PSYMPATHIC : Jurnal Ilmiah Psikologi
Vol. 9, No. 1, 2022: 103-116
eISSN: 2502-2903, pISSN: 2356-3591
DOI: 10.15575/psy.v9i1.14977

Quality Analysis of Intelligence Structure Test 2000 Revision (IST 2000R) Items in Indonesian

Medianta Tarigan[1]*, Fadillah[2]
[1]Department of Psychology, Universitas Pendidikan Indonesia, Bandung, Indonesia
[2]Department of Visual Communication Design, Institut Teknologi Bandung, Indonesia
*e-mail: medianta@upi.edu

Abstract

This study aimed to test the quality of item analysis in the Indonesian IST 2000R using the Item Response Theory (IRT) method and participants comprising 1780 students aged 16-20. IRT 3PL analysis and DIF examination showed that the Indonesian IST 2000R had good item quality, with an improvement in IST 70. The discriminatory power parameter (ɑ) of the verbal category items was high. Furthermore, the difficulty level parameter (b) indicated that more than 50% of the items on the IST 2000R are in the average category. In the guessing probability parameter (c), 8 of 9 subtests showed that more than 50% of the items are in a low category. Overall, the Indonesian IST 2000R has a small gender bias because all subtests obtained more than 60% in category A.

Keywords: Intelligence Structure Test (IST), Item Response Theory, item analysis

Abstrak

Penelitian ini bertujuan untuk melakukan uji analisis kualitas item pada IST 2000R berbahasa Indonesia dengan metode *Item Response Theory* (IRT). Partisipan penelitian adalah mahasiswa (usia 16-20 tahun, N=1780 subjek). Dengan analisis IRT 3PL dan pemeriksaan DIF, diperoleh bahwa IST 2000R Bahasa Indonesia memiliki kualitas butir item yang baik dan ada perbaikan terhadap kualitas IST 70. Parameter daya beda (ɑ) item kategori verbal tergolong tinggi. Berdasarkan parameter tingkat kesulitan (b), lebih dari 50% item pada IST 2000R berada pada kategori rata-rata. Pada parameter peluang menebak (c), 8 dari 9 subtes menunjukkan hasil bahwa lebih dari 50% item berpeluang menebak rendah. Secara keseluruhan, IST 2000R versi Bahasa Indonesia memiliki bias jenis kelamin yang kecil karena seluruh subtes mendapatkan persentase lebih dari 60% pada kategori A.

Kata Kunci: *Intelligence StructureTest* (IST), Teori Respon Butir, analisis item

Introduction

Intelligence Structure Test (IST) is an intelligence measuring tool created and developed by Rudolf Amthauer in 1953, with the earliest version named IST 70. This tool was adapted in Indonesia in 1971 by the Faculty of Psychology, University of Padjadjaran (Adinugroho, 2016). Practitioners in Psychology often use IST to measure individual intelligence in education, industry, organization, or clinical fields. It is an intelligence test preferred by practitioners because it measures more intelligence aspects than other test tools such as CFIT and Raven's Progressive Matrics. Additionally, IST provides information on verbal, numerical, and figural intelligence abilities rarely obtained comprehensively using other intelligence measuring tools.

In Indonesia, studies to test the psychometric properties of IST 70 are conducted to determine its feasibility. The quality of items on the Indonesian IST 70 version has been tested through classical test theory analysis. The results showed that almost half the items in each IST subtest need improvement. This is because IST 70's discrimination index is not very good, where many items cannot distinguish between subjects with high and low abilities (Sirodj, 2018). Furthermore, a study tested the quality of the IST 70-item questions using IRT analysis. The results showed that 53.125% of the IST items analyzed had poor characteristics according to psychometric limits (Rahmawati, 2014).

Several other studies also examined the quality of the items for each IST 70 subtest. An examination of the quality of Figurenauswahl (FA) subset items found the need for revisions regarding answer choices because distractors cannot outwit individual responses (Adinugroho, 2016). A review of the IST 70 Rechnaufgaben (RA) and Zahlenreihen (ZR) subtest found that the items' quality is quite decent. However, 17.5% of the total items have a measurement bias (Tarigan & Fadillah, 2021a). Investigating the quality of the IST verbal subtest showed that 71.67% of the 60 items had fairly good quality and an acceptable estimation of different power. This shows that the IST 70 is valid but requires revising 25% of the items with a measurement bias (Tarigan & Fadillah, 2021b).

The previous psychometric test studies recommended revising or improving the IST 70 items used in Indonesia. Other things have also raised concerns among practitioners in using this measuring instrument. The IST 70 has been the field's most widely used intelligence measuring instrument since nearly 50 years ago. Therefore, some experts considered that this test tool has reached obsolescence or expired conditions. For instance, the question items in the Wortauswahl (WA) subtest regarding verbal insights was irrelevant. The participants' access to IST 70 questions on the internet and the test takers' familiarity with the items are also a concern. It raises the question of whether the participants answered using their knowledge and intelligence or based on the information in the widely circulated answers. Therefore, a new alternative test tool is needed to measure individual intelligence comprehensively.

IST has been developed with version 2000R in Germany and adapted into English (Beauducel et al., 2010). Previous studies re-examined the IST 2000R construct by analyzing the relationship between items and each construct domain. The results showed that the IST 2000R was better than IST 70 (Beauducel et al., 2001; Schulze et al., 2005). IST 2000R has three modules, with the main module measuring verbal, numerical, and figural aspects. In the latest version, each aspect is represented by three subtests, making it more proportional. However, no study tested the strength of the Indonesian IST 2000R psychometric properties regarding item quality analysis.

