JP3I (Jurnal Pengukuran Psikologi dan Pendidikan Indonesia), 11(2), 2022, 123-136
DOI: http://dx.doi.org/10.15408/jp3i.v10i1. 25947
http://journal.uinjkt.ac.id/index.php/jp3i

Psychometric Characteristics of the Culture Fair Intelligence Test Scale 2

Novina Sabila Zahra, Widyastuti, Ahmad Ridfah

Faculty of Psychology, Universitas Negeri Makassar, Indonesia

novinasabila@gmail.com

Abstract

This study aims to investigate the psychometric characteristics of the CFIT-scale-2, including investigation of difficulty level, discrimination, distractor effectiveness, DIF (Differential item functioning), IRT (Item Response Theory) 2PL model, validity, and reliability. Respondents were 507 students with 245 (48%) females and 262 (52%) males, with ages ranged from 8-13 years (M= 10.88; SD = 1.35). The classical test theory (CTT) analysis showed that the item difficulty level on the CFIT-scale-2 had varying item difficulty levels, from easy to difficult. However, the test arrangement was not structured according to the suggested difficulty level from easy to medium to difficult. The discrimination of items was poor because 28 items were not included in the very good category (p > .40). In addition, 47 (25%) of the 184 distractors are ineffective. CFIT-scale-2 did not contain DIF ($Adj.p$ > .05), and IRT analysis showed that the CFIT-scale-2 was not structured according to the difficulty pattern from easy, medium, and difficult. The CFIT-scale-2 based on IRT analysis contained 44 items or 96% with good discrimination. The results of a construct validity test using the CFA technique showed a good fit model (p < .001; RMSEA < .08; SRMR < .08), which was acceptable and supported the fit between the theoretical and empirical model. The reliability coefficient value of Cronbach's alpha was 0.88 (α > .70), indicating that the CFIT-scale-2 had good reliability, but the construct reliability was below an acceptable value (CR < .69). According to this study, the psychometric characteristics of the CFIT-scale-2 should be revised and reevaluated.

Keywords: CFIT, CTT, DIF, IRT, psychometric characteristics

Abstrak

Penelitian ini bertujuan untuk mengetahui karakteristik psikometrik CFIT-skala-2, meliputi investigasi tingkat kesukaran, diskriminasi, efektivitas distraktor, model DIF (Differential Item Function), IRT (Item Response Theory) 2PL, validitas, dan reliabilitas. Responden sebanyak 507 siswa dengan 245 (48%) perempuan dan 262 (52%) laki-laki, dengan rentang usia 8-13 tahun (M= 10.88; SD = 1.35). Analisis teori tes klasik menunjukkan bahwa tingkat kesukaran soal pada CFIT-skala-2 memiliki tingkat kesukaran soal yang bervariasi, dari yang mudah sampai yang sulit. Namun, susunan soal tidak terstruktur sesuai dengan tingkat kesulitan yang disarankan dari mudah ke sedang hingga sulit. Diskriminasi item kurang baik karena 28 item tidak termasuk dalam kategori sangat baik (p > .40). Selain itu, 47 (25%) dari 184 distraktor tidak efektif. CFIT-skala-2 tidak mengandung DIF (Adj.p > .05), dan analisis IRT menunjukkan bahwa CFIT-skala-2 tidak tersusun menurut pola kesukaran dari mudah, sedang, dan sukar. Skala CFIT-2 berdasarkan analisis IRT terdapat 44 item atau 96% dengan daya pembeda yang baik. Hasil uji validitas konstruk dengan teknik CFA menunjukkan model yang baik (p < .001; RMSEA < .08; SRMR < .08) dapat diterima dan mendukung kesesuaian antara model teoritis dan empiris. Nilai koefisien reliabilitas Cronbach's alpha adalah 0,88 (α > 0.70), menunjukkan bahwa CFIT-skala-2 memiliki reliabilitas yang baik, tetapi reliabilitas konstruk berada di bawah nilai yang dapat diterima (CR < 0.69). Menurut penelitian ini, karakteristik psikometrik CFIT-skala-2 harus direvisi dan dievaluasi ulang.

***Kata kunci**: CFIT, CTT, DIF, IRT, karakteristik psikometri*

Introduction

Tellegen & Laros (1993) suggest that intelligence testing is widely used to diagnose children, refer children to special education programs, and treat specific disabilities. The intelligence test is a neutral test tool widely used by various groups. However, there may be a cultural bias in intelligence tests. Nenty & Dinero (1981) suggest that cultural experience affects individual responses to test items, leading to cultural bias. Therefore, Raymond B. Cattell developed a Culture Fair Intelligence Test (CFIT) based on the ability of fluid intelligence to measure an individual's analytical and abstract thinking capacity without being based on a particular culture (Gregory, 2013). CFIT aims to measure cognitive abilities free from cultural and environmental influences (Shetty, Pashine, Jose, & Mantha, 2018).

CFIT is a culture-free test tool that has been investigated in many studies. For example, a research conducted by Smith et al. (1977) and Hays & Smith (1980), who examined the CFIT-scale-2 and WISC-r tests in minority adolescents, found that the CFIT-scale-2 reduced cultural bias compared to the WISC-r. Nenty & Dinero (1981) analyzed the CFIT-scale-2 items with the Rasch Model on Nigerian and American students, and most of the CFIT items proved to be free of bias. Heine et al. (2018) found that CFIT can also be used to test individuals with learning disabilities.

