on 151 students. When comparing male and female students, the analysis showed no statistically significant DIF for most items, indicating test invariance. Only in two out of 32 items DIF contrast was more noticeable than 0.64 logits. Also, in comparing DIF of students based on completed high school, only two items had statistically significant DIF.

We encountered that the students accomplished best on card A, which required a transition from graphics to verbal. This leads to the conclusion that they are able to describe the graph in words. They had the worst result on card B (from graphic to algebraic), with a similar result on card C (from algebraic to graphic). This result might be possible due to the lack of description in words of the given concept. Card D enables them to transition between algebraic and verbal representations. It is clear to them, from the set task, what they need to conclude. The majority of students (62.0%) completed the test successfully.

Limitations of the Research

The first limitation of the study is that these 32 items do not provide much information about very successful students in the region above 2.03 logits (17.0% of students who are above the hardest item of the Wright map). Another limitation relates to the sample size when it comes to DIF analysis, which requires a larger sample and a larger number of respondents in the groups we compare.

CONCLUSIONS

The content domain covered introductory mathematics at the university and high school level. We wanted to investigate further and verify the psychometric characteristics of CURF item bank to offer a trustworthy and proper instrument for estimating the understanding of concepts related to real functions. For the purposes of this research, we developed a web application that was used to assess students through a set of well-prepared and in advance designed items. The validity of CURF item bank was evaluated through a study conducted by Hrnjičić et al. (2022) using Rasch model. The study confirmed that Rasch model assumptions were met, and it was determined that 32 items could be combined to create a reliable scale for measuring students' CURF. We discovered that our 32 items measured the construct in the same direction, and excellent item fit statistics characterized them. In general, our set of items fulfils the requirements for estimating students' CURF. Items are locally independent, and most items are unbiased in relation to the respondent's gender and completed high school.

Our item bank can be utilized to measure high school students' CURF, as it covers situations faced by grammar and technical school graduates. The outcomes of this paper are important because if a math teacher wants to discover if a student understands, for example, the concept of function's zeros (or any of the concepts included), they can select four items from CURF item bank (in this case, these are the items A2, B2, C2, and D2) and check whether students can move from one representation to another of the same concept without falling into contradictions.

Author contributions: Both authors have sufficiently contributed to the study and agreed with the results and conclusions.

Funding: No funding source is reported for this study.

Acknowledgements: The authors would like to thank all students & professors who took part in this study.

Ethical statement: The authors stated that, in accordance with local legislation and the requirements of state universities, ethical review approval was not required, for this study. The authors briefed the deans of the faculties where the research was conducted on the details of the study. Written informed consents were obtained from the participants before the research was conducted. All participants were adults and informed about the research, anonymity and their right to withdraw consent without adverse repercussions.

Declaration of interest: No conflict of interest is declared by authors.

Data sharing statement: Data supporting the findings and conclusions are available upon request from the corresponding author.

REFERENCES

Ashton, H. S., Schofield, D. K., & Woodger, S. C. (2003). Pilot summative web assessment in secondary education. In *Proceedings of the 7th International Computer Assisted Assessment Conference* (pp. 19-29).

Bardelle, C., & Ferrari, P. L. (2011). Definitions and examples in elementary calculus: The case of monotonicity of functions. *ZDM, 43*(2), 233-246. https://doi.org/10.1007/s11858-010-0303-4

Bardini, C., Pierce, R., Vincent, J., & King, D. (2014). Undergraduate mathematics students' understanding of the concept of function. *Indonesian Mathematical Society Journal on Mathematics Education, 5*(2), 85-107. https://doi.org/10.22342/jme.5.2.1495.85-107

Bell, C. J. (2001). *Conceptual understanding of functions in a multi-representational learning environment* [PhD thesis, The University of Texas at Austin].

Bezuidenhout, J. (2001). Limits and continuity: Some conceptions of first-year students. *International Journal of Mathematical Education in Science and Technology, 32*(4), 487-500. https://doi.org/10.1080/00207390010022590

Bond, T. G., & Fox, C. M. (2015). *Applying the Rasch model: Fundamental measurement in the human sciences*. Routledge.