An analysis is important in observing the items' characteristics and improving the test quality (Quaigrain & Arhin, 2017). The test could use Item Response Theory (IRT) analysis, a modern approach to analyzing items from a test kit (Sadhu & Laksono, 2018). The Item Characteristic Curve used in IRT shows the relationship between the subject's ability and the item's quality (Adedoyin & Mokobi, 2013). IRT analysis obtains information on item parameters according to the Logistics Parameters (PL) used. The parameters include item discriminatory power, difficulty level, and guessing probability (McGrory et al., 2014). This study used item analysis of 3 Logistics Parameters (3PL). According to Obinne (2012), item analysis estimates the relationship between the probability of the correct answer to the item and the individual's ability. Therefore, IRT analysis was used to determine the discriminatory power, difficulty level, and probability of guessing the IST 2000R items.

Item bias occurs when the test is inconsistent, unfair, and influenced by factors other than the ability to be tested. This study investigated whether gender is a factor that causes bias. There is no bias identification in other aspects, such as education level and age, because they are considered variables related to intelligence. In IST, the norms are arranged based on age and education level. A Differential Item Functioning (DIF) bias analysis was performed to identify item bias in the IST 2000R. The DIF analysis results provided recommendations for reviewing biased items. Therefore, this study aimed to conduct item analysis on the IST 2000 R intelligence test using the IRT method.

Methods

Design

The IST 2000R instrument was adapted into Indonesian and tested for item validity using IRT analysis.

Participants

This study used 1780 participants, comprising 791 male and 989 female students from State University undergraduate education levels. They were aged between 16-20 and spread across Medan, Padang, Pekanbaru, Bogor, Jakarta, Bandung, Malang, Semarang, and Yogyakarta Cities. The sample was selected because it understands the instructions on cognitive instruments. Rewards were given to participants after completing all IST 2000R questions. Data from 2018 to 2020. Table 1 shows the distribution of participant data.

Instruments

The measuring instrument used is the IST 2000R, comprising the verbal, numerical, and figural aspects. The aspects contain three subtests each, and every subtest has 20 questions. The IST 2000R instrument was adapted into Indonesian with a back-to-back translation adaptation process through several stages. In the first stage, the measuring instrument was translated by two people.

Table 1
Overview of Subjects

Category	Total	Percentage
Gender		
Male	791	44%
Female	989	56%
City		
Bandung	388	22%
Bogor	197	11%
Depok	195	11%
Malang	177	10%
Medan	118	7%
Padang	59	3%
Pekanbaru	55	3%
Semarang	194	11%
Yogyakarta	393	22%

The first person was an informed translator knowledgeable of the IST 2000R instrument. The second person translated the IST 2000R instrument without knowing its concept. The second stage was synthesis, where the translation results were further processed to be compared with the word of which the meaning was closest to the original. In the third stage, the items translated into Indonesian were translated back into English, followed by a review from three linguists. The final stage considered the spoken language's suitability to determine the cultural differences between the two translators.

The test time was 77 minutes. Eight of the nine subtests on the IST 2000R were multiple choice questions. In the Numerical Calculations (CA) subtest, subjects wrote their answers freely. After testing, scoring was conducted according to the provisions of each subtest in the IST 2000R module. Table 2 explains the subtests in question.

Study Procedure

Data were collected through offline and online methods. In the offline method, data were administered according to the module using the IST 2000R. The process was conducted by a tester, a psychologist who attended a workshop on the use of the IST 2000R. A minimum of 2 psychologists administered the test with a maximum of 20 participants per session. Furthermore, the online method was conducted through a web-based test platform (https://psikotes.anargya.id/). In this case, the test was designed and developed concerning the provisions of the IST 2000R module. There was no difference in scoring and interpretation between the two methods. Preliminary trials were conducted to ensure that online and offline data produced identical scores and final results. During online data collection, proctoring tests were also performed concerning the Proctoring & Security System standards recommended by The Association of Test Publishers and The National College Testing Association.

Table 2
IST 2000R Subtests

Aspects		Subtests	Description	Number of Questions	Question Form	Answer Score
Verbal	1.	*Sentence Completion* (SC)	Contains sentences with one word missing	20	*Multiple choice*	True (1)/ False (0)
	2.	*Verbal Analogies* (VA)	The relationship between two words and finding words with a similar relationship	20	*Multiple choice*	True (1)/ False (0)
	3.	*Similarities Subtest* (VS)	Presents six-word groups to find two words with the same term	20	*Multiple choice*	True (1)/ False (0)
Numerical	4.	*Numerical Calculations* (CA)	It contains arithmetic tasks with real numbers	20	*Free Answer*	True (1)/ False (0)
	5.	*Number Series* (NS)	Presents numbers formed according to a certain pattern and are asked to continue the pattern	20	*Multiple choice*	True (1)/ False (0)
	6.	*Numerical Signs* (SI)	Choosing the correct mathematical operators for equations	20	*Multiple choice*	True (1)/ False (0)
Figural	7.	*Figure Selection* (FS)	Geometric shapes are presented with several pieces resulting from cutting one shape to identify all the shapes built from the pieces	20	*Multiple choice*	True (1)/ False (0)
	8.	*Cubes* (CU)	Identify the rotated cube	20	*Multiple choice*	True (1)/ False (0)
	9.	*Matrices* (MA)	Presented a set of images arranged according to certain rules	20	*Multiple choice*	True (1)/ False (0)

Data Analysis

Data were analyzed by presenting the descriptive analysis results to describe the participant data. The item quality testing procedure used IRT. The IRT statistical tool is a response model to educational and psychological test items and latent traits that determine how individuals respond to those items (Foster et al., 2017). Presently, 4 IRT models are quite popular (Ogunsakin & Shogbesan, 2018), namely (1) 1 Parameter Logistics (1PL) only describes the discriminatory power in items; (2) 2PL describes discriminatory power and item difficulty level; (3) 3PL describes discriminatory power, difficulty level, and guessing probability on items; and (4) 4PL describes the item's discriminatory power, difficulty level, guessing probability, and carelessness probability.