Gregory (2013) suggests that the CFIT consists of three versions: The CFIT-scale-1 is used by adults with mental limitations and children between four to eight years old, while the CFIT-scale-2 is used by adults with average intelligence and children aged eight to 13. Scale 3 is used by adults with high abilities and high school and college students. The CFIT-scale-2 and -3 consist of four subtests: series, classification, matrices, and typology (Saptoto, 2018). The CFIT-scale-2 and -3 are more widely used because it is a classical test, so it does not take long to administer.

CFIT-scale-2 and -3 indicate acceptable test-retest reliability, alternative forms, and internal consistency with a score of .70, and test reliability has a score of .80 (Gregory, 2013). Ruiz (2009) suggests that the CFIT-scales-2 and -3 correlated with the general intelligence factor or *g* factor (r= .80) and showed a consistently strong relationship with other intelligence tests such as the WISC, WAIS, Raven's Progressive Matrices, Stanford-Binet, Otis, and General Aptitude Test Battery (r= .70 - .80). Goldstein & Hersen (2000) suggest that the CFIT-scale-2 had a good internal consistency (r= .82) and test-retest (r= .84) coefficient and correlated with other intelligence tests of .50 - .70. Therefore, the CFIT-scale-2 test can adequately measure intelligence and provide consistent results.

However, CFIT was revised in 1961 (Gregory, 2013; Sukadji, 2005). The CFIT-scale-2 is an intelligence test that is still used today. The data obtained by the researcher suggest that the CFIT-scale-2 is still used by psychological services and bureaus for psychological assessment, such as the Psikomorfosa Psychology Bureau, Dwipayana Psychology Bureau, LPPT Widya Prasthya, and the Psychology Service Center of the Faculty of Psychology UNM. The CFIT-scale-2 is also widely used in children's cognitive abilities research. Research conducted by Gottschling et al. (2019), Habersaat et al. (2018), Heine et al. (2018), Putra & Soetikno (2018), Diahsari (2017), Fitriyani & Nulanda (2017), and Pranungsari (2010) were studies that used the CFIT-scale-2 to measure children's cognitive abilities.

The CFIT-scale-2 is still popularly used to measure individual intelligence but has not been revised since 1961. One of the issues that must be addressed in psychological measurement is the measurement bias that can affect the results. Valid and reliable test tools may not necessarily be used in various conditions and research subjects. Therefore, the validity and reliability of test equipment must continue to be studied (Tavakol & Dennick, 2011;Yusup et al., 2018). Thus, the psychometric characteristics of the CFIT-scale-2 could be doubtful, contain bias, and impact the test results. Meanwhile, the CFIT-scale-2 in Indonesia is still underexplored, and the number of studies is relatively small.

Researchers only found a few studies that examined the psychometric properties of the CFIT-scale-3 in Indonesia, namely Fitriani (2018), who examined the psychometric characteristics of the CFIT-scale-3, Nurhadini (2017), who reviewed the validity of the CFIT-scale-3, and Mustari (2015) who studied the CFIT-scale-3 norm. There have been no recent findings for the CFIT-scale-2, and it requires further research.

http://journal.uinjkt.ac.id/index.php/jp3i

Therefore, it is necessary to investigate the CFIT-scale-2's psychometric characteristics and examine its measurement construct. Thus, the analysis results in this study can be considered in using the CFIT-scale-2 to provide robust psychological test services to the public. In addition, this study can be used as a reference in revising the CFIT-scale-2 items.

Methods

Measures

This study employed a quantitative method with a psychometrics approach that focused on evaluating the psychometric properties of the CFIT scale 2. CFIT-scale-2 is used by adults with average intelligence and children aged eight to 13. The CFIT-scale-2 consist of four subtests: series, classification, matrices, and typology.

Participant and Procedures

Data were collected from randomly recruited respondents from various schools and tutoring groups. Respondents were under 18 years old, and parental or guardian permission was required before taking the test. The researcher provided information about the study to the teachers and guardians before administering the test.

Data were collected in class or group, where respondents were given a test simultaneously. The test administration included test booklets, answer sheets, and pencils. The test instruction was delivered by explaining the rules and instructions for working in a structured manner with a time limit for each subtest. In addition, this study also documented respondents' gender (1=male or 0=female), age, class, and school name using a self-report survey. This research obtained 521 responses, but only 507 were included in the data analysis. The respondents were 507 students, 245 (48%) female and 262 (52%) male with ages ranged from eight to 13 years (M= 10.88; SD = 1.35) and they were between grade three Elementary School to grade seven Junior High School (M = 5.32; SD = 1.37).

Data Analysis

Classical Test Theory (CTT)

CTT is readily satisfied by conventional psychometric procedures and its primary focus is on both test-level information and item statistics (Coaley, 2010). This study used the CTT consisting of item difficulty level, item discrimination, and distractor effectiveness

Item Response Theory (IRT)

IRT uses a mathematical model to study the behavior of things based on the item characteristic curve. This is a "consistent model" as it could be more challenging to verify its premises when examining test data (Coaley, 2010). This study used the IRT 2PL model consisting of item difficulty and discrimination analysis.

Differential Item Functioning (DIF)

The DIF measures a function of an item in a test influenced by the respondent's gender even though the respondent has the same ability. This analysis was measured using Raju's area analysis (Raju, 1988). The Raju's area measure was estimated using the Marginal Maximum Likelihood Estimation (MMLE), and the area was obtained using the unsigned area between the ICCs.