Boone, W. J., & Staver, J. R. (2020). Test information function (TIF). In W. J. Boone, & J. R. Staver (Eds.), *Advances in Rasch analyses in the human sciences* (pp. 39-55). Springer. https://doi.org/10.1007/978-3-030-43420-5_4

Boone, W. J., Staver, J. R., & Yale, M. S. (2013). *Rasch analysis in the human sciences*. Springer. https://doi.org/10.1007/978-94-007-6857-4

Brown, J. D. (2013). Research on computers in language testing: Past, present and future. In M. Thomas, H. Reinders, & M. Warschauer (Eds.), *Contemporary computer-assisted language learning* (pp. 73-94). Bloomsbury Publishing.

Bull, J., & McKenna, C. (2004). *A blueprint for computer-assisted assessment*. Routledge. https://doi.org/10.4324/9780203464687

Carlson, M., Oehrtman, M., & Engelke, N. (2010). The precalculus concept assessment (PCA) instrument: A tool for assessing reasoning patterns, understandings and knowledge of precalculus level students. *Cognition and Instruction, 28*(2), 113-145. https://doi.org/10.1080/07370001003676587

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory*. Holt, Rinehart and Winton.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52*, 281-302. https://doi.org/10.1037/h0040957

Day, J., & Bonn, D. (2011). Development of the concise data processing assessment. *Physical Review Special Topics-Physics Education Research, 7*(1), 010114. https://doi.org/10.1103/PhysRevSTPER.7.010114

Dreyfus, T., & Eisenberg, T. (1982). Intuitive functional concepts: A baseline study on intuitions. *Journal for Research in Mathematics Education, 13*(5), 360-380. https://doi.org/10.5951/jresematheduc.13.5.0360

Elia, I., & Spyrou, P. (2006). How students conceive function: A triarchic conceptual-semiotic model of the understanding of a complex concept. *The Mathematics Enthusiast, 3*(2), 256-272. https://doi.org/10.54870/1551-3440.1053

Even, R. (1998). Factors involved in linking representations of functions. *The Journal of Mathematical Behavior, 17*, 105-121. https://doi.org/10.1016/S0732-3123(99)80063-7

Fisher Jr, W. P. (1994). The Rasch debate: Validity and revolution in educational measurement. In M. Wilson (Ed.), *Objective measurement: Theory into practice* (pp. 36-72). Ablex.

Gagatsis, A., & Shiakalli, M. (2004). Ability to translate from one representation of the concept of function to another and mathematical problem solving. *Educational Psychology, 24*(5), 645-657. https://doi.org/10.1080/0144341042000262953

Glamočić, D. S., Mešić, V., Neumann, K., Sušac, A., Boone, W. J., Aviani, I., Hasović, E.,Erceg, N., Repnik, R., & Grubelnik, V. (2021). Maintaining item banks with the Rasch model: An example from wave optics. *Physical Review Physics Education Research, 17*(1), 010105. https://doi.org/10.1103/PhysRevPhysEducRes.17.010105

Habre, S., & Abboud, M. (2006). Students' conceptual understanding of a function and its derivative in an experimental calculus course. *The Journal of Mathematical Behavior, 25*(1), 57-72. https://doi.org/10.1016/j.jmathb.2005.11.004

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory*. SAGE.

Hitt, F. (1998). Difficulties in the articulation of different representations linked to the concept of function. *The Journal of Mathematical Behavior, 17*, 123-134. https://doi.org/10.1016/S0732-3123(99)80064-9

Hrnjičić, A., Alihodžić, A., Čunjalo, F., & Kamber Hamzić, D. (2022). Development of an Item Bank for Measuring Students' Conceptual Understanding of Real Functions. *European Journal of Science and Mathematics Education*, *10*(4), 455-470. https://doi.org/10.30935/scimath/12222

Linacre, J. M. (1998). Detecting multidimensionality: Which residual data-type works best? *Journal of Outcome Measurement, 2*, 266-283. https://winsteps.com/a/Linacre-multidimensionality-residuals.pdf

Linacre, J. M. (2002). What do infit and outfit, mean-square and standardized mean? *Rasch Measurement Transactions, 16*, 878.

