

Measurement of Psychometric Properties Numerical Aptitude Assessment Scale for Prospective High School Students: A Rasch Model Analysis

M. Rais Ridwan^{1,2}, Samsul Hadi³, Jailani Jailani⁴

¹Study Program of Research and Evaluation Education, Graduate School, Universitas Negeri Yogyakarta, Jl. Colombo No. 1 Yogyakarta, Yogyakarta

²Study Program of Mathematics Education, STKIP YPUP Makassar, Jl. Andi Tonro No. 17 Makassar, Makassar, Indonesia

³Department of Electrical Engineering Education, Universitas Negeri Yogyakarta, Jl. Colombo No.1 Yogyakarta, Yogyakarta, Indonesia

⁴Department of Mathematics Education, Universitas Negeri Yogyakarta, Jl. Colombo No.1 Yogyakarta, Yogyakarta, Indonesia

Abstract – Developing structured test instruments is essential in forming a measurable and consistent ability assessment construct. This study examines the psychometric properties of the numerical aptitude test instrument (NAPTIN) utilizing Rasch model analysis to evaluate standardized numerical aptitude assessment test instruments for prospective high school (PHS) students based on validation and construct reliability. Data was collected using an online survey of 228 PHS students. The Rasch model analyses person and item fit, item measure, dimensionality and local independence testing, reliability, and differential item functioning. The results showed that NAPTIN had appropriate and consistent items for measuring the numerical aptitude of PHS students based on gender and school type. This study recommends a systematic measurement tool for researchers. Another contribution is measurement instruments for Indonesian schools and policymakers to choose student program specialties.

Keywords – Differential item functioning, numerical aptitude test, Rasch model analysis, validity, reliability.

1. Introduction

Validity is the leading and essential factor for a research study to develop test or non-test instruments for all fields of science (such as health, social sciences, and education). The instrument needs to identify the relevance of the theoretical study evidence with the data from field trials and the measuring instrument's consistency based on the results of repeated tests over a long period of time and different research samples. Validity is the relevance of evidence and theory to interpreting scores and the test's purpose. At the same time, reliability is the consistency of test results that are carried out repeatedly [1]. Validity is evidence and theory support for interpreting test scores according to the test's purpose [2]. Validity refers to what a test is intended to measure or the purpose of the test. In addition, validity also refers to how far the test score provides accurate information for decisions based on the test score [3]. Reliability describes how the treatment, test, or measurement procedure produces the same results for repeated treatments. Reliability describes a measure of the consistency of an instrument based on the interpretation of the resulting score [4]. Reliability describes the extent to which the measurement results have a level of trust, reliability, stability, consistency, and stability [5].

Although many researchers have validated the development of aptitude test instruments, validation using the Rasch model analysis approach still needs to be done.

DOI: 10.18421/TEM124-54

<https://doi.org/10.18421/124-54>


Corresponding author: Samsul Hadi,
Department of Electrical Engineering Education,
Universitas Negeri Yogyakarta, Jl. Colombo, Indonesia
Email: samsul_hd@uny.ac.id

Received: 05 June 2023.

Revised: 07 September 2023.

Accepted: 13 September 2023.

Published: 27 November 2023.

 © 2023 M. Rais Ridwan, Samsul Hadi & Jailani Jailani; published by UIKTEN. This work is licensed under the Creative Commons Attribution-NonCommercial-NoDerivs 4.0 License.

The article is published with Open Access at <https://www.temjournal.com/>

Several researchers have carried out the development of aptitude test instruments for several purposes, such as the selection of prospective student admissions to higher education [6], [7], selection of prospective high school students [8], [9], as well as mapping the specializations of high school students in Indonesia [10], [11]. Development of other test instruments to assess language aptitude abilities with research samples of foreign language learners [12], university students [13], elementary school students [14], and native language speakers [15]. Other relevant research studies have been conducted to assess the talents of vocational high school students in Taiwan [16] and to evaluate students' mathematical abilities in Pakistan [17]. Additionally, there have been studies examining indicators of high school student graduation in Saudi Arabia [18].

These researchers carried out several studies on the development of measuring instruments, namely construct validation of the development of talent instruments using a confirmatory factor analysis approach [6], [8], [18]. Other studies have only identified studies related to content validity and test instruments [9], [10], [13] and only some studies examine differential validity using differential item functioning analysis [7] to identify differences in students' abilities towards the same item characteristics. Several researchers in Asia and Europe have used Rasch model analysis to validate the development of aptitude or intelligence instrument constructs and their association with learning. Identification of the quality of the talent instrument with the Rasch model analysis approach has been carried out by Kara, et al. [19] was used to assess the spatial aptitude of elementary school students in Turkey and for the secondary school level in Australia [20]. Cramman, et al. [21] used an analytical technique based on the Rasch measurement model to evaluate learning instruments based on Hindu-Arabic numeric symbols to improve students' mathematical ability in England and Scotland. Other relevant research studies to identify the suitability of language proficiency assessment instruments with the theory of language aptitude tests involving student respondents in several countries in Europe and Asia [13]. Evaluation of the development of an inductive reasoning test instrument to predict academic success and cognitive intelligence processes of students in Vietnam was carried out by Van Vo and Csapó [22] using the Rasch model analysis approach based on identifying the suitability of items with the test instrument.

Subsequent research studies used the same analytical technique to identify the psychometric properties and quality of numerical comprehension instruments for elementary school students in

Indonesia [23] and New Zealand [24]. Weller, et al. [25] also evaluated psychometric properties in developing a numeracy ability scale instrument with research samples aged 18-89 years in the United States. Ilić, et al. [26] and Endler and Bond [27] used Rasch model analysis to validate the test instrument construct on the results of developing a measurement tool for assessing operational logic understanding in Serbian and Australian primary and secondary school students. Subsequent research by Vasilyeva, et al. [28], namely evaluating instruments with the same analytical approach to developing intelligence assessment measuring instruments consisting of analytical and conceptual-based abilities for elementary school students in the United States. Evaluation of other aptitude instruments, namely the word knowledgeability subtest, is used for placement of career abilities using undergraduate and graduate-level student respondents at the University of Florida [29]. Evaluation of psychometric properties using the Rasch model analysis approach to develop intelligence tests using Raven's Advanced Progressive Matrix construct [30] with a study sample of students at the University of Toronto.

The validity of developing numerical aptitude test instruments using the Rasch model analysis approach is still rarely carried out, with validity and reliability studies taking into account the characteristics of persons and items and based on demographic variables of the research sample (such as gender and school status). The Rasch model is a psychometric method that aims to increase accuracy in item construction and item quality and measure respondents' ability [31]. The Rasch measurement model is valid for rating scale items and can compute the likelihood of correct test item responses [32]. The approach using the Rasch model analysis is a structured and comprehensive approach to determining the reliability and validity of the test instrument. According to Bond and Fox [33], the Rasch model analysis comprehensively evaluates an instrument's validity. Determining the construct validity of an instrument with the Rasch model analysis approach provides criteria or indicators related to how much the match level for each item defines the underlying construct [34].

Furthermore, the Rasch model analysis method yields detailed data for respondents and instruments [35], [36]. Other analysis results from this model technique can use Rasch probability to categorize fit responses depending on skill levels. In addition, the Rasch model analysis results can categorize the difficulty level for each item in the instrument and identify biased items and rating scales. The results of the subsequent analysis obtained item and person reliability, item separation index and person, and Cronbach's alpha coefficient.