Validity

Construct validity is the extent to which a group of measured variables genuinely represents the theoretical latent construct that those variables are supposed to test (Hair, J. F., Black, Babin, Anderson, & Tatham, 2014). Confirmatory factor analysis (CFA) was used to test the fit between a theoretical construct and empirical data among a 1-factor model (series, classification, matrices, and typology) with first-order and second-order models.

Reliability

Reliability is how the consistency of measurement scores can be trusted. This study used Cronbach's alpha coefficient and construct reliability.

Results and Discussion

CTT

The item difficulty parameter is the ratio between the number of correct answers and the total number of individuals who answered the item. So it can also be said that the item difficulty level is the same as the proportion of individuals with the correct answer. Table 1 describes the criteria for the difficulty level of an item with three categories.

Table 1. Item Difficulty Level Category.

Item Difficulty	Category
.00 - .32	Difficult
.33 - .66	Medium
.67 – 1.00	Easy

Sources: Purwanto (2013)

The item discrimination parameter indicates the ability of an item to discriminate between individuals. These parameters can be estimated by computing the total item correlation coefficient, item discrimination power, and mean score comparison. The evaluation criteria for item discrimination are divided into several categories. Please see Table 2.

Table 2. Item Discrimination Category.

Item Discrimination	Category
.40 or more	Very good
.30 - .39	Pretty good
.20- .29	Not satisfied
Less than .20	Items must be discarded

Sources: Azwar (2016)

Table 3. Item Difficulty and Discrimination with CTT in Subtest 1.

Item	Diff		Disc	
	p	*Cat*	*DI*	*Cat*
CFIT_S1_1	.73	Easy	.54	Very good
CFIT_S1_2	.68	Easy	.55	Very good
CFIT_S1_3	.65	Medium	.49	Very good
CFIT_S1_4	.64	Medium	.73	Very good
CFIT_S1_5	.54	Medium	.62	Very good
CFIT_S1_6	.48	Medium	.51	Very good
CFIT_S1_7	.33	Medium	.53	Very good
CFIT_S1_8	.31	Medium	.53	Very good
CFIT_S1_9	.09	Difficult	.23	Not satisfied
CFIT_S1_10	.06	Difficult	.13	Item must be discarded
CFIT_S1_11	.05	Difficult	.16	Item must be discarded
CFIT_S1_12	.03	Difficult	.11	Item must be discarded

N= 507, Cat= Category, Diff = Difficulty, Disc= Discriminant

Sources: Research data (2022)

http://journal.uinjkt.ac.id/index.php/jp3i

Subtest 1 (Series)

Subtest 1 is a subtest that requires test takers to sort images. Subtest 1 consists of 12 items with an administration time of 3 minutes. The results of the psychometric analysis on subtest one can be seen in the Table 3.

CFIT-scale-2 subtest 1 consists of 12 items with five answer choices. Each item has one correct answer and four distractors. Based on the results of analysis, it can be seen that 33 (69%) distractors were effective, and 15 (31%) were ineffective.

Subtest 2 (Classification)

Subtest 2 is a subtest that requires test takers to discriminate a different image from the others. Subtest 2 consists of 14 items with an administration time of 4 minutes. The results of the psychometric analysis on subtest two can be seen in the following Table 4:

Table 4. Item Difficulty and Discrimination with CTT in Subtest 2.

Item	Diff		Disc	
	p	*Cat*	*DI*	*Cat*
CFIT_S2_1	.88	Easy	.54	Very good
CFIT_S2_2	.86	Easy	.56	Very good
CFIT_S2_3	.72	Easy	.43	Very good
CFIT_S2_4	.61	Medium	.34	Pretty good
CFIT_S2_5	.46	Medium	.26	Not satisfied
CFIT_S2_6	.49	Medium	.38	Pretty good
CFIT_S2_7	.30	Difficult	.25	Not satisfied
CFIT_S2_8	.40	Medium	.38	Pretty good
CFIT_S2_9	.38	Medium	.34	Pretty good
CFIT_S2_10	.28	Difficult	.17	Items must be discarded
CFIT_S2_11	.15	Difficult	.23	Not satisfied
CFIT_S2_12	.06	Difficult	.02	Items must be discarded
CFIT_S2_13	.13	Difficult	.16	Items must be discarded
CFIT_S2_14	.02	Difficult	.03	Items must be discarded

N= 507, Cat= Category, Diff = Difficulty, Disc= Discriminant

Sources: Research data (2022)

The CFIT-scale-2 subtest 2 consists of 14 items with five answer choices. Each item has one correct answer and four distractors. Based on the results of analysis, it can be seen that 44 (79%) distractors were effective, and 12 (21%) were ineffective.

Subtest 3 (Matrices)

Subtest 3 is a subtest that requires test takers to complete a picture pattern. Subtest 3 consists of 12 items with an administration time of 3 minutes. The following Table 5 provides the results of the psychometric analysis of subtest 3.

CFIT-scale-2 subtest 3 consists of 12 items with five answer choices. Each item has one correct answer and four distractors. Based on the results of analysis, it can be seen that 34 (71%) distractors were effective, and 14 (29%) were ineffective.

Table 5. Item Difficulty and Discrimination with CTT in Subtest 3.