This study used item quality analysis with an IRT approach with 3PL, including discriminatory power, difficulty level, and guessing probability. According to Martín et al. (2006), the 3PL model could be used for binary data (true or false). It is not a test that applies a score reduction to questions answered incorrectly. Moreover, the 3PL model is recommended in multiple-choice performance speed tests. This is because the tendency of participants to guess the answers is higher, requiring the third parameter (Ogunsakin & Shogbesan, 2018). Therefore, the IRT 3PL analysis was applied to all subtests, except the independent response Numerical Calculations (CA). This subtest used the IRT 2PL model that allows the items' discriminatory power to be known and is suitable for use on independent response items. To determine and ensure the best parameter model for each subtest, an Akaike Information Criteria (AIC) analysis was conducted. This was followed by analyzing the quality of the items for each subtest based on the smallest AIC value among the three models tested (Ayanwale, 2019).

Several criteria were used to interpret the IRT model. For the discriminatory power (ɑ), a score of 0-2 was included in the normal category. Scores of less than -2, -2 to 2, and more than 2 were classified as low, average level, and difficult, respectively, for the difficulty level (b). For the guessing probability (c), scores of 0 - .35 and above .35 were in the acceptable and unacceptable categories, respectively (Baker, 2001). The IRT results also showed the item characteristic curve (ICC) to facilitate data interpretation.

After analyzing the IRT items, a DIF analysis was conducted to determine the possibility of a measurement bias on these items. DIF was conducted based on the participants' sex because males and females have significant differences in behavioral perceptions. Males think that their performance is significantly better than others. In contrast, females consider their performance equal to their female counterparts (Ring et al., 2016). Reilly et al. (2022) found that sex role identification significantly contributed to intellectual self-image. Masculine personality traits increase self-esteem that significantly and independently affects self-intelligence estimates. The DIF parameter index category referred to Gierl et al. (2001). In this case, DIF categories A, B, and C are ignored, moderate, and high when the null hypothesis is rejected with $| \beta | < .059$, $.059 \leq | \beta | < .088$, and $| \beta | \geq .088$, respectively. The data were analyzed using Jmetric software because it is user-friendly and performs IRT analysis up to 4PL model.

Results and Discussion

Result

Descriptive Analysis

The data were cleaned to remove outliers that could interfere with the analysis process. Therefore, 1780 data were processed for further analysis. Descriptive analysis was performed first to describe the data as a whole. The analysis results in Table 3 showed that the CA subtest has the highest average score of 18.954. This means the average score of participants in this subset is quite high. Of the nine IST 2000R subtests, CA, SI, and FS have the highest scores. The standard deviation value most away from 0 is in the FS subtest, implying heterogeneous or diverse data. Moreover, the NS, CU, and MA subtests have the largest value range of 20. It implies a considerable distance between the highest and lowest scores obtained by the participants. On the skewness value, the CA subtest has a fairly extreme negative value of -2.374. The data distribution with a negative slope has a longer left tail in the negative direction (Cain et al., 2017). This shows more items with values above the median, making the curve tail longer to the left.

Before calculating the IRT model parameters to determine the items' quality, an *Akaike Information Criterion (*AIC) analysis was conducted to determine the most suitable IRT model. The analysis was based on the smallest AIC value among the three models tested (Ayanwale, 2019). The recommendations from the AIC analysis in Table 4 showed that almost all subtests are in the IRT 3PL model. The CA subtest is an exception, which uses IRT 2PL for item quality analysis.

Table 5 shows that the calculation recapitulation of each parameter for items in all subtests meets Baker's (2001) criteria. In the test category measuring verbal aspects, 95% or 57 items were indicated according to the item discriminatory power parameter. Only three items need reviewing because they contradict the discriminatory power parameter index. All items in the VA subtest have an appropriate discriminatory power index. Analyzing the items' difficulty levels on the subtests measuring verbal ability showed that easy, average, and difficult items are spread proportionally with a total of 18.3%, 48.3%, and 33.4%, respectively. Based on the guessing probability index on the verbal ability subtest, only 7 of 60 items need to be studied because they do not fulfill the criteria. The other 88.33% have item

Table 3
Descriptive Analysis Results

Subtests	Valid	Mean	Median	Mode	SD	Skewness	Kurtosis	Range	Min	Max	Sum
SC	1780	9.3831	9	8	2.373	.113	-.198	17	1	18	16702
VA	1780	11.0483	11	11	1.961	-.09	.146	15	3	18	19666
VS	1780	11.382	12	12	2.49	-.939	1.866	16	1	17	20260
CA	1780	18.577	19	20	1.733	-2.374	9.278	17	3	20	33067
NS	1780	16.3253	17	19	3.536	-1.147	.905	20	0	20	29059
SI	1780	16.7028	17	20	3.133	-1.046	.912	19	1	20	29731
FS	1780	14.5084	15	20	4.212	-.434	-.765	18	2	20	25825
CU	1780	12.2697	13	13	3.463	-.485	.199	20	0	20	21840
MA	1780	12.6017	13	13	2.764	-.134	.568	20	0	20	22431

Table 4
AIC results

Subtests	AIC			Recommendation IRT Model
	1PL	2PL	3PL	3PL
SC	2636.143	1244.92	1174.649	3PL
VA	2359.335	805.0756	793.432	3PL
VS	2305.576	1659.464	1651.074	3PL
CA	1155.846	1342.811	-	2PL
NS	1868.016	2797.706	2764.938	3PL
SI	1824.675	2497.339	2495.513	3PL
FS	2306.3	2958.651	2778.424	3PL

Table 5
Item Quality Recapitulation Based on IRT 3PL in Verbal Subtests

	Discriminatory power (α) $0 \le a_i \le 2$		Difficulty level (b) $-2 \le b_i \le 2$			Guessing Probability (c) $0 \le c_i \le .35$	
Subtests	Appropriate	Need Improvement	Easy	Average	Difficult	Appropriate	Need improvement
SC	19	1	3	11	6	18	2
	(95%)	(5%)	15%	55%	30%	90%	10%
VA	20	0	6	6	8	17	3
	(100%)	(0%)	30%	30%	40%	85%	15%
VS	18	2	2	12	6	18	2
	(90%)	(10%)	10%	60%	30%	90%	10%
Total	57 (95%)	3 (5%)	11 (18.3%)	29 (48.3%)	20 (33.4%)	53 (88.33%)	7 (11.67%)

quality according to the criteria, meaning they are not easily guessed by the participants.