Reliability and validity are measured in classical theory tests based on Cronbach's alpha coefficients and factor analysis. However, the Rasch model analysis approach, which emphasizes measure reproducibility over raw score reproducibility, requires the determination of the two measurement qualities [37]. This research study analyses the Rasch model to assess the psychometric properties of the NAPTIN development results. The steps are (1) person and item fit analysis, (2) measure items, (3) dimensionality and local independence tests, (4) reliability analysis and separation index, and (5) differential item function (DIF) analysis.

DIF analysis is a type of differential validity that refers to differences in individual abilities in different sub-groups based on the nature or characteristics of the same item [38]. Items are identified as DIF if they have statistical characteristics different from individuals in different subgroups. DIF is a statistical term used to describe test items having different item difficulty estimates across different subgroups [39], [40], [41]. The DIF theory refers to the problem of a measuring instrument or test instrument related to differences in the functioning of items for different groups of test takers. Suppose two test takers with the same ability but different attributes (such as gender, school status, ethnicity, or language) have a varying chance of responding correctly to an item. In that case, that item is known as DIF. DIF-identified items lead to biased ability measurements that are influenced by confounding factors [42] and can potentially affect the results of score validity [43].

A specific educational or psychological test has many items identified by DIF with the possibility of unfair conditions for certain groups, so it is necessary to identify these items and correct or remove them from the test instrument [44]. The difference in the functioning of these items for groups of test takers with different characteristics provides information that the test instrument has biased items. A test is said to be biased if there is evidence of an interaction between group members and the performance of the test items taking into account differences in abilities or psychological conditions between groups [45]. The DIF procedure is designed to identify each item that functions differently relative to some of the identified criteria.

Practically and procedurally, this study aims to determine the quality of standardized numerical aptitude assessment test instruments based on the validation results and construct reliability using the Rasch model analysis approach. Determination of construct validation uses person analysis and item fit, item measure, dimensionality, local independence tests, and DIF analysis. The construct reliability study consists of identifying the reliability coefficients of persons, items, and tests and

determining the quality of other instruments based on the value of the separation index.

This research study significantly contributes to the construct validation process resulting from developing test and non-test measuring instruments for various studies in other fields of science. Specifically, this study contributed to the determination of respondents who met the unfit criteria who participated in the test of the instrument, identification of items that did not fit used for the assessment of numerical giftedness, determination of the number of measurable dimensions in the instrument, identification of respondents' independence in responding to instrument items and reporting instrument quality based on reliability coefficient and separation index. In addition, this study has contributed to determining valid items by identifying differences in respondent groups in responding to items based on gender and school-type variables. The characteristics of valid items indicated no significant difference in the ability to respond to items for the two groups of respondents.

2. Methodology

The research methodology section of this article consists of the research instruments used to collect data on the NAPTIN trials. The second part consists of data on the characteristics of test takers or respondents consisting of prospective high school students in several regions in Indonesia. The next part, namely data analysis, involves using programs or software to analyze instrument trial data and data analysis techniques consisting of analytical approaches and procedures to obtain research results. In addition, the data analysis section describes the criteria based on the theory that supports the research results.

2.1. Instruments

Data collection used a numerical talent ability assessment instrument based on the results of developing a test measuring instrument using a review of relevant literature consisting of research articles and textbooks. A literature review was conducted to determine the measurable variables that define the observed variables (namely, numerical aptitude), the indicators for each measurable variable, and the preparation of instrument items based on indicators. The assessment instrument's content validation determines the appropriateness scale of measured variables supporting observed variables, indicators of measured variables, and question items with indicators. Content validation used three material experts relevant to the field of mathematics consisting of 1 lecturer with a doctoral degree.

In contrast, the other lecturers and one teacher with the last educational qualification was a master’s degree. The validation results used the content validation index (CVI) based on the number of expert approvals obtained for each item with a CVI value of 1.00. Valid item criteria use a CVI value equal to 1.00 [46].

Table 1. NAPTIN constructs and items

Constructs	Number of items	Total items
Algebraic ability (ALJ)	ALJ1.1, ALJ1.2, ALJ1.3, ALJ2.4, ALJ2.5, ALJ3.6, ALJ3.7, ALJ4.8, ALJ4.9, ALJ4.10	10
Arithmetic ability (ART)	ART1.1, ART1.2, ART1.3, ART1.4, ART2.5, ART2.6, ART3.7, ART3.8, ART4.9, ART4.10	10
Geometric ability (GEO)	GEO1.1, GEO1.2, GEO1.3, GEO1.4, GEO2.5, GEO2.6, GEO2.7, GEO3.8, GEO4.9, GEO4.10	10
Total items		30

2.2. Participants

This study’s respondents were prospective high school pupils. The sampling technique used purposive random sampling, which pays attention to the types of public and private schools. The purposive random sampling technique is also called judgment sampling, namely the selection of research samples that are deliberately carried out by paying attention to the characteristics of the respondents based on age, background, and culture [47]. Data collection uses an online survey via Google Forms. The research participants who responded were 228 respondents from 3 regional locations in Indonesia, namely West Java, South Kalimantan, and South Sulawesi. Table 2 describes the data related to the characteristics of research respondents based on gender and school-type variables.

Table 2. Demographic characteristics of research respondents (N = 228)

Variable	Category	Frequency	Percentage
Gender	Male	98	42.98
	Female	130	57.02
Type of School	Public school	128	56.14
	Private school	100	43.86

2.3. Data Analysis

The response dataset of 228 prospective high school students was procedurally analyzed using the Winsteps 3.7.3 program. Data analysis techniques used the Rasch model analysis approach with procedures consisting of person fit analysis and person point measure correlation, item fit analysis

and item point measure correlation, item measure, dimensionality, and local independence tests, DIF analysis, and reliability analysis, and separation index.

The first stage involved analyzing data from 228 respondents’ responses to 30 questions to determine person fit using the infit mean-square (MNSQ) and z-standardized (ZSTD) criteria. MNSQ infit score criteria at intervals of 0.50 and 1.50 [48], [49], [50] and scores for ZSTD infit at intervals of -2.00 and 2.00 [33], [39]. In the second stage, people with unfit criteria were eliminated and excluded in the person point measure correlation analysis stage. The analysis at this stage uses the criteria for a positive point measure correlation (PMC) value [39], [51]. The dataset from the PMC person analysis that meets the fit criteria is then subjected to an item fit analysis in the third step. This stage aims to identify fit items with the criteria of using the same MNSQ and ZSTD infit values in the person fit analysis. Items that did not fit were eliminated before PMC item analysis. Criteria with a positive PMC value are also used for the PMC item analysis stage—the item measure’s value is used in the fifth analysis stage. If items have the same measure value, then one item is retained and used for the subsequent analysis [52]. Items were maintained using fit criteria based on the MNSQ and ZSTD infit values close to 1.00 and zero, respectively [51].

Items that match the fit criteria based on the previous analysis stages are utilized in the sixth stage to assess the number of dimensions and identify the numerical aptitude test instrument’s local independence. The unidimensional determination criterion in this study used the results of the principal component analysis of the residuals (PCAR) based on the second residual contrast eigenvalue, which is smaller than 2.00 [53]. Then, identify the conditions of local independence using the standard residual correlation values criteria between items smaller than 0.70 [54]. The seventh stage is the determination of bias items using DIF analysis. DIF analysis uses the type of DIF uniform effect using two groups with different respondent abilities based on school type and gender variables. The uniform DIF effect indicates differences in abilities for each member of the two groups based on school type and gender. The criteria for determining bias items for the two DIF effects use a contrast DIF value greater than 0.64 or a probability value based on the Mantel-Haenszel analysis, which is smaller than 0.05 [55].

The final analysis stage, reporting the results of the statistical summary, consists of Cronbach’s alpha value, the person and item reliability coefficient, and the person and item separation index. The Cronbach’s alpha coefficient criterion greater than or equal to 0.70 fulfills the reliable criteria [56], [57].