Item	Diff		Disc	
	p	*Cat*	*DI*	*Cat*
CFIT_S3_1	.79	Easy	.48	Very good
CFIT_S3_2	.80	Easy	.46	Very good
CFIT_S3_3	.75	Easy	.39	Pretty good
CFIT_S3_4	.05	Difficult	-.24	Items must be discarded
CFIT_S3_5	.51	Medium	.47	Very good
CFIT_S3_6	.41	Medium	.52	Very good
CFIT_S3_7	.36	Medium	.55	Very good
CFIT_S3_8	.24	Difficult	.40	Very good
CFIT_S3_9	.19	Difficult	.43	Very good
CFIT_S3_10	.27	Difficult	.40	Very good
CFIT_S3_11	.24	Difficult	.31	Pretty good
CFIT_S3_12	.16	Difficult	.30	Pretty good

N= 507, Cat= Category, Diff = Difficulty, Disc= Discriminant

Sources: Research data (2022)

Subtest 4 (Tipology)

Subtest 4 is a subtest that requires test takers to choose an image that meets certain conditions among the options. Subtest 4 consists of 8 items with an administration time of 2.5 minutes. The following Table 6 provides the results of the psychometric analysis of subtest 4:

Table 6. Item Difficulty and Discrimination with CTT in Subtest 4.

Item	Diff		Disc	
	p	*Cat*	*DI*	*Cat*
CFIT_S4_1	.50	Medium	.40	Very good
CFIT_S4_2	.54	Medium	.38	Pretty good
CFIT_S4_3	.43	Medium	.37	Pretty good
CFIT_S4_4	.44	Medium	.22	Not satisfied
CFIT_S4_5	.28	Difficult	.38	Not satisfied
CFIT_S4_6	.22	Difficult	.08	Items must be discarded
CFIT_S4_7	.21	Difficult	.27	Not satisfied
CFIT_S4_8	.15	Difficult	.24	Not satisfied

N= 507, Cat= Category, Diff = Difficulty, Disc= Discriminant

Sources: Research data (2022)

The CFIT-scale-2 subtest 4 consists of 8 items with five answer choices. Each item has one correct answer and four distractors. Based on the results of analysis, it can be seen that 26 (81%) distractors were effective, and 6 (19%) were ineffective.

IRT

IRT 2PL model consisted of item difficulty and discrimination analysis.The criteria for a good item difficulty level is that the value of parameter *b* is in the range of -2 to +2. While, the criteria for a good item discrimination is that the value of parameter *a* is in the range of 0 to +2 (Setiawati, Izzaty, & Hidayat, 2018).

Subtest 1 (Series)

The results of the psychometric analysis for subtest 1 can be seen in the following Table 7:

Table 7. Item Difficulty and Discrimination with IRT in Subtest 1.

Item	IRT			
	a (SE)	***Cat***	***b (SE)***	***Cat***
CFIT_S1_1	1.57 (.16)	Good	-.90 (.09)	Medium
CFIT_S1_2	1.67 (.16)	Good	-.66 (.08)	Medium
CFIT_S1_3	1.39 (.14)	Good	-.58 (.09)	Medium
CFIT_S1_4	2.71 (.16)	Not good	-.40 (.05)	Medium
CFIT_S1_5	1.99 (.18)	Good	-.11 (.06)	Medium
CFIT_S1_6	1.53 (.15)	Good	.11 (.07)	Medium
CFIT_S1_7	1.85 (.17)	Good	.62 (.07)	Medium
CFIT_S1_8	1.89 (.18)	Good	.71 (.07)	Medium
CFIT_S1_9	.99 (.18)	Good	2.66 (.39)	Difficult
CFIT_S1_10	.29 (.16)	Good	9.35 (5.05)	Difficult
CFIT_S1_11	.62 (.20)	Good	4.98 (1.48)	Difficult
CFIT_S1_12	.49 (.24)	Good	7.28 (3.37)	Difficult

N= 507, IRT= Item Response Theory, a= Discriminant Parameter, b= Difficulty Parameter, SE= Standard Error.

Sources: Research data (2022)

Subtest 2 (Classification)

The results of the psychometric analysis for subtest 2 can be seen in the following Table 8.

Table 8. Item Difficulty and Discrimination with IRT in Subtest 2.

Item	IRT			
	a (SE)	***Cat***	***b (SE)***	***Cat***
CFIT_S2_1	1.97 (.22)	Good	-1.62 (.12)	Medium
CFIT_S2_2	2.22 (.23)	Not good	-1.42 (.09)	Medium
CFIT_S2_3	1.38 (.14)	Good	-.90 (.10)	Medium
CFIT_S2_4	.87 (.11)	Good	-.57 (.13)	Medium
CFIT_S2_5	.71 (.10)	Good	.25 (.13)	Medium
CFIT_S2_6	.87 (.11)	Good	.07 (.11)	Medium
CFIT_S2_7	.62 (.11)	Good	1.46 (.27)	Medium
CFIT_S2_8	.92 (.11)	Good	.55 (.12)	Medium
CFIT_S2_9	.76 (.11)	Good	.72 (.15)	Medium
CFIT_S2_10	.47 (.10)	Good	2.11 (.47)	Difficult
CFIT_S2_11	.70 (.14)	Good	2.68 (.47)	Difficult
CFIT_S2_12*	.33 (.17)	Good	8.56 (4.32)	Difficult
CFIT_S2_13*	.40 (.13)	Good	4.83 (1.49)	Difficult
CFIT_S2_14*	.22 (.20)	Good	17.28 (15.12)	Difficult

N= 507, IRT= Item Response Theory, a= Discriminant Parameter, b= Difficulty Parameter, SE= Standard Error.