The numerical category results in Table 6 show that the quality of 83.3% of the items is consistent with the parameters because they have a discriminatory power index of less than .00 and more than 2.00. The remaining 16.7% require improvement. Most items in the NS subtest need reviewing. Similarly, all items in the CA subtest have a discriminatory power index according to the parameters. Regarding the difficulty level, no item is indicated as difficult, 38.3% are easy, and 61.7% are average. The highest number of easy items is in the CA subtest. In the guessing probability index, 60% of the 20 items in the NS and SI subtests are consistent with the parameter index. Therefore, 40% of items need reviewing because they contradict the recommended guessing probability index parameter.

The analysis results of the discriminatory power parameter in the figural category are shown in Table 7. The discriminatory power of almost all items in the CU subtest is consistent with the parameter index, and only

Table 6
Item Quality Recapitulation based on 3PL IRT in Numerical Subtests

	Discriminatory power (ɑ) $0 \leq a_i \leq 2$		Difficulty level (b) $-2 \leq b_i \leq 2$			Guessing Probability (c) $0 \leq c_i \leq .35$	
Subtests	Appropriate	Need improvement	Easy	Average	Difficult	Appropriate	Need improvement
CA	20	0	16	4	0	-	-
	100%	0%	80%	20%	0%		
NS	14	6	3	17	0	12	8
	70%	30%	15%	85%	0%	60%	40%
SI	16	4	4	16	0	12	8
	80%	20%	20%	80%	0%	60%	40%
Total	50 (83.3%)	10 (16.7%)	23 (38.3%)	37 (61.7%)	0 (0%)	24 (60%)	16 (40%)

Table 7
Item Quality Recapitulation based on 3PL IRT in Spatial Figural Subtests

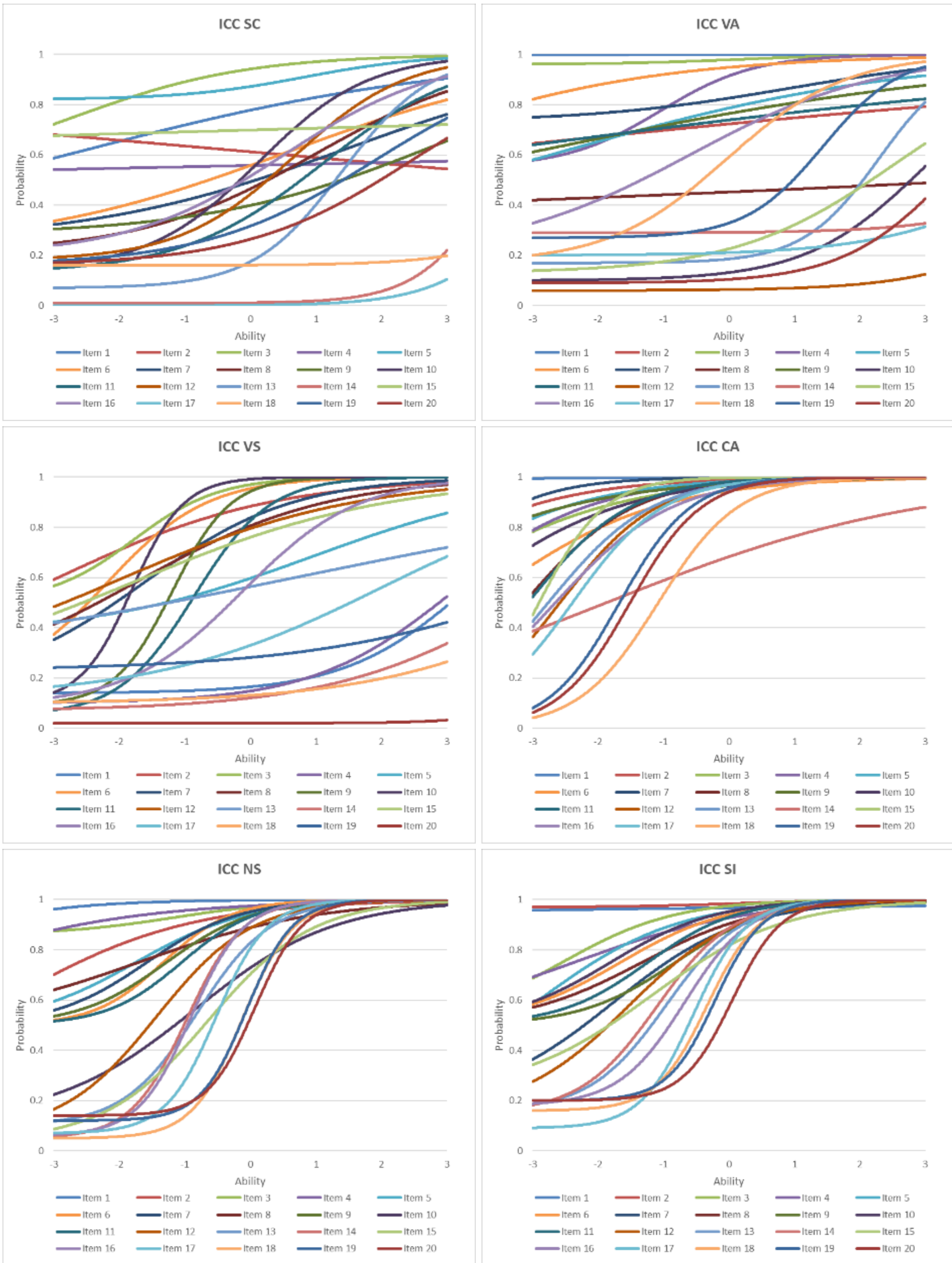
	Discriminatory power (ɑ) $0 \leq a_i \leq 2$		Difficulty level (b) $-2 \leq b_i \leq 2$			Guessing Probability (c) $0 \leq c_i \leq .35$	
Subtests	Appropriate	Need improvement	Easy	Average	Difficult	Appropriate	Need improvement
FS	11	9	0	20	0	14	6
	55%	45%	0%	100%	0%	70%	30%
CU	19	1	1	19	0	18	2
	95%	5%	5%	95%	0%	90%	10%
MA	15	5	4	14	2	18	2
	75%	25%	20%	70%	10%	90%	10%
Total	45 (75%)	15 (25%)	5 (8.33%)	53 (88.33%)	2 (3.34%)	50 (83.33%)	10 (16.67%)