Then, the criteria for instruments that meet reliable conditions based on the reliability coefficient of person and item are each value greater than 0.70 [58]. As for determining the quality of other instruments based on the separation of person and item indexes, each uses a value greater than 1.50 and 2.00 [39], [59].

3. Results

The results of data analysis on 228 research respondents who were test responses to NAPTIN with a test instrument length of 30 items using the Rasch model analysis approach consisting of a summary table of the analysis results of item and person fit, person and item PMC, item measure, dimensionality test, and local independence. The following results are uniform DIF analyses based on gender and school-type variables. Other analysis results are statistical summary tables of Cronbach's alpha values, person and item reliability coefficients, and person and item separation indexes.

Table 4. Summary of PMC person analysis results on NAPTIN

Analysis	Respondents of category	PMC	Total respondents excluded
First	Male respondents with public school status	-0.30 to -0.07	11
	Female respondents with public school status	-0.23 to -0.01	13
	Male respondents with private school status	-0.28 to -0.01	14
	Female respondents with private school status	-0.28 to -0.01	14
Second	Female respondents with public school status	-0.02	1
	Female respondents with private school status	-0.01	1
Third	No respondents were excluded.	0.00 to .68	0
Overall total			54

Overall, the findings of the person fit study in Table 4 employing the PMC value criteria consisting of three stages of investigation yielded 54 respondents who were not fit with negative PMC values. The PMC score criteria in this study were positive [39], [51]. The analysis procedure was carried out three times in stages by identifying the PMC value for each stage of the analysis. In the first stage of analysis obtained 52 respondents with negative PMC values. The 52 respondents were not included in the second phase of the analysis, with the analysis results obtained from two respondents who needed to be fit. Then, for the third stage, the two respondents were eliminated so that 174 met the fit criteria. Respondents who met the fit criteria were used for item fit analysis of as many as 30 items.

Table 3 shows the results of the person fit analysis consisting of 1 stage of analysis obtained by all respondents, as many as 228 students meeting the fit criteria with MNSQ infit scores at intervals of 0.80 and 1.17 and for ZSTD infit values at intervals of -1.80 and 1.40. This investigation determined the fit response criteria using the MNSQ infit value at intervals between 0.50 and 1.50 [48], [49], [50] and the ZSTD infit value at intervals between -2.00 and 2.00 [33], [39]. These respondents were utilized in the next-person fit study based on PMC values.

Table 3. Summary of the results of person fit analysis on NAPTIN

Analysis	Respondents of category	Infit ZSTD	Infit MNSQ	Total respondents excluded
First	No respondents were excluded.	-1.80 to 1.40	0.80 to 1.17	0
Overall total				0

Table 5. Summary of the results of the analysis of item fit on NAPTIN

Analysis	Item number	Infit ZSTD	Infit MNSQ	Total items excluded
First	ART1.3	-2.50	0.83	2
	ALJ4.10	4.10	1.20	
Second	ALJ2.4	2.10	1.25	1
Third	No items were excluded.	-1.90 to 1.80	0.87 to 1.17	0
Overall total				3

The fit item analysis results in Table 5 obtained three items that needed to be fitted with two items in the first analysis and 1 item in the second analysis. Meanwhile, in the third analysis stage, 27 fit items were obtained. The fit criteria use the same MNSQ and ZSTD infit value categories in the person fit analysis.

Analysis in the first stage obtained items ART1.3 and ALJ4.10 with respective ZSTD infit values of -2.50 and 4.10, so these two items were not fit and were not included in the second analysis stage. Then, for the second analysis stage, without including the two items, it was found that the ALJ2.4 item did not fit with an infit ZSTD value of 2.10 which is greater than 2.00. The twenty-seven fit items and 174 fit respondents were used for the following fit item analysis based on the PMC item value.

Table 6. A summary of the PMC analysis results for NAPTIN items

Analysis	Item number	PMC	Total items excluded
First	No items were excluded.	0.20 to 0.54	0
Overall total			0

The following fit item analysis using PMC items obtained 27 fit items with PMC values of 0.20 and 0.54, all of which were positive. The analysis described in Table 6 was only carried out in 1 stage of analysis. Twenty-seven fit items were obtained by analyzing the previous item fit stage. The item fit criteria use a positive PMC value [39], [51]. The following procedure still uses item fit analysis based on the measured value for each item. Identify fit items based on items in the same construct with the same measure value. For the first analysis stage in Table 7, 2 items were obtained in the algebraic and arithmetic ability constructs with the same item measure values of 0.15 and -1.05. As a result, MNSQ and ZSTD infit values must be used to identify fit items. As for the second analysis stage, it was found that ART3.7 and ART4.9 had the same measure value of 0.08 in the arithmetic ability construct. Item fit criteria use the MNSQ and ZSTD infit values close to 1.00 and 0.00, respectively [51]. The analysis results in the first stage contained items ALJ1.3 and ART2.5, as well as items ART3.7 in the second analysis stage. The ZSTD value was close to 0.00, and the MNSQ infit value was close to 1.00 compared to items ALJ1.1, ART2.6, and ART4.9. Thus, 24 fit items were obtained with 174 respondents for use in the analysis of determining the number of dimensions in NAPTIN, identification of local independence, and DIF analysis as well as reliability analysis and analysis of instrument quality based on separation index.

Table 7. A summary of the results of the NAPTIN item measurement analysis

Construct/Analysis	IN	MV	Infit ZSTD	Infit MNSQ	Results
First analysis					
ALJ	ALJ1.1	0.15	0.70	1.07	Excluded
	ALJ1.3	0.15	-0.10	0.99	Retained
ART	ART2.5	-1.05	-1.20	0.94	Retained
	ART2.6	-1.05	-1.90	0.90	Excluded
Second analysis					
ART	ART3.7	0.08	0.80	1.08	Retained
	ART4.9	0.08	-1.30	0.87	Excluded
Total items excluded					3

Note. IN – Item Number; MV – Measure Value

Table 8. Results of the PCAR analysis of NAPTIN

	Eigen value	Expected (%)	Observed (%)
Total raw variance in observations	29.80	100	100
Raw variance explained by measures	5.80	19.40	19.50
Raw variance explained by persons	2.10	7.10	7.10
Raw Variance explained by items	3.70	12.30	12.40
Raw unexplained variance (total)	24.00	80.60	80.50
Unexplained variance in 1st contrast	2.00	8.20	6.60
Unexplained variance in 2nd contrast	1.60	6.70	5.40

Table 8 displays the findings of the subsequent analysis conducted to determine the number of dimensions in NAPTIN using PCAR. Unidimensional identification using eigenvalue criteria for unexplained variance is explained in the second contrast, which is smaller than 2.00 [53]. The PCAR analysis obtained an eigenvalue of 1.60 so that NAPTIN only measures one dimension. The following analysis procedure is the identification of local independence in NAPTIN. Table 9 shows the analysis results of correlation items between items as many as 23 pairs of items, each with the largest standardized residual correlation values between items at intervals of -0.0546 and 0.2275. The correlation value for the twenty-three pairs of items is less than 0.70, showing that NAPTIN meets the local independence criteria.