Sources: research data (2022)

Subtest 3 (Matrices)

The following Table 9 provides the results of the psychometric analysis for subtest 3:

Table 9. Item Difficulty and Discrimination with IRT in Subtest 3.

Item	IRT			
	a (SE)	***Cat***	***b (SE)***	***Cat***
CFIT_S3_1	1.84 (.19)	Good	-1.12 (.09)	Medium
CFIT_S3_2	1.91 (.19)	Good	-1.14 (.09)	Medium
CFIT_S3_3	1.30 (.14)	Good	-1.13 (.12)	Medium
CFIT_S3_4	.04 (.03)	Good	76.54 (68.27)	Difficult
CFIT_S3_5	1.04 (.12)	Good	-.06 (.10)	Medium
CFIT_S3_6	1.62 (.15)	Good	.35 (.07)	Medium
CFIT_S3_7	1.26 (.13)	Good	.61 (.09)	Medium
CFIT_S3_8	1.33 (.15)	Good	1.14 (.11)	Medium
CFIT_S3_9	1.27 (.16)	Good	1.46 (.15)	Medium
CFIT_S3_10	1.15 (.14)	Good	1.08 (.13)	Medium
CFIT_S3_11	.79 (.12)	Good	1.69 (.25)	Medium
CFIT_S3_12	.94 (.15)	Good	2.08 (.28)	Difficult

N= 507, IRT= Item Response Theory, a= Discriminant Parameter, b= Difficulty Parameter, SE= Standard Error.

Sources: Research data (2022)

Subtest 4 (Tipology)

The following Table 10 provides the results of psychometric analysis for subtest 4:

Table 10. Item Difficulty and Discrimination with IRT in Subtest 4

Item	IRT			
	a (SE)	***Cat***	***b (SE)***	***Cat***
CFIT_S4_1	.45 (.09)	Good	.04 (.20)	Medium
CFIT_S4_2	.50 (.09)	Good	-.33 (.19)	Medium
CFIT_S4_3	.56 (.10)	Good	.53 (.18)	Medium
CFIT_S4_4	.66 (.10)	Good	.43 (.15)	Medium
CFIT_S4_5	.73 (.11)	Good	1.43 (.23)	Medium
CFIT_S4_6	.25 (.10)	Good	5.29 (2.04)	Difficult
CFIT_S4_7	.58 (.12)	Good	2.47 (.47)	Medium
CFIT_S4_8	.68 (.14)	Good	2.74 (.49)	Medium

N= 507, IRT= Item Response Theory, a= Discriminant Parameter, b= Difficulty Parameter, SE= Standard Error.

Sources: Research data (2022)

DIF

The results of the analysis shows the value of Adj. P was in the range of .84 – .96 (*Adj.p* > .05). The results showed that the CFIT-scale-2 items functioned well and were not influenced by gender. This study confirmed that the CFIT-scale-2 items only provided scores representing the test taker's ability and were not influenced by subgroup differences.

Construct Validity

Validity measures the extent to which the score of a measurement follows the theoretical construction that underlies the preparation of the CFIT scale 2. The validity test in this study used a confirmatory factor analysis (CFA). The following are the results of the CFA analysis with a 1-factor model (series, classification, matrices, and typology) with first-order and second-order:

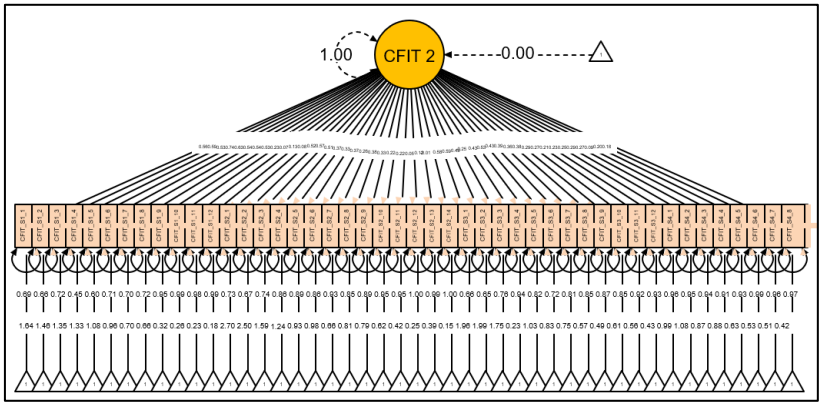
http://journal.uinjkt.ac.id/index.php/jp3i

Table 11. CFA Analysis

Model	χ^2	df	χ^2/df	CFI	TLI	GFI	RMSEA	SRMR
Model 1 (First Order)	5972.67	1036	5.77	.60	.59	.86	.06	.06
Model 2 (Second Order)	5972.67	1035	5.77	.72	.71	.88	.05	.06

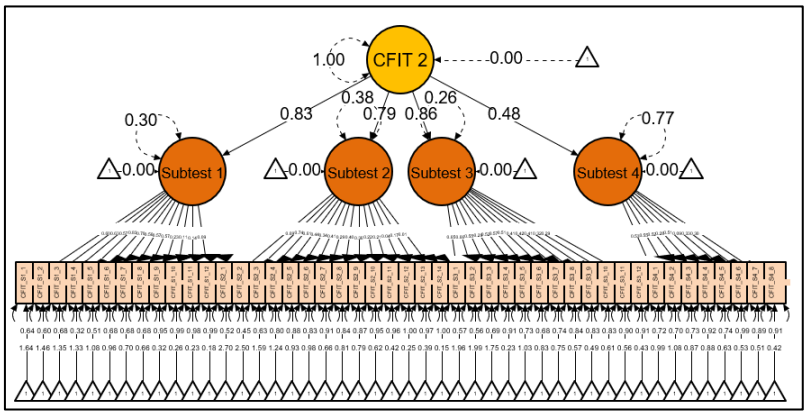
Sources: Research data (2022)

The result of the CFA showed that the first-order and second-order have a suitable model (p < .001; RMSEA < .08; SRMR < .08), which was acceptable and supported the fit between the theoretical and empirical models. Please see Table 1.