1 item has an index outside the range of 0-2. Overall, 25% of the 60 items need reviewing. Regarding the difficulty level, 88.33%, 8.33%, and 3.34% are average, easy, and considered difficult, respectively. Only the MA subtest has two items with a high difficulty index. All items in the FS subtest belong to the average difficulty level. Moreover, 90% of items in the CU and MA subtests are consistent with the guessing probability index. Six items in the FS subtest need reviewing because they have a probability of being guessed.

Information on the distribution of item quality is shown in an item characteristic curve (ICC) graph. The graph describes the item characteristics to help understand the item quality analysis results in each subtest.

Figure 1 shows the visualization of IRT parameters for each IST item. In the SC subtest, item 2 has a curve different from the others. The curve forms a descending straight line, indicating a negative discriminant value of -.12. This means that the item cannot distinguish the participants' abilities. Participants with low abilities are more likely to answer correctly than those with high. Furthermore, item 5 has the highest guessing value. Participants with low abilities have a high probability of answering correctly, close to 1. In the VA subtest, items 1 and 3 have a higher constant probability than others. The curves of these two items form a continuous straight line at a probability close to 1. Item 1 has a low difficulty level, meaning participants with low to high abilities are possible to answer the question correctly. This is also supported by the extreme difficulty parameter value of -12.73. In item 3, the guessing value is very high at .96, meaning the item could be easily guessed by participants.

In the VS subtest, items 2 and 3 have a curve with a higher probability value than the others. The probability value is close to .6 at the ability level of -3. Items 2 and 3 need further analysis because they have a low difficulty level and excessivel yhigh a guessing probability of -2.97 and .5, respectively.


ICC SC
ICC VA
ICC VS
ICC CA
ICC NS
ICC SI
Probability
Ability
Item 1 Item 2 Item 3 Item 4 Item 5
Item 6 Item 7 Item 8 Item 9 Item 10
Item 11 Item 12 Item 13 Item 14 Item 15
Item 16 Item 17 Item 18 Item 19 Item 20