Table 9. A summary of NAPTIN item's largest standardized residual correlation values

Item number	Item number	Correlation
ALJ1.2	ALJ4.8	0.1118
ALJ1.3	GEO1.3	0.0542
ALJ2.5	ALJ4.8	0.2275
ALJ3.6	ART2.5	0.1968
ALJ3.7	GEO2.5	0.0625
ALJ4.8	GEO4.9	0.1413
ALJ4.9	GEO1.4	0.0441
ART1.1	GEO2.6	0.0274
ART1.2	GEO1.1	0.0945
ART1.4	ART3.8	0.0895
ART2.5	GEO1.1	0.0699
ART3.7	GEO3.8	0.1026
ART3.8	GEO1.2	0.1023
ART4.10	GEO4.10	0.1440
GEO1.1	GEO2.6	0.0673
GEO1.2	GEO2.6	0.0234
GEO1.3	GEO2.5	0.0656
GEO1.4	GEO2.6	-0.0283
GEO2.5	GEO4.10	0.0085
GEO2.6	GEO2.7	0.1065
GEO2.7	GEO3.8	-0.0219
GEO3.8	GEO4.10	0.0089
GEO4.9	GEO4.10	-0.0546

Table 10. Summary of DIF uniform of NAPTIN test results by gender

Item Number	DIF Measure		DIF Contrast	Mantel-Haenszel (Prob.)
	Females (F)	Males (M)		
ALJ1.2	-0.18	0.25	-0.44	0.1370
ALJ1.3	0.42	-0.27	0.69	0.1796
ALJ2.5	0.29	0.99	-0.70	0.3935
ALJ3.6	1.05	-0.28	1.32	0.0014
ALJ3.7	0.71	0.99	-0.28	0.9435
ALJ4.8	-0.39	-0.27	-0.12	0.8250
ALJ4.9	0.63	0.25	0.38	0.8672
ART1.1	-1.67	-1.48	-0.20	0.6449
ART1.2	-0.88	-0.57	-0.31	0.4233
ART1.4	-0.39	-0.03	-0.37	0.6730
ART2.5	-1.08	-1.10	0.02	0.8838
ART3.7	0.49	-0.43	0.91	0.0640
ART3.8	-0.49	-0.78	0.28	0.3691
ART4.10	0.35	0.46	-0.11	0.9022
GEO1.1	0.56	0.36	0.20	0.8166
GEO1.2	0.29	0.16	0.13	0.4006
GEO1.3	0.42	0.06	0.35	0.7703
GEO1.4	0.63	0.36	0.28	0.8425
GEO2.5	-0.39	-0.03	-0.37	0.4572
GEO2.6	0.16	0.58	-0.42	0.6628
GEO2.7	-0.24	0.46	-0.70	0.0941
GEO3.8	-0.44	-0.11	-0.33	0.3856
GEO4.9	0.87	0.46	0.41	0.5715
GEO4.10	-0.29	0.06	-0.35	0.3981

The following analysis results are the DIF analysis of NAPTIN items by considering gender and school-type variables. Tables 10 and 11, respectively, show the results of the uniform DIF analysis based on school type (private vs. public) and gender (male vs. female) variables. In addition, Figures 1(a) and 1(b), respectively, show the person DIF measure plot for each item based on these two variables.

Table 10 and Figure 1(a) show no statistically significant difference in item difficulty for each pair of female and male respondents, except for items ALJ1.3, ALJ3.6, and ART3.7. These three items each have a DIF measure value of 0.69 ($p = 0.1797$), 1.32 ($p = 0.0014$), and 0.91 ($p = 0.0640$), which is greater than 0.64, indicating that the item is biased [55]. These three items provide significantly different difficulty levels for groups of male and female students. Furthermore, these three items had a more accessible difficulty level for the male student group than the female group. The uniform DIF analysis then revealed that for the variable type of school, there were three NAPTIN items, namely ALJ4.9, ART4.10, and GEO2.5, each of which presented a statistically significant difference for each pair of student responder groups at public and private schools. The magnitude of the difference based on the contrast DIF values were 1.02 ($p = 0.0391$), 0.66 ($p = 0.0890$), and 0.69 ($p = 0.1634$), respectively. Table 11 and Figure 1(b) respectively show that the three items have a more complex level of difficulty for groups of students in public schools than private schools.

Table 11. Summary of NAPTIN uniform DIF test results based on school type

Item Number	DIF Measure		DIF Contrast	Mantel-Haenszel (Prob.)
	Public school (S)	Private school (V)		
ALJ1.2	-0.18	0.21	-0.39	0.4198
ALJ1.3	-0.06	0.40	-0.46	0.5817
ALJ2.5	0.25	0.96	-0.71	0.2635
ALJ3.6	0.39	0.50	-0.11	0.6912
ALJ3.7	0.98	0.60	0.37	0.4825
ALJ4.8	-0.55	-0.04	-0.50	0.3113
ALJ4.9	0.98	-0.04	1.02	0.0391
ART1.1	-1.69	-1.43	-0.26	0.4821
ART1.2	-0.88	-0.57	-0.31	0.5494
ART1.4	-0.34	-0.12	-0.22	0.6749
ART2.5	-0.97	-1.23	0.26	0.5561
ART3.7	-0.12	0.40	-0.52	0.4202
ART3.8	-0.39	-0.91	0.51	0.3769
ART4.10	0.70	0.04	0.66	0.0890
GEO1.1	0.39	0.60	-0.22	0.7042
GEO1.2	0.18	0.30	-0.12	0.6562
GEO1.3	0.18	0.40	-0.22	0.6691
GEO1.4	0.46	0.60	-0.14	0.8714
GEO2.5	0.06	-0.64	0.69	0.1634
GEO2.6	0.46	0.13	0.34	0.6892
GEO2.7	0.25	-0.28	0.53	0.2945
GEO3.8	-0.34	-0.28	-0.06	0.9827
GEO4.9	0.71	0.71	0.00	0.5706
GEO4.10	-0.18	-0.12	-0.05	0.8622

Thus, the uniform DIF analysis results based on the gender variable obtained 21 items with characteristics of differences that were not significant for the two groups of females vs. male. These items meet the valid criteria with the construct of 5 items of algebraic ability (ALJ1.2, ALJ2.5, ALJ3.7, ALJ4.8, ALJ4.9), six items of arithmetic ability (ART1.1, ART1.2, ART1.4, ART2.5, ART3.8, ART4.10), and for geometry ability variables as many as ten items namely GEO1.1, GEO1.2, GEO1.3, GEO1.4, GEO2.5, GEO2.6, GEO2.7, GEO3.8, GEO4.9, and GEO4.10. Whereas the DIF analysis based on school type variables (public vs. private) showed that there were 21 valid items with details of the algebraic ability construct as many as six items (ALJ1.2, ALJ1.3, ALJ2.5, ALJ3.7, ALJ4.8, ALJ4.10), six items for arithmetic skills (ART1.1, ART1.2, ART1.4, ART2.5, ART3.7, ART3.8), and for geometry ability variables as many as nine items namely GEO1.1, GEO1.2, GEO1.3, GEO1.4, GEO2.6, GEO2.7, GEO3.8, GEO4.9, and GEO4.10.