Sources: Research data (2022)

Figure 1. CFIT Scale 2 with First-order Model



Sources: Research data (2022)

Figure 2. CFIT Scale 2 with Second-order Model

Reliability

The reliability coefficient value of Cronbach's alpha was 0.88 (α>.70), which indicated that the CFIT scale 2 had good reliability. The CR value was .57, indicating that the CFIT-scale-2 was not reliable because the value was less than .69. The different reliability coefficients between Cronbach alpha and CR did not result in significantly different reliability estimates (Hair, J. F. et al., 2014). There are differences in assumptions between Cronbach's alpha and CR, so each reliability is not on the same baseline to be compared. However, exploration must continue to gather information about the characteristics of psychometrics.

Discussion

This study aims to determine the psychometric characteristics of the CFIT-scale-2 test. Psychometric characteristics are an analysis to determine item analysis with classical test theory (CTT), item response theory (IRT), differential item functioning (DIF), validity, and reliability on the CFIT-scale-2 test. Five hundred seven respondents participated in this study. In line with Coaley (2010), this study argued that the number of participants to evaluate items must be large enough, five times more than the number of items. The CFIT-scale-2 has 46 items, so this study's sample is sufficient for the item psychometric analysis. The approach to item analysis is divided into two parts, namely the CTT and IRT. The CTT consisted of item difficulty level, item discrimination, and distractor effectiveness. Meanwhile, the IRT used the 2PL item difficulty and discrimination analysis model.

The CTT analysis showed that the CFIT-scale-2 was not adequately structured according to the suggested difficulty level (i.e., easy, medium, and difficult). The item difficulty level could be seen from the participant's number of items answered correctly (Price, 2017). The IRT analysis showed that the CFIT-scale-2 items fell between medium and difficult, and there were no items in the easy category. The CFIT-scale-2 items were not structured according to the suggested difficulty level (i.e., easy, medium, and difficult).

The item discrimination of the CFIT-scale-2 did not function properly. Item discrimination measures the size of the test item that distinguishes the examinee with the highest and lowest scores on the test (Price, 2017). From the analysis, 61% of the items included did not satisfy the standard and should be discarded from the test. These items could not distinguish between individuals with high and low abilities. On the other hand, the IRT showed that the CFIT-scale-2 items could discriminate 96% of individuals between high and low abilities. The IRT uses the ICC to indicate the difficulty and discrimination for each CFIT-scale-2 item. The ICC showed the relationship between an individual's ability and the probability of answering the item correctly.

This study showed differences between the CTT and IRT in the item discrimination analysis. The CTT analysis showed that only 39% of the items had good discrimination, while IRT showed that 69% had good discrimination. The advantage of the IRT compared to the CTT is that the standard error measurement (SEM) is different in its measurement.

A measure's difficulty level is proven to be related to item discrimination (Sim & Isaiah Rasiah, 2006). This is indicated by the fact that difficult items tend to have fewer items with high discrimination. A non-functioning discrimination item suggests that a measure may contain ambiguous question instructions, poor administration procedures, or even incorrect answer keys. Therefore, one must evaluate the factors contributing to poor discrimination before deciding to drop items with poor discrimination.

The distractors in multiple-choice questions are designed to contain reasonable but incorrect answers based on common errors so that the options can measure the level of personal knowledge (Shin, Guo, & Gierl, 2019). The CFIT-scale-2 consists of five answer choices, with one correct answer, so the other four answer choices serve as distractors. Individuals with low abilities potentially opt for the ideal distractor, and those with high abilities do not select the distractors (Azwar, 2016). The results indicated that 47 (25%) of the 184 distractors were ineffective. Many ineffective distractors could still affect item discrimination in a test (Shin et al., 2019). In other words, if an item's distractors are ineffective, it cannot distinguish between individuals with and without the ability.

The results revealed the distinctions between the CTT and IRT item analysis approaches. CTT has a SEM that applies to all scores in a given population, while IRT has a SEM that varies across all item scores but may be generalized across the population (De Mars, 2010). Therefore, item analysis using IRT tends to be consistent if given to different samples (Suryabrata, 2005). In this study, participants were

http://journal.uinjkt.ac.id/index.php/jp3i

recruited from various groups (e.g., ages 8-13). Therefore, this study can investigate the ability within the same sample group as well as items from other sample groups.

The CFIT-scale-2 is an intelligence test prepared to be independent of socio-cultural and environmental influences. This study aimed to analyze the CFIT-scale-2 items using the DIF test to see how the functioning of the items in the test was determined by gender in order to prove that the CFIT-scale-2 is free from unintended discriminations (e.g., gender, culture). The DIF analysis assesses whether items behave differently in different subgroups (Fayers & Machin, 2016). The results indicated that the CFIT -scale-2 functioned well, or the results were not influenced by gender. This study confirmed that the CFIT-scale-2 items only provided scores representing the test taker's ability and were not influenced by subgroup differences.