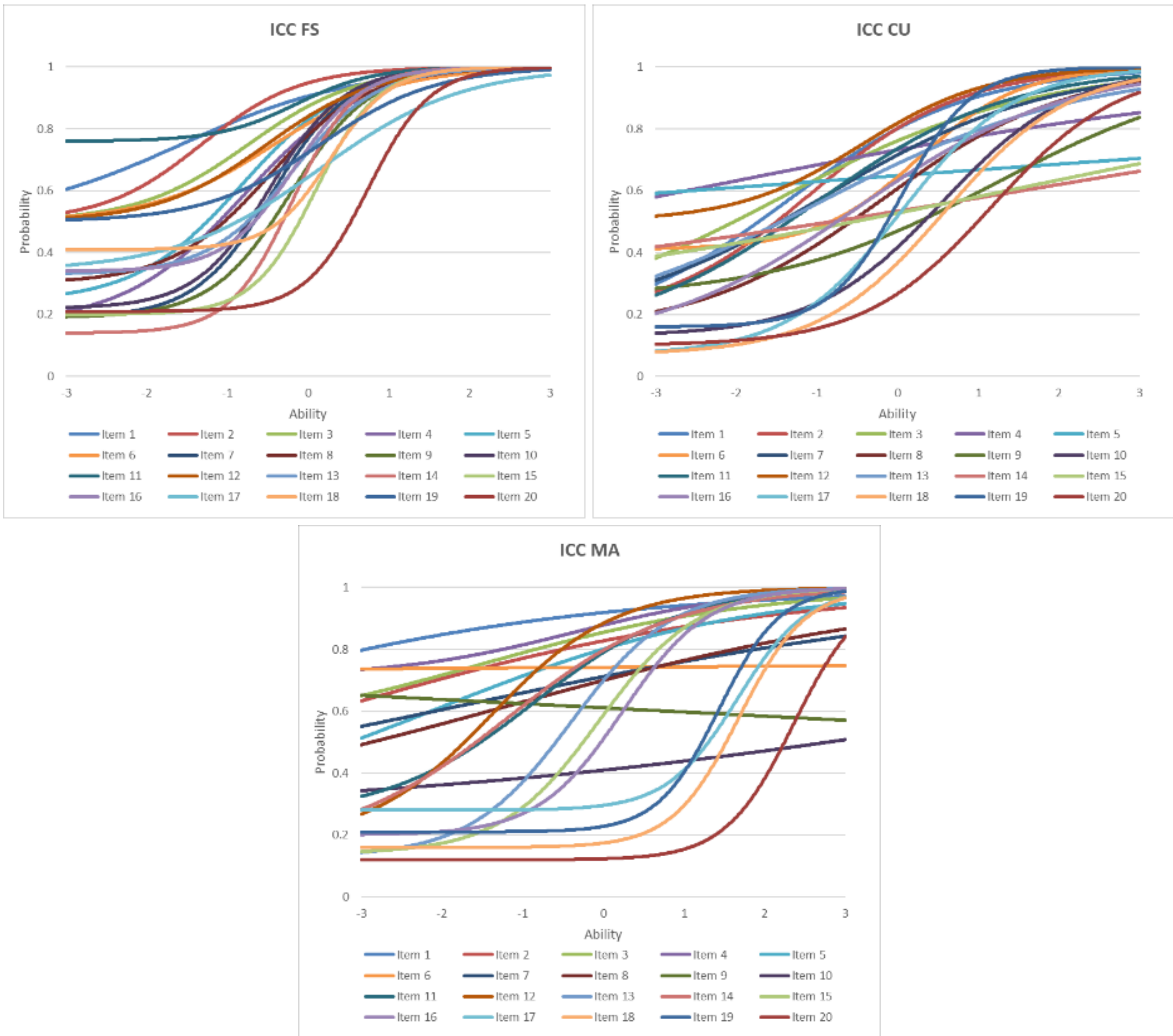


Figure 1. Item characteristic curve (ICC)

In the CA subtest, item 1 has a flat curve with a constant probability value close to 1. This item needs reviewing because it has a very low difficulty level. Participants with low and high abilities are more likely to answer the question correctly. This is also seen at the extreme difficulty parameter value of -11.04. In the NS subtest, item 1 has a higher graphic pattern than others, meaning it has a very low difficulty level of -5.67. Therefore, they could be answered correctly by all participants, including those with low abilities.

In the NS subtest, items 1 and 2 have a flat graph with a high probability value close to 1. It means these two items could be answered correctly by participants with low to high abilities. Item 1 has a very low difficulty level of -34.61, while item 2 has a high guessing probability of .97. In the FS subtest, item 11 has a higher probability than others, with a curve forming a sloping S pattern. It means that item 11 has a high guessing power, and participants with low abilities have a high probability of close to .8 to answer correctly.

In the CU subtest, item 5 has a different and more sloping curve than others. The item has a fairly low discriminatory power of .1, an average difficulty level of -1.52, and a guessing probability of .24. In the MA subtest, item 9 has a curve that forms a descending straight line, indicating a negative discriminant value of -.07. It means this item

Table 8
Recapitulation of DIF IST 2000R Results

Subtests	Category	Total Item	Percentage	Question Number
SC	A	13	65%	1, 2, 3, 4, 7, 8, 11, 12, 14, 16, 17, 18, 19
	B	6	30%	5, 6, 9, 10, 15, 20
	C	1	5%	13
VA	A	19	95%	1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 20
	B	1	5%	19
VS	A	18	90%	1, 2, 4, 5, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20
	B	2	10%	3, 6
CA	A	18	90%	1, 2, 3, 4, 5, 6, 7, 8, 9, 11, 12, 14, 15, 16, 17, 18, 19, 20
	B	2	10%	10, 13
NS	A	20	100%	all items
SI	A	20	100%	all items
FS	A	20	100%	all items
CU	A	20	100%	all items
MA	A	19	95%	1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 13, 14, 15, 16, 17, 18, 19, 20
	B	1	5%	12

Description: Category A (ignored) $|\beta| < .059$; Category B (moderate) $.059\ |\beta| < .088$; Category C (high) when $|\beta| \geq .088$

cannot distinguish participants' abilities. Participants with low abilities are more likely to answer the question correctly than those with high, meaning the item should be re-examined.

The items in the VS, CA, NS, SI, FS, CU, and MA subtests have good parameters because they form a rising curve. This shows that the items' quality is quite good. The higher the participants' ability, the greater their probability of answering correctly. Furthermore, each item has good discriminatory power because it distinguishes between participants with low and high abilities. The items are difficult to guess and average to the difficult difficulty level.

Differential Item Functioning (DIF) Analysis

The DIF statistical technique is based on the principle that variables outside the measured construct do not affect the measurement results. This study conducted the DIF analysis to check for an item bias generated by the influence of the gender variable. Table 8 shows the DIF analysis results.

The results showed that 35% or seven items in the SC subtest indicated bias, six were biased with moderate significance, and 1 had a high bias category. The VA, VS, CA, and MA subtests contain 1-2 items with moderate bias. All items in the NS, SI, FS, and CU subsets have insignificant bias and were ignored. Therefore, 92.7% of the items in IST 2000R were not infected with DIF due to the influence of the gender variable.

Discussion

The analysis using the IRT 3PL method showed that the Indonesian version of the IST 2000R has better quality items than the IST 70. The item discriminatory parameter analysis (ɑ) showed that 84% of the items on the IST 2000R have an appropriate discriminatory power parameter index. In contrast, 16% of the items were inappropriate and needed further analysis. The item quality analysis with the difficulty level parameter (b) showed that 21.67%, 66.11%, and 12.2% of items had low, moderate, and high difficulty levels, respectively. The analysis showed that 79.37% and the remaining 20.62% of the items have an appropriate and incorrect guessing probability index, respectively.