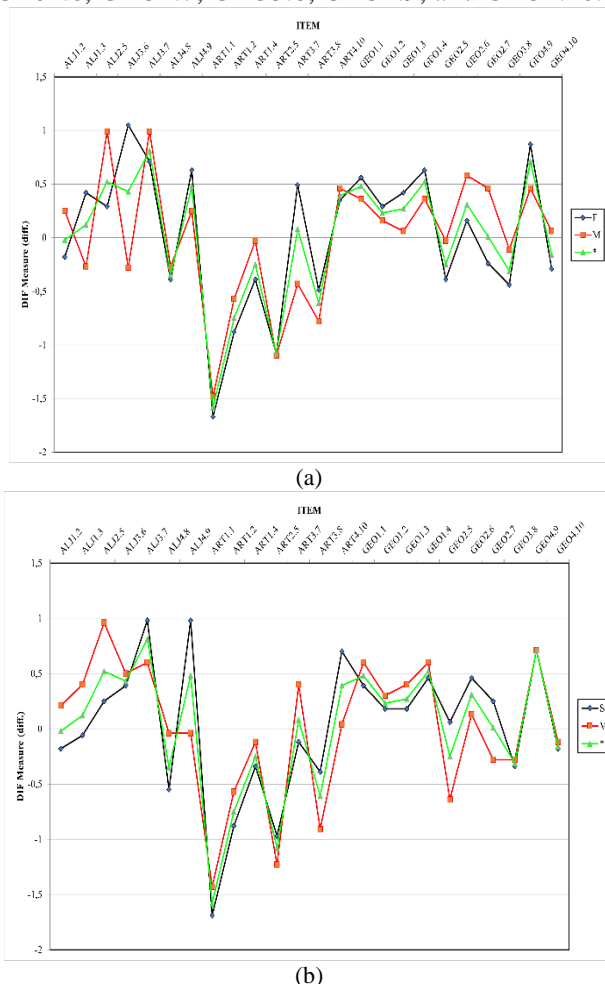


Figure 1. Plot the person DIF measure for each item based on the (a) gender and (b) school type variables

Figure 2 shows the ICC plot representing items identified as biased based on differences in the ability of groups of students (male vs. female) and (public vs. private).

Specifically, Figures 2(a) and 2(b) respectively show ICC plots for items ALJ3.6 and ALJ4.9, where the three items were identified as biased with positive contrast DIF values of 1.32 ($p = 0.0014$) and 1.02 ($p = 0.0391$). Each of these two items has two different ICCs for groups of students based on male (red curve) and female (blue curve) gender and for groups of students in public (blue curve) and private schools (red curve). The ICC shift to the right for the group of female students showed a more negligible probability of success than for male students to get a high score with a more complex level of item difficulty. In this case, the item indicates that female students have a more challenging level of difficulty than male students. Then, to shift the ICC to the right for groups of students in private schools, it shows a greater probability of success than students in public schools to get high scores with a more complex level of item difficulty so that the three items indicate that the group of students in private schools has a more accessible level of difficulty than students in public schools.

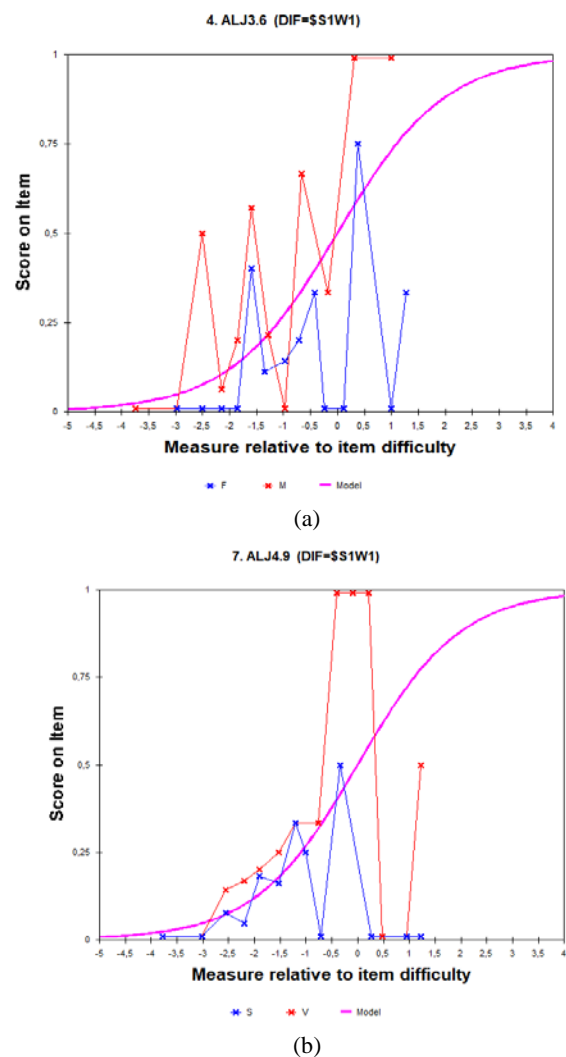


Figure 2. The plot of characteristic curve items for (a) ALJ3.6 and (b) ALJ4.9 items with uniform DIF based on gender variables and type of school variables

The results of the following analysis related to the statistical summary consist of the reliability coefficient and separation index. The statistical summary in Table 12 consists of a Cronbach's alpha value of 0.78, indicating a reliable NAPTIN with a reliability value criterion greater than 0.70 [56], [57]. The reliability analysis results using Cronbach's alpha show the consistency of NAPTIN with the same results even though they use different research samples. The following statistical summary consists of the item and person reliability coefficients, each worth 0.88 and 0.68, indicating that NAPTIN has a low level of reliability for use on respondents with different characteristics. In contrast, the use of items has a high level of reliability for assessing numerical aptitude abilities. Then, the results of the quality of the other instruments based on the separation index consisting of persons and items each obtained a coefficient value of 1.46 and 2.75. The results of the separation person index show that NAPTIN can distinguish between two groups (number of groups = $(4.00 \times \text{separation person index})/3.00$) based on the different ability levels of the respondents. Then, for the separation item index, it shows that NAPTIN has varying levels of item difficulty to measure the test takers' ability.

Table 12. Summary of NAPTIN's separation index and reliability coefficient

Summary of statistic	Value
Cronbach's alpha	0.78
Person reliability	0.68
Item reliability	0.88
Person separation	1.46
Item separation	2.75

4. Discussion

The psychometric components of the NAPTIN construct's validity and reliability are examined in this study. The study used a response dataset of 228 prospective high school students and a test instrument with 30 items for algebraic ability (ALJ), arithmetic ability (ART), and geometry ability (GEO). Construct validity consisted of person and item fit analysis, item measure, dimensionality, local independence tests, and DIF analysis, while construct reliability consisted of reliability analysis and separation index. The results showed that there were 174 respondents and 24 items that met the fit criteria; NAPTIN fulfills the conditions of one-dimensionality and local independence; the reliability of items, persons, and Cronbach's alpha coefficients were 0.88, 0.68, and 0.78, respectively; and the separation person and item indexes were 1.46 and 2.75, respectively.

Then, the DIF analysis found that 21 items met the valid criteria and were suitable for assessing the numerical aptitude of prospective high school students in Indonesia based on gender and school status variables, respectively.

Relevant research studies related to the use of the Rasch model analysis to identify the psychometric properties of the construct of talent or intelligence instrument development and also its relationship to ability assessment and implementation in the analysis process have been carried out by several researchers in several countries such as in mainland Asia [22], [23] and Europe [21], [25], [26]. Other relevant research studies have also been carried out by several researchers for assessing the talent abilities of elementary school students [28], evaluating career placement surveys [29], and identifying the development of intelligence instruments [30]. The same analytical approach has also been used by several researchers in mainland Oceania [20], [24], [27] and in Eurasia [13], [19].

Van Vo and Csapó [22] used the Rasch measurement model analysis to research the construction of inductive reasoning test instruments to discover cognitive intelligence processes and predict student academic progress in Vietnam. Identifying test item features using the Rasch model analysis to create valid and reliable instruments to measure students' inductive reasoning abilities at each school level. The results of the analysis of item characteristics for the four measurable variables (namely, the ability to complete a series of pictures, series of numbers, picture analogies, and numbers) using the ACER Conquest program use dichotomous data consisting of the discriminant index, difficulty level, and MNSQ infit value. The results of the following analysis use the output of the Rasch model analysis, namely the Wright map, to identify predictions of the suitability of the test instrument based on theoretical studies. The evaluation reports test item evaluation findings using the correct map, defines the measurable variables, and compares item difficulty levels based on expected and actual outcomes in the data set.