Linacre, J. M. (2008). Variance in data explained by Rasch measures. *Rasch Measurement Transactions, 22*(3), 1164. https://www.rasch.org/rmt/rmt221j.htm

Linacre, J. M. (2017). *A users guide to Winsteps: Rasch-model computer programs*. MESA Press.

Linacre, J. M. (2019). A user's guide to WINSTEPS. *Winsteps. com*. https://www.winsteps.com/manuals.htm

Linacre, J. M. (2022). A user's guide to WINSTEPS MINISTEP Rasch-model computer programs. *Winsteps.com*. https://www.winsteps.com/winman/copyright.htm

Liu, X. (2010). *Using and developing measurement instruments in science education: A Rasch modeling approach*. IAP.

Lloyd, G., Beckmann, S., Zbiek, R. M., & Cooney, T. (2010). *Developing essential understanding of functions for teaching mathematics in grades 9-12*. National Council of Teachers of Mathematics.

Mantel, N., & Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the National Cancer Institute, 22*(4), 719-748. https://doi.org/10.1093/jnci/22.4.719

Marcus, B., & Bühner, M. (2009). *Grundlagen der testkonstruktion* [*Basics of test construction*]. FernUniversität.

Mešić, V., Neumann, K., Aviani, I., Hasović, E., Boone, W. J., Erceg, N., Grubelink, V., Sušac, A., Glamočić, S. D., Karuza, M., Vidak, A., Alihodžić, A., & Repnik, R. (2019). Measuring students' conceptual understanding of wave optics: A Rasch modelling approach. *Physical Review Physics Education Research, 15*(1), 010115. https://doi.org/10.1103/PhysRevPhysEducRes.15.010115

Nitsch, R., Fredebohm, A., Bruder, R., Kelava, A., Naccarella, D., Leuders, T., & Wirtz, M. (2015). Students' competencies in working with functions in secondary mathematics education–Empirical examination of a competence structure model. *International Journal of Science and Mathematics Education, 13*(3), 657-682. https://doi.org/10.1007/s10763-013-9496-7

Nugent, G. (2003). On-line multimedia assessment for K-4 students. In *Proceedings of the World Conference on Educational Multimedia, Hypermedia and Telecommunications* (pp. 1051-1057).

Oon, P. T., Spencer, B., & Kam, C. C. S. (2016). Psychometric quality of a student evaluation of teaching survey in higher education. *Assessment & Evaluation in Higher Education, 42*(5), 788-800. https://doi.org/10.1080/02602938.2016.1193119

Overton, W. F. (1999). *Construct validity: A forgotten concept in psychology?* [Paper presentation]. The Annual Meeting of the American Psychological Association.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Danmarks Paedagogiske Institut.

Schmid, H. (1992). *Psychologische tests: Theorie und konstruktion* [*Psychological testing: Theory and construction*]. Hogrefe AG.

Sealey, C., Humphries, P., & Reppert, D. (2003). 'At the coal face': Experiences of computer-based exams. In *Proceedings of the 7th CAA Conference*.

Sierpinska, A. (1992). On understanding the notion of function. In E. Dubinsky, & G. Harel (Eds.), *The concept of function: Aspects of epistemology and pedagogy* (pp. 25-58). Mathematical Association of America.

Smith Jr, E. V. (2001). Evidence for the reliability of measures and validity of measure interpretation: A Rasch measurement perspective. *Journal of Applied Measurement, 2*(3), 281-311.

Sumintono, B., & Widhiarso, W. (2015). *Aplikasi pemodelan Rasch pada assessment pendidikan* [*Application of Rasch modeling to educational assessment*]. Trim Komunikata Publishing House.

Szabó, G. (2008). *Applying item response theory in language test item bank building*. Peter Lang. https://doi.org/10.3726/978-3-653-01167-8

Tennant, A., Horton, M., & Pallant, J. F. (2011). Introductory Rasch analysis: A workbook. *ScienceOpen*. https://www.scienceopen.com/document?vid=1a824134-b7f6-45f2-bc9c-a688c27b2e3e

Wolfe, E. W. (2000). Equating and item banking with the Rasch model. *Journal of Applied Measurement, 1*(4), 409-434.

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