The results of the validity analysis using the CFA showed that the CFIT-scale-2 measures the appropriate theoretical construct. The CFIT-scale-2 is based on fluid intelligence's ability to measure individuals' analytical and abstract thinking capacity without significant cultural effects (Gregory, 2013). Fluid intelligence is seen in a series of ability, classification, analogy, topology, and other well-known intelligence tests. Therefore, this study analyzed the CFIT-scale-2 items construct using the CFA analysis. The results showed that the CFIT-scale-2 was valid in reflecting the theoretical construct that underlined the test.

In addition, this study also found the level of consistency of the CFIT-scale-2 in measuring individual intelligence. The results showed that the CFIT-scale-2 had good internal consistency. This study was supported by some previous studies (Goldstein & Hersen, 2000; Gregory, 2013; Ruiz, 2009), which showed that the CFIT-scale-2 had a high level of reliability. Coaley (2010) also suggested that individual cognitive ability tests, such as IQ tests, must have a high-reliability coefficient.

Overall, the CFIT-scale-2 still has varying difficulty levels, but it did not follow a well-balanced pattern from easy, medium, to difficult. The discriminant of items functions still need some improvements because 61% of items were not included in the good category, and distractors were not generally effective. The CFIT-scale-2 did not contain DIF, so individuals with the same ability also had the same opportunity to answer correctly without being influenced by confounding variables.

The IRT 2PL analysis of the CFIT-scale-2 model showed that the arrangement of the CFIT-scale-2 items was not structured according to the level of difficulty, from easy, medium, and difficult. The discriminant of items that had not functioned clearly because there were still 4% of items not included in the good category. In addition, the CFA and Cronbach's alpha analysis results indicated that all the CFIT-scale-2 items were valid and reliable for measuring intelligence.

This study still has some limitations in its implementation. Firstly, data collection was carried out during the Covid-19 pandemic. The CFIT-scale-2 is a test that still uses a paper-and-pencil test administration. Consequently, data collection during the COVID-19 pandemic restricted the recruitment of some respondents. In addition, the participants in this study were children aged 8 to 13 years old, so the CFIT-scale-2 could not be administered to a significant number of individuals in a single session. Therefore, further research is needed to determine the psychometric characteristics of the CFIT-scale-2 in larger sample size and area.

This study's empirical findings indicated that the psychometric properties of the CFIT-scale-2 should be changed and reexamined. Therefore, psychologists and psychology bureaus should use the test with caution and reevaluate the CFIT-scale-2 psychometric properties before conducting psychological testing. Alternatively, the CFIT-scale-2 should be used with other psychological assessment instruments.

Conclusion

Classical test theory (CTT) of the CFIT-scale-2 showed some varying difficulty levels, with easy to difficult items. Unfortunately, the item arrangement was not structured according to the easy, medium, and difficult categories. The CFIT-scale-2's item discrimination did not function clearly because some items did not meet the criteria for a good discrimination category. The effectiveness of distractors of the CFIT-scale-2 did not meet the desired standard because there were still 47 (25%) distractors that did not show a practical distractor function.

Item response theory (IRT) of the CFIT-scale-2 with the 2PL model showed that the arrangement of the CFIT-scale-2 items was structured according to the difficulty levels, from easy, medium, and difficult. The CFIT-scale-2 items also did not have good item discrimination.

Differential item functioning (DIF) was not included in the CFIT-scale-2 items. The construct validity is classified as fit, indicating the CFIT-scale-2 measured the appropriate theoretical construct. The CFIT-scale-2 showed internal consistency reliability but not construct reliability.

Acknowledgment

We are grateful to Hillman Wirawan for his enthusiastic support, equally enthusiastic critique of this manuscript, and excellent statistical advice. We are also grateful to LPPT Widya Prasthya, who has provided test instruments for the research data collecting.

References

Azwar, S. (2016). *Konstruksi tes kemampuan kognitif*. Yogyakarta: Pustaka Pelajar.

Coaley, K. (2010). *An introduction to psychological assessment and psychometrics*.

De Mars, C. (2010). *Item response theory understanding statistics measurement*.

Diahsari, E. Y. (2017). Memotret kemampuan intelektual siswa SD di pedusunan. *Seminar Nasional dan Gelar Produk*, (9), 746–752.

Fayers, P. M., & Machin, D. (2016). Quality of life: the assessment, analysis and reporting of patient-reported outcomes. *John Wiley & Sons Inc*, *1*, 648.

Fitriani, S. D. (2018). *Karakteristik psikometri pada culture fair intelligence test skala 3*. Universitas Negeri Makassar.

Fitriyani, E., & Nulanda, P. Z. (2017). Efektivitas media flash cards dalam meningkatkan kosakata bahasa Inggris. *Psympathic : Jurnal Ilmiah Psikologi*, *4*(2), 167–182. https://doi.org/10.15575/psy.v4i2.1744

Goldstein, G., & Hersen, M. (2000). *Handbook of psychological assessment*. Elsevier.

Gottschling, J., Hahn, E., Beam, C. R., Spinath, F. M., Carroll, S., & Turkheimer, E. (2019). Socioeconomic status amplifies genetic effects in middle childhood in a large German twin sample. *Intelligence*, *72*(November 2018), 20–27. https://doi.org/10.1016/j.intell.2018.11.006

Gregory, R. J. (2013). *Tes psikologi sejarah, prinsip, dan aplikasi* (6th ed.). Jakarta: Penerbit Erlangga.