Another relevant research is identifying the psychometric aspects of numerical understanding instruments for elementary school students in Indonesia [23] and New Zealand [24]. A research study by Suranata, et al. [23] used the results of polytomous data responses to identify psychometric properties consisting of construct reliability and validity, test rubric quality, and analysis of instrument item identification based on differences in abilities for different groups of respondents.

Item quality analysis used the MNSQ and ZSTD outfit criteria; reliability studies and person-item separation indexes are used to evaluate other instruments; and PCA dimensionality test using raw variance criteria explained by measurements. Then, the quality of the test rubric is determined based on the assessment criteria used to classify the abilities of each student in response to each item. DIF analysis was also carried out, but only using criteria based on DIF size with categorical variables of respondents based on gender. The research study by Irwin and Irwin [24] compared the abilities of two to three different groups of students based on differences in treatment by considering the results of the analysis of the characteristics of the numerical comprehension test instrument items. Analysis of item characteristics consisted of difficulty levels for the three subtests: addition and subtraction, multiplication and division, and proportional reasoning.

Identification of the quality of the following talent instrument using the Rasch model analysis approach has been carried out by Ramful, et al. [20] to obtain a standardized test instrument for measuring the spatial aptitude of secondary school students in Australia and primary school level in Turkey [19]. A research study by Ramful, et al. [20] analyses item characteristics based on item difficulty level for the usefulness of the test instrument used at each grade level. Another psychometric characteristic is analyzing item quality or suitability for each measured variable (spatial visualization, spatial orientation, and mental rotation sub-tests) using the MNSQ infit and outfit criteria and the infit and outfit statistical scores. Then, the research study by Kara, et al. [19] identified the psychometric properties of the construct of the visual-spatial ability (VS) test instrument consisting of analysis of item quality using the MNSQ infit and outfit criteria; the criteria for data compatibility with the Rasch model used point measure correlation (PMC), expected correlation (EC), and root mean square error (RMSE). Other psychometric properties based on the grain separation index and the reliability of each test indicate the test instrument's quality. The dimensionality test employs a PCA-based eigenvalue criterion.

An approach using Rasch model analysis to identify the quality of cognitive development test instruments [27] and instruments for adapting mathematical logic structures to formal operations in Piaget's theory [26] respectively using research samples of elementary school and school students middle school in Australia and Serbia. Implementation of Rasch model analysis with the ACER Quest program by Endler and Bond [27] was carried out separately on age group data, which was used to calculate estimates of student abilities based on the age group variables.

Then, research studies by Ilić, et al. [26] consisted of an analysis of the characteristics of the logical operation test instrument items using fit criteria based on the MNSQ and ZSTD infit and outfit values, the following psychometric properties using item reliability analysis, and item invariant conditions using a comparison of the three versions of the test instrument's difficulty levels.

Another relevant research study was conducted by Cramman, et al. [21] to evaluate learning instruments using Hindu-Arabic numeric symbols to improve students' mathematical abilities in England and Scotland. Bokander and Emanuel [13] evaluated the suitability of the subtest on the language skills assessment instrument with the theory of the language aptitude test using the Rasch model analysis approach by involving student respondents in several countries in Europe and Asia. Analysis with the Winsteps 3.90 program was carried out separately using learning outcome data based on the pre and post-test stages to explore the psychometric properties of learning assessment using Hindu-Arabic numeric symbols [21]. Psychometric properties include MNSQ infit and outfit statistical criteria for item suitability with the model, PCA criteria for dimensionality, separation person index, and test reliability for instrument quality.

Weller, et al. [25] also evaluated psychometric properties to develop a numeracy ability scale instrument with research samples aged 18-89 in the United States. Psychometric properties consist of item analysis carried out procedurally in 2 stages of analysis based on statistical fit criteria using infit and outfit values. Each of the two stages of the analysis identified the quality of the instrument consisting of person reliability and test reliability using Cronbach's alpha coefficient. Another study by Vasilyeva, et al. [28] evaluated the intelligence assessment test instrument for elementary school students in the United States using the Rasch model. The measurable variables in the assessment instrument consist of analytical and conceptual-based abilities. Evaluation of psychometric properties for the test instrument with the two measurable variables each uses a map variable to determine the quality of the items based on the level of difficulty of the items. The results of another analysis are to identify the score of the respondent's ability based on the response data for each instrument item.

The quality of another aptitude tool, the word knowledge subtest, was evaluated to place career abilities with respondents from the University of Florida's graduate and undergraduate programs [29]. Analysis of the Rasch model using the Winsteps 3.31 program consists of a dimensionality test, a description of the items ordered in a hierarchy, and an analytical study based on the reliability coefficient and separation person index.

The dimensionality test based on the item analysis results using statistical criteria consists of an infit mean square standardized residual (MS) value that is less than or equal to 1.30 and a ZSTD score that is less than 2.00. Analysis of sorting items in a hierarchical manner using a person and item map, which describes the distribution of items and each person based on the difficulty level of the item and the test takers' ability. This hierarchical ordering of items aims to identify inconsistencies in the presentation of items in the ability subtest in the word knowledge subtest. The person reliability coefficient shows that the items consistently produce test takers' ability scores. In contrast, the separation person index shows that the instrument's condition can differentiate the test takers' abilities into three different ability levels. Evaluation of psychometric properties against the development of other intelligence test instruments using Raven's Advanced Progressive Matrix construct [30]. The analysis uses the Rasch measurement model based on student research sample data at the University of Toronto. The research study uses statistical criteria Q1 and Q2 to examine the test instrument's dimensionality depending on item difficulty and the test taker's ability. The unidimensional criterion uses a statistical significance test using the Rasch model with a p -value greater than 0.001. Identify other psychometric properties, namely item validation, by identifying differences in respondents' abilities based on gender using Q1 statistical values.

Overall, research studies on testing the quality of test instruments by several researchers in Asian, European, Oceanian, and Eurasian countries using the Rasch model analysis approach have characteristics and analytical adequacy that align with the research objectives. However, the current research study contributes to a procedural analytical study to identify psychometric properties of the constructs resulting from the development of aptitude test instruments. The analytical study uses the Rasch model of measurement, which pays attention to person and item variables. Another analytical study in the current research study used DIF analysis to obtain valid items based on the identification of no difference in the ability of test takers to respond to an item based on gender and school-type variables.

5. Conclusion

Standardized tests and non-test instruments have criteria for the relevance of theoretical study evidence with data from field trials and the consistency of

measuring instruments based on the same test results on the frequency of repeated test execution and the use of different research samples.

One alternative to identify the criteria for the relevance and consistency of the test or non-test instrument is the Rasch model analysis approach, which considers the characteristics of persons and items. The purpose of implementing the Rasch model analysis in this research project is to obtain a valid and reliable numerical aptitude test instrument for measuring the numerical aptitude of prospective high school pupils in Indonesia. The analysis consisted of a person fit analysis using the MNSQ and ZSTD infit value criteria and the person point measure correlation value. Person fit analysis is performed until all persons meet the fit criteria. The data from the fit response analysis were used to identify fit items with the MNSQ and ZSTD infit value criteria and point measure correlation items. Other item fit criteria use item measures to identify items that define the same measure variable. Data from person analysis and item fit are used for instrument prerequisite tests consisting of dimensionality tests and identification of local independence. The next psychometric trait uses DIF analysis to discover gender and school-type differences in test takers' item-response abilities. The reliability coefficients of items, persons, and Cronbach's alpha, as well as the value of the separation index of item and person, are used to assess the quality of other instruments. The limitation of this study is the use of a research sample consisting of prospective high school students in Indonesia. As a result, the research findings only contribute to adopting reliable and acceptable aptitude test instruments in Indonesia for determining majors in specialization programs at the senior high school level based on assessing students' numerical aptitude abilities. Nonetheless, the results of this research study contribute to science and become a reference for relevant research related to the validity and reliability of the constructs of test and non-test instruments using the Rasch measurement model approach.