Previous studies analyzed the quality of the Indonesian language adaptation item IST 70. In comparison, the IST 2000R has a better item eligibility condition, specifically on the discriminant index or the ability to distinguish between individuals. The discriminant index (ɑ) of all items on the verbal, numerical, and figural subtest IST 70 showed that only 45.72% is good, while the remaining 54.28%

needs reviewing or improvement (Sirodj, 2018). Items with low discriminatory power usually occur for several reasons. First, the items may not function similarly in measuring psychological constructs. Second, the use of sentences on items may not be appropriate. Third, the questions given could be too complex for certain educational backgrounds. Fourth, the items could be made of facets. Fifth, there may be a cultural bias on the items when used in certain groups (Cohen & Swerdlik, 2010). However, the IST 2000R has 21.67% items with low difficulty levels, and 20.62% were easy to guess. Questions with a low difficulty level and easy to be guessed affect the measuring instruments' ability to distinguish individual characteristics. Some items could be answered by participants with low and high abilities.

This study compared the quality of the subtests that measure verbal, numerical, and spatial figural aspects. Based on the ability to distinguish items, 57 items in the subtest that measures the verbal aspect match the discriminatory power index. This contradicts studies on IST 70, where items with good discriminatory power are more common in subtests that measure numerical aspects (Sirodj, 2018). In the subtest measuring the verbal aspect, 95% of all items on the IST 2000R verbal subtest have appropriate discriminatory power, and only 5% need improvement. This contradicts the item quality analysis at IST 70, where 48.33% of verbal subtest items need improvement (Sirodj, 2018). Other results also showed that all items on the verbal subtest IST 70 have good quality according to the discriminatory power index (Tarigan & Fadillah, 2021b).

The IST 2000R subtest measuring the spatial figural aspect has more items with the best difficulty level than subtests in other aspects. Overall, 88.33%, 8.33%, and 3.34% of items on the figural subtest have average, easy, and difficult levels, respectively. The difficulty level in the subtest measuring the numerical aspect is spread out, but 38.3% of the items are considered easy. This contradicts IST 70, where two items are easy while seven are in the difficult range. Tarigan & Fadillah (2021a) found that 12.5% of items are outside the difficulty level index range.

In the study of IST 2000R, the numerical aspect subtest has the most items to be revised due to an inappropriate guessing probability index. The IST 70 analysis showed that the guessing index on this subtest does not indicate items that need improvement. All items have a good guessing probability index (Sirodj, 2018). Overall, 39 items in the IST 70 numerical subtest have a correct guessing probability index, and only 1 item needs revision (Tarigan & Fadillah, 2021a). The difficulty level and the guessing probability between the subtests measuring different aspects of the IST 2000R indicate differences in quality that need observation. Similarly, there are distinctive characteristic differences between IST 2000R and IST 70 regarding each parameter index problem.

The participants' abilities and the guessing probability could also be studied based on visualizing the ICC. When the slope is large, the curve is steeper and tends to rise. This indicates that high-ability participants are more likely to answer correctly. The probability is smaller for low-ability participants and vice versa (Yacob et al., 2014). Therefore, when the curve gives a high probability value but low ability, the answer to the item is easy to guess, characterized by a relatively easy difficulty level. This reduces the discriminatory power because the item cannot distinguish between high and low-ability participants. The ICC results showed that the items in the VS, CA, NS, SI, FS, CU, and MA subtests have good parameters because they form a rising curve. This also shows that the items' quality is quite good. The higher the participants' ability, the greater their probability of answering the items correctly in the subtests. Furthermore, the results on the curve indicated that the items have good discriminatory power because they distinguish between high and low-ability participants. The items are also difficult to guess and have difficulty levels from average to difficult.

This study also analyzed the data using the DIF statistical technique. The technique is based on the principle that different test takers with the same knowledge level should have similar results on individual test items regardless of group membership (Ibrahim et al., 2018). The bias in some items indicated the possibility of differences in skills in women and men. Many studies showed that women are better than men in verbal skills and obtain higher scores on math tests suited to coursework. Men outperform women in questions of geometry, arithmetic reasoning, and algebra (Abedalaziz et al., 2014). Furthermore, masculine and feminine sex roles contribute to cognitive development. In this case, masculinity predicts visual-spatial performance (Reilly & Neumann, 2013). The criteria used to detect bias in this study is when the Chi-square value is significant (p-value $\leq$ .05) (Meyer, 2014). The positive sign in category B or C indicates that the item is easier for the focal group. In contrast, the negative sign implies that the item is easier for the reference group. This study focused on a female subject, with the male subject as the reference group. It did not investigate differences in the subjects with socioeconomic levels and public or private educational institutions that might cause a test score bias.

Conclusion

The Indonesian version of the IST 2000R has good quality items and is suitable for measuring individual intelligence. There was an improvement in the item quality from the previous IST 70 version. Each subtest showed different qualities in discriminatory power, difficulty level, and guessing probability. Some items require further review or revision to improve their quality. The verbal subtest group had the best quality of discriminatory power and the best guessing probability compared to the numerical and spatial figural subtest group. The numerical subtest had the most items with the highest difficulty level. This study recommends examining items affected by gender DIF bias, specifically those in the SC subtest. Furthermore, future studies could examine IST 2000R's unknown construct validity and develop standard norms for Indonesian subjects based on age category.

References

Abedalaziz, N., Leng, C. H., & Alahmadi, A. (2014). Detecting a gender-related differential item functioning using transformed item difficulty. *Malaysian Online Journal of Educational Sciences, 2*, 16-22.

Adedoyin, O. O., & Mokobi, T. (2013). Using IRT psychometric analysis to examine the quality of junior certificate mathematics multiple choice test items. *International Journal of Asian Social Science, 3*(4), 992–1011.

Adinugroho, I. (2016). Pengujian properti psikometrik intelligenz struktur test subtes kemampuan spasial dua dimensi (form auswahl): Studi pada dua SMA swasta di Jakarta. *Jurnal Ilmiah Psikologi MANASA*, *5*(2).

Ayanwale, M. A. (2019). *Efficacy of item response theory in the validation and score ranking of dichotomous and polytomous response mathematics achievement tests in Osun State, Nigeria* (Doctoral Thesis Unpublished). Institute of Education, University of Ibadan.