Habersaat, S., Ramain, J., Mantzouranis, G., Palix, J., Boonmann, C., Fegert, J. M., Urben, S. (2018). Substance-use disorders, personality traits, and sex differences in institutionalized adolescents. *American Journal of Drug and Alcohol Abuse*, *44*(6), 686–694. https://doi.org/10.1080/00952990.2018.1491587

Hair, J. F., Black, W. C., Babin, B. J., Anderson, R. E., & Tatham, R. L. (2014). *Multivariate data analysis*.

USA: Pearson.

Hays, J. R. A. Y., & Smith, A. L. (1980). Comparison of wisc-r. *Psychological Reports*, *46*, 1–4.

Heine, J.-H., Gebhard, M., Schwab, S., Neumann, P., Gorges, J., & Wild, E. (2018). Testing psychometric properties of the CFT 1-R for students with special educational needs. *Psychological Test and Assessment Modeling*, *60*(1), 3–28.

Mustari, M. A. (2015). *Penyusunan norma tes inteligensi culture fair intelligence test (CFIT)*. Universitas Negeri Makassar.

Nenty, H. J., & Dinero, T. E. (1981). *A cross-cultural analysis of the fairness of the cattell culture fair intelligence test using the rasch Model*. *5*(3), 355–368.

Nurhadini, D. (2017). *Studi pendahuluan: uji validitas konstruk culture fair intelegency test (CFIT) dengan metode confirmatory factor analysis (CFA) di Pusdikbang SDM Perum Perhutani Madium*. Universitas Negeri Makassar.

Pranungsari, D. (2010). Hubungan antara kecerdasan dengan perfeksionisme pada anak gifted di kelas akselerasi. *Humanitas: Indonesia Psychological Journal*, *7*(1).

Price, L. R. (2017). Psychometric methods: theory into practice. In *Measurement: interdisciplinary research and perspectives*. Retrieved from https://lccn.loc.gov/2016013346

Purwanto. (2013). *Evaluasi hasil belajar*. Yogyakarta: Pustaka Pelajar.

Putra, A. S., & Soetikno, N. (2018). Pengaruh intervensi psikoedukasi untuk meningkatkan achievement goal pada kelompok siswi underachiever. *Jurnal Muara Ilmu Sosial, Humaniora, Dan Seni*, *2*(1), 254. https://doi.org/10.24912/jmishumsen.v2i1.1514

Raju, N. S. (1988). The area between two item characteristic curves. *Psychometrika*, *53*(4), 495–502. https://doi.org/10.1007/BF02294403

Ruiz, P. E. (2009). Measuring fluid intelligence on a ratio scale: Evidence from nonverbal classification problems and information entropy. *Behavior Research Methods*, *41*(2), 439–445. https://doi.org/10.3758/BRM.41.2.439

Saptoto, R. (2018). *Pengaruh adaptasi waktu administrasi yang disebabkan penggunaan lembar jawaban komputer terhadap hasil CFIT 3 A dan 3 B*. *45*, 52–65. https://doi.org/10.22146/jpsi.30853

Setiawati, F. A., Izzaty, R. E., & Hidayat, V. (2018). Analisis respon butir pada tes bakat skolastik. *Jurnal Psikologi*, *17*(1), 1–17.

Shetty, R. M., Pashine, A., Jose, N. A., & Mantha, S. (2018). Role of Intelligence Quotient (IQ) on anxiety and behavior in children with hearing and speech impairment. *Special Care in Dentistry*, *38*(1), 13–18. https://doi.org/10.1111/scd.12264

Shin, J., Guo, Q., & Gierl, M. J. (2019). Multiple-choice item distractor development using topic modeling approaches. *Frontiers in Psychology*, *10*(APR), 825. https://doi.org/10.3389/FPSYG.2019.00825/BIBTEX

Sim, S.-M., & Isaiah Rasiah, R. (2006). *Relationship between item difficulty and discrimination indices in true/false-type multiple choice questions of a para-clinical multidisciplinary paper*. *35*(2). Retrieved from http://www.ams.edu.sg

Smith, A. L., Hays, J. R., & Solway, K. S. (1977). Comparison of the wisc-r and culture fair intelligence test in a juvenile delinquent population. *Journal of Psychology: Interdisciplinary and Applied*, *97*(2), 179–182. https://doi.org/10.1080/00223980.1977.9923959

Sukadji, S. (2005). *Tes kecerdasan fair skala 2 dengan Petunjuk Penyajian*.

Suryabrata, S. (2005). *Pengembangan alat ukur psikologi*. Yogyakarta: ANDI.

Tavakol, M., & Dennick, R. (2011). Post-examination analysis of objective tests. *Medical Teacher*, *33*(6), 447–458. https://doi.org/10.3109/0142159X.2011.564682

Tellegen, P. J., & Laros, J. A. (1993). The construction and validation of a nonverbal test of intelligence: the revision of the snijders-oomen tests desenvolvimento da teoria psicometrica view project. *Article in European Journal of Psychological Assessment*, *2*(January), 147–157. Retrieved from https://www.researchgate.net/publication/233387445

Yusup, F., Studi, P., Biologi, T., Islam, U., & Antasari, N. (2018). Uji validitas dan reliabilitas instrumen penelitian kuantitatif. *Jurnal Tarbiyah: Jurnal Ilmiah Kependidikan*, *7*(1), 17–23.