Acknowledgements

The author thanks Indonesia's Ministry of Education, Culture, Research, and Technology for funding Yogyakarta State University's Ph.D. program. The author acknowledges the Indonesian Ministry of Education, Culture, Research, and Technology's Directorate of Research, Technology, and Community Service for providing a doctoral dissertation research grant under research contract number 042/E5/PG.02.00.PL/2023.

References:

- [1]. Thorndike, R. M., & Thorndike-Christ, T. (2014). Measurement and evaluation in psychology and education. *Journal of the American Statistical Association*, 56(296), 1029. Doi:10.2307/2282039
- [2]. Mardapi D. (2017). *Pengukuran, penilaian, dan evaluasi pendidikan (Measurement, assessment, and evaluation of education)*. Yogyakarta: Parama Publishing.
- [3]. Messick, S. (1995). Standards of validity and the validity of standards in performance assessment. *Educational Measurement: Issues and Practice*, 14(4), 5-8. Doi:10.1111/j.1745-3992.1995.tb00881.x
- [4]. Crocker, L. & Algina, J. (2008). *Introduction to classical and modern test theory*. Cengage Learning.
- [5]. Chakrabartty, S. N. (2013). Best split-half and maximum reliability. *OSR Journal of Research & Method in Education (IOSR-JRME)*, 3(1), 1-8.
- [6]. Carlstedt, B., & Gustafsson, J.-E. (2005). Construct validation of the Swedish Scholastic Aptitude Test by means of the Swedish Enlistment Battery. *Scandinavian Journal of Psychology*, 46(1), 31-42. Doi:10.1111/j.1467-9450.2005.00432.x
- [7]. Sirikit, R., & Mahalawalert, P. (2021). A Study of the Quality of Scholastic Aptitude Test by Applying Modern Test Theories. *Turkish Journal of Computer and Mathematics Education*, 12(11), 3423-3431.
- [8]. Ridwan, M. R., Hadi, S., Jailani, J., & Retnawati, H. (2023). The instrument development to measure the verbal ability of prospective high school students. *International Journal of Evaluation and Research in Education (IJERE)*, 12(1), 357-368. Doi:10.11591/ijere.v12i1.22736
- [9]. Yudha, I. W. P., Candiasa, I. M., & Indrawan, G. (2020). The development of online vocational aptitude test. *Journal of Physics: Conference Series*, 1516(1), 012036. Doi:10.1088/1742-6596/1516/1/012036
- [10]. Wulandari, F., Mardapi, D., & Haryanto. (2019). The Development of Students' Aptitude Test in Online and Multimedia Based Interests Group Selection. *Journal of Physics: Conference Series*, 1339(1), 012110. Doi:10.1088/1742-6596/1339/1/012110
- [11]. Jati, H., Ristanto, R. D., & Nurkhamid. (2019). Implementation of CBT (Computer-Based Test) System on Aptitude Test Development Using C4.5 Algorithm as Potential Detection Tool for Choosing High School Major. *Journal of Physics: Conference Series*, 1413(1), 012037. Doi:10.1088/1742-6596/1413/1/012037
- [12]. Li, L. & Luo, S. S. (2019). Development and preliminary validation of a foreign language aptitude test for Chinese learners of foreign languages. In Wen, Z. E., Skehan, P. Biedroń, A., Li, S., & Sparks, R. L. (Eds.), *Language aptitude: Advancing theory, testing, research and practice*. 33-55. Routledge.
- [13]. Bokander, L., & Emanuel, B. (2019). Probing the Internal Validity of the LLAMA Language Aptitude Tests. *Language Learning*, 70(1) 1-37. Doi:10.1111/lang.12368
- [14]. Kiss, C., & Nikolov, M. (2005). Developing, Piloting, and Validating an Instrument to Measure Young Learners' Aptitude. *Language Learning*, 55(1), 99-150. Doi:10.1111/j.0023-8333.2005.00291.x
- [15]. Sedaghatgoftar, N., Karimi, M. N., Babaii, E., & Reiterer, S. M. (2019). Developing and validating a second language pragmatics aptitude test. *Cogent Education*, 6(1), 1654650. Doi:10.1080/2331186X.2019.1654650
- [16]. Wen, J.-R., & Shih, W.-L. (2003). Information aptitude scale development for vocational high school in Taiwan. Proceedings. *3rd IEEE International Conference on Advanced Learning Technologies*.
- [17]. Hashmi, M. A., Zeeshan, A., Saleem, M., & Akbar, R. A. (2012). Development and Validation of an Aptitude Test for Secondary School Mathematics Students. *Bulletin of Education and Research*, 34(1), 65-76.
- [18]. Dimitrov, D. M., & Shamrani, A. R. (2015). Psychometric Features of the General Aptitude Test-Verbal Part (GAT-V). *Measurement and Evaluation in Counseling and Development*, 48(2), 79-94. Doi:10.1177/0748175614563317
- [19]. Kara, C., Coşkun, K., & Coskun, M. (2022). Development of Visual-Spatial Ability Test (VSAT) for Primary School Children: Its Reliability and Validity. *Interchange*, 53(2), 335-352. Doi:10.1007/s10780-022-09462-8
- [20]. Ramful, A., Lowrie, T., & Logan, T. (2017). Measurement of Spatial Ability: Construction and Validation of the Spatial Reasoning Instrument for Middle School Students. *Journal of Psychoeducational Assessment*, 35(7), 709-727. Doi:10.1177/0734282916659207
- [21]. Cramman, H., Gott, S., Little, J., Merrell, C., Tymms, P., & Copping, L. T. (2020). Number identification: a unique developmental pathway in mathematics?. *Research Papers in Education*, 35(2), 117-143. Doi:10.1080/02671522.2018.1536890
- [22]. Van Vo, D., & Csapó, B. (2020). Development of inductive reasoning in students across school grade levels. *Thinking Skills and Creativity*, 37, Article 100699. Doi:10.1016/j.tsc.2020.100699
- [23]. Suranata, K., Rangka, I. B., Ifdil, I., Ardi, Z., Susiani, K., Prasetyaningtyas, W. E., Daharnis, D., Alizamar, A., Erlinda, L., & Rahim, R. (2018). Diagnosis of students zone proximal development on math design instruction: A Rasch analysis. *Journal of Physics: Conference Series*, 1114, 012034. Doi:10.1088/1742-6596/1114/1/012034
- [24]. Irwin, K. C., & Irwin, R. J. (2005). Assessing development in numeracy of students from different socio-economic areas: A rasch analysis of three fundamental tasks. *Educational Studies in Mathematics*, 58(3), 283-298. Doi:10.1007/s10649-005-6425-x
- [25]. Weller, J. A., Dieckmann, N. F., Tusler, M., Mertz, C. K., Burns, W. J., & Peters, E. (2013). Development and Testing of an Abbreviated Numeracy Scale: A Rasch Analysis Approach. *Journal of Behavioral Decision Making*, 26(2), 198-212. Doi:10.1002/bdm.1751