Baker, F. B. (2001). *The basics of item response theory second edition*. ERIC Clearinghouse on Assessment and Evaluation.

Beauducel, A., Brocke, B., & Liepmann, D. (2001). Perspectives on fluid and crystallized intelligence: Facets for verbal, numerical, and figural intelligence. *Personality and Individual Differences*, *30*(6), 977–994. https://doi.org/10.1016/S0191-8869(00)00087-8

Beauducel, A., Liepmann, D., Horn, S., & Brocke, B. (2010). *Intelligence structure test*. Hogrefe.

Cain, M. K., Zhang, Z., & Yuan, K. H. (2017). Univariate and multivariate skewness and kurtosis for measuring nonnormality: Prevalence, influence, and estimation. *Behavior Research Methods, 49*(5), 1716–1735. https://doi.org/10.3758/s13428-016-0814-1

Cohen, & Swerdlik. (2010). *Psychological testing and assessment: An introduction to tests and. measurement, seventh edition*. McGraw-Hill.

Foster, G. C., Min, H., & Zickar, M. J. (2017). Review of item response theory practices in organizational research: Lessons learned and paths forward. *Organizational Research Methods, 20*(3), 465-486. https://doi.org/10.1177/1094428116689708

Gierl, M. J., Bisanz, J., Bisanz, G. L., Boughton, K. A., & Khaliq, S. N. (2001). Illustrating the utility of differential bundle functioning analyses to identify and interpret group differences on achievement tests. *Educational Measurement: Issues and Practice, 20*(2), 26-36. https://doi.org/10.1111/j.1745-3992.2001.tb00060.x

Ibrahim, A. (2018). Differential item functioning: The state of the art. *Jigawa Journal of Multidisciplinary Studies (JJMS), 1*(1), 37-50.

Martín, E. S., Del Pino, G., & De Boeck, P. (2006). IRT models for ability-based guessing. *Applied Psychological Measurement, 30*(3), 183–203. https://doi.org/10.1177/0146621605282773

McGrory, S., Doherty, J. M., Austin, E. J., Starr, J. M., & Shenkin, S. D. (2014). Item response theory analysis of cognitive tests in people with dementia: A systematic review. *BMC Psychiatry*. https://doi.org/10.1186/1471-244X-14-47

Meyer, J. P. (2014). Applied measurement with jmetrik. Routledge. https://doi.org/10.4324/9780203115190

Obinne, A. D. E. (2012). Using IRT in determining test items prone to guessing. *World Journal of Education, 2*(1). https://doi.org/10.5430/wje.v2n1p91

Ogunsakin, I. B., & Shogbesan, Y. O. (2018). Item response theory (IRT): A modern statistical theory for solving measurement problems in the 21st century. *International Journal of Scientific Research in Education (IJSRE), 11*(3B), 627-635.

Quaigrain, K., & Arhin, A. K. (2017). Using reliability and item analysis to evaluate a teacher-developed test in educational measurement and evaluation. *Cogent Education, 4*(1), 1–11. https://doi.org/10.1080/2331186X.2017.1301013

Rahmawati, E. (2014). *Evaluasi karakteristik psikometri intelligenz struktur test (IST)* [Proceeding]. Seminar Nasional Psikometri, UMS, Indonesia.

Reilly, D., & Neumann, D. L. (2013). Gender-role differences in spatial ability: A meta-analytic review. *Sex Roles, 68*(9–10), 521–535. https://doi.org/10.1007/s11199-013-0269-0

Reilly, D., Neumann, D. L., & Andrews, G. (2022). Gender differences in self-estimated intelligence: Exploring the male hubris, female humility problem. *Frontiers in Psychology, 13,* 812483. https://doi.org/10.3389/fpsyg.2022.812483

Ring, P., Neyse, L., David-Barett, T., & Schmidt, U. (2016). Gender differences in performance predictions: Evidence from the cognitive reflection test. *Frontiers in Psychology, 7,* 1680. https://doi.org/10.3389/fpsyg.2016.01680

Sadhu, S., & Laksono, E. W. (2018). Development and validation of an integrated assessment for measuring critical thinking and chemical literacy in chemical equilibrium. *International Journal of Instruction, 11*(3), 557–572. https://doi.org/10.12973/iji.2018.11338a

Schulze, D., Beauducel, A., & Brocke, B. (2005). Semantically meaningful and abstract figural reasoning in the context of fluid and crystallized intelligence. *Intelligence, 33*(2), 143–159.

https://doi.org/10.1016/j.intell.2004.07.011

Sirodj, D. A. N. (2018). Analisis kualitas item intelligence structure test (IST) melalui metode item response theory (IRT). *SCHEMA (Journal of Psychological Research)*, *4*(2), 98–108. https://doi.org/10.29313/schema.v4i2.4420

Tarigan, M., & Fadillah. (2021a). Properti pikometrik intelligenz struktur test subtes kemampuan numerik (rechenaufgaben dan zahlen reihen). *Jurnal Psikologi Ilmiah*, *13*(2), 155–170. https://doi.org/10.15294/intuisi.v13i2.31839

Tarigan, M., & Fadillah, F. (2021b). Properti psikometri struktur inteligensi IST subtes verbal (Satzergaenzung, Wortauswahl, dan Analogien) berbahasa Indonesia. *Jurnal Muara Ilmu Sosial, Humaniora, dan Seni*, *5*(1), 63-72. https://doi.org/10.24912/jmishumsen.v5i1.9623.2021

Yacob, A., Ali, N., Yusoff, M. H., MohdSaman, M. Y., & Hamzah, W. M. (2014). Personalized learning: An analysis using item response theory. *World Academy of Science, Engineering and Technology, International Journal of Social, Behavioral, Educational, Economic, Business and Industrial Engineering, 8*, 1107-1113.