- [26]. Ilić, I. S., Baucal, A., & Bond, T. G. (2012). Parallel Serbian Versions of BLOT Test: An empirical examination. *Psihologija*, 45(2), 121-137. Doi:10.2298/PSI1202121S
- [27]. Endler, L. C., & Bond, T. (2000). Cognitive Development in a Secondary Science Setting. *Research in Science Education*, 30(4), 403-416. Doi:10.1007/BF02461559
- [28]. Vasilyeva, M., Ludlow, L. H., Casey, B. M., & Onge, C. S. (2009). Examination of the psychometric properties of the measurement skills assessment. *Educational and Psychological Measurement*, 69(1), 106-130. Doi:10.1177/0013164408318774
- [29]. Pomeranz, J. L., Byers, K. L., Moorhouse, M. D., Velozo, C. A., & Spitznagel, R. J. (2008). Rasch analysis as a technique to examine the psychometric properties of a career ability placement survey subtest. *Rehabilitation Counseling Bulletin*, 51(4), 251-259. Doi:10.1177/0034355208317317
- [30]. Vigneau, F., & Bors, D. A. (2005). Items in context: Assessing the dimensionality of raven's advanced progressive matrices. *Educational and Psychological Measurement*, 65(1), 109-123. Doi:10.1177/0013164404267286
- [31]. Boone, W. J. (2016). Rasch Analysis for Instrument Development: Why, When, and How? *CBE Life Sci Educ*, 15(4). Doi:10.1187/cbe.16-04-0148
- [32]. Shea, T., Cooper, B. K., De Cieri, H., & Sheehan, C. (2012). Evaluation of a perceived organisational performance scale using Rasch model analysis. *Australian Journal of Management*, 37(3), 507-522. Doi:10.1177/0312896212443921
- [33]. Bond, T. G., & Fox, C. M. (2007). *Applying the Rasch Model: Fundamental Measurement in the Human Sciences, Second Edition (2nd ed.)*. Psychology Press. Doi:10.4324/9781410614575
- [34]. Saad, R., Yusuff, R. Z., Abas, Z., Aziz, A. A., & Masodi, M. S. (2011). Validating The ISO 9000 Construct of Measurement Instrument Through Application of RASCH Model. *The Asian Journal of Technology Management*, 4(1), 28-40.
- [35]. Coe, R. (2008). Comparability of GCSE examinations in different subjects: an application of the Rasch model. *Oxford Review of Education*, 34(5), 609-636. Doi:10.1080/03054980801970312
- [36]. Wilmot, D. B., Schoenfeld, A., Wilson, M., Champney, D., & Zahner, W. (2011). Validating a Learning Progression in Mathematical Functions for College Readiness. *Mathematical Thinking and Learning*, 13(4), 259-291. Doi:10.1080/10986065.2011.608344
- [37]. Aziz, A. A., Mohamad, A., Arshad, N., Zakaria, S., Ghulman, H. A., & Masodi, M. (2008). Application of Rasch Model in validating the construct of measurement instrument. *International Journal of Education and Information Technologies*, 2(2), 105-112.
- [38]. Jones, R. N. (2019). Differential Item Functioning and its Relevance to Epidemiology. *Current Epidemiology Reports*, 6(2), 174-183. Doi:10.1007/s40471-019-00194-5
- [39]. Bond, T. G., & Fox, C. M. (2015). *Applying the Rasch Model: Fundamental Measurement in the Human Sciences, Third Edition (3rd ed.)*. Routledge. Doi:10.4324/9781315814698
- [40]. Ferne, T., & Rupp, A. A. (2007). A Synthesis of 15 Years of Research on DIF in Language Testing: Methodological Advances, Challenges, and Recommendations. *Language Assessment Quarterly*, 4(2), 113-148. Doi:10.1080/15434300701375923
- [41]. Tennant, A., & Pallant, J. (2007). DIF matters: A practical approach to test if differential item functioning makes a difference. *Rasch Measurement Transactions*, 20(4), 1082-1084.
- [42]. Ackerman, T. A. (1992). A didactic explanation of item bias, item impact, and item validity from a multidimensional perspective. *Journal of Educational Measurement*, 29, 67-91. Doi:10.1111/j.1745-3984.1992.tb00368.x
- [43]. Martinkova, P., Drabinova, A., Liaw, Y.-L., Sanders, E. A., McFarland, J. L., Price, R. M., & Nehm, R. (2017). Checking equity: Why differential item functioning analysis should be a routine part of developing conceptual assessments. *CBE Life Sciences Education*, 16(2), 1-13. Doi:10.1187/cbe.16-10-0307
- [44]. Westers, P., & Kelderman, H. (1992). Examining differential item functioning due to item difficulty and alternative attractiveness. *Psychometrika*, 57(1), 107-118. Doi:10.1007/BF02294661
- [45]. Osterlind, S. J. (1983). *Test item bias*. Sage Publication Inc.
- [46]. Polit, D. F., & Beck, C. T. (2006). The content validity index: Are you sure you know what's being reported? critique and recommendations. *Research in Nursing & Health*, 29(5), 489-497. Doi:10.1002/nur.20147
- [47]. Etikan, I., Musa, S. A., & Alkassim, R. S. (2016). Comparison of convenience sampling and purposive sampling. *American Journal of Theoretical and Applied Statistics*, 5(1), 1-4. Doi:10.11648/j.ajtas.20160501.11
- [48]. Linacre, J. M. (2017). Teaching Rasch Measurement. *Rasch Measurement Transactions*, 31(2), 1639-1640.
- [49]. Boone, W. J., Staver, J. R., & Yale, M. S. (2014). *Rasch analysis in the human sciences*. Springer. Doi:10.1007/978-94-007-6857-4
- [50]. Wright, B. D., & Linacre, J. M. (1994). Reasonable mean-square fit values. *Rasch Measurement Transactions*, 8(3), 370.
- [51]. Aziz, A. A., Masodi, M. S., & Zaharim, A. (2013). *Asas model pengukuran rasch: pembentukan skala dan struktur pengukuran*. Universiti Kebangsaan Malaysia.
- [52]. Azrilah, A. A., Mohd Saidfudin, M., & Azami, Z. (2013). *Asas Model Pengukuran Rasch: Pembentukan Skala & Struktur Pengukuran*. Universiti Kebangsaan Malaysia.
- [53]. Raïche, G. (2005). Critical eigenvalue sizes (variances) in standardized residual principal components analysis (PCA). *Rasch Measurement Transactions*, 19(1), 1012.

- [54]. Linacre, J. M. (2002). What do infit and outfit mean-square and standardized mean? *Rasch Measurement Transaction*, 16(2), 878.
- [55]. Paek, I., & Holland, P. (2015). A Note on Statistical Hypothesis Testing Based on Log Transformation of the Mantel–Haenszel Common Odds Ratio for Differential Item Functioning Classification. *Psychometrika*, 80(2), 406-411. Doi:10.1007/s11336-013-9394-5
- [56]. Sarstedt, M., & Mooi, E. (2019). Regression analysis. *A Concise Guide to Market Research: The Process, Data, and Methods Using IBM SPSS Statistics*, 209-256.
- [57]. Wells, C. S., & Wollack, J. A. (2003). *An instructor's guide to understanding test reliability*. University of Wisconsin.
- [58]. Boone, W. J., & Noltemeyer, A. (2017). Rasch analysis: A primer for school psychology researchers and practitioners. *Cogent Education*, 4(1), 1416898. Doi:10.1080/2331186X.2017.1416898
- [59]. Chang, K.-C., Wang, J.-D., Tang, H.-P., Cheng, C.-M., & Lin, C.-Y. (2014). Psychometric evaluation, using Rasch analysis, of the WHOQOL-BREF in heroin-dependent people undergoing methadone maintenance treatment: further item validation. *Health and Quality of Life Outcomes*, 12(1), 148. Doi:10.1186/s12955-014-0148-6