

Psychometric Properties Analysis of IST (Intelligenz Struktur Test)

Nur Akmal¹, Widyastuti², Haerani Nur³

Psychology, Universitas Negeri Makassar

e-Mail korespondensi: nurakmal@unm.ac.id

Abstract. This study aims to examine the characteristics of the intelligence test on the Intelligenz Structure Test (IST) version 70. The study processes were carried out to examine the psychometric characteristics of each subtest on the IST, including: item difficulty index, item discrimination index, distractor effectiveness, tool reliability measuring, and construct validity. The data sources in this study are 1112 people with undergraduate educational backgrounds from several cities in Eastern Indonesia. The results of the analysis of psychometric characteristics show that in the SE subtest there are 11 items accepted, in the WA subtest there are 13 items accepted, the AN subtest shows that 17 items are acceptable, in the GE subtest shows that all the items contained in it are still eligible to be used. In the RA subtest, 18 items are eligible to be accepted, in the ZR subtest, all items are still suitable for use, in the FA subtest, which shows 18 items are acceptable. in the WU and ME subtests, it can be seen that all items contained in them are acceptable. The reliability of the IST construct shows the omega coefficient (ω) is 0.845 and the construct validity shows that the nine subtests contained in the IST fit are based on three criteria of goodness, namely RMSEA, CFI, and TLI.

Keywords: *IST, Psychometric Properties*

INDONESIAN JOURNAL OF EDUCATIONAL STUDIES (IJES)

E-ISSN: 2621-6736

P-ISSN: 2621-6744

Submitted : 2nd October 2021

Revised : 10th October 2021

Accepted : 1st November 2021



This work is licensed under a [Creative Commons Attribution-NonCommercial 4.0 International License](https://creativecommons.org/licenses/by-nc/4.0/)

INTRODUCTION

A psychological test is an objective measurement method to find out differences between individuals or the reaction of the same individual in different situations. Psychological testing is a quantification process carried out by psychologists in providing assessments to individuals according to the goals and objectives for which the test is given. In administering psychological tests, there are standardized procedures, namely, preparation of test material, presenting instructions, scoring and interpretation.

The psychological test contains items that are scored based on the answers of the test takers, then the score will be compared with the established norms so that it can help facilitate the interpretation of the score. Thus psychological tests can help predict behaviour and provide information about a person's description following the purpose of carrying out the test (Kaplan & Saccuzo, 2005). The administrators of the tests need to develop the knowledge on how to evaluate such as recognizing the basic characteristics of the tools, namely validity, reliability, level of difficulty, and norms. So that the psychological test tool is expected to work well (Anastasi & Urbina, 2006).

Currently, the psychological test tools have been used and made rapid progress and have made a very large contribution in various fields of life. Starting from the educational sectors, industries and organizations, even to the social, the psychological tests have been involved. As the schools has usually done in selecting and determining which students deserve to take part in accelerated or regular programs. In the industrial and organizational fields, psychological tests are used as a tool for selecting or classifying human resources, such as employee recruitment, assignment, transfer, promotion, and even termination of employment. Psychological tests are also commonly used to assess victims of sexual harassment. Thus there is almost no job that does not require the analysis of psychological tests (Anastasi & Urbina, 2006).

Basically, there are two kinds of psychological tests, namely, personality tests and ability tests. The personality test is a test that reveals the image from within a person. The ability test is a test that measures the speed and accuracy of answers. One of them is the intelligence test, which is a test used to measure general abilities such as problem-solving potential and adaptability in the environment (Kaplan & Saccuzo, 2005). IST (Intelligenz Structure Test) is one of the tools used to measure a person's level of intelligence. IST popularized by Rudolf Amthauer in 1953 in Germany. The IST has been revised several times in 1973 which was referred to as IST-70, in 1999 it was referred to as IST 2000 and the most recent one in 2007 was known as IST 2000R. IST was designed to measure verbal intelligence, numerical intelligence and spatial intelligence which are assessed in 9 subtests.

IST in Indonesia had adapted by the Faculty of Psychology, the University of Padjadjaran in 1973. Until now, the use of IST in Indonesia is still the result of the adaptation of the first version. IST is the most frequently used test in Indonesia, including in the Eastern Indonesia region. The use of IST is based on recommendations from the test administration bureau that wants to conduct

employee selection. The problem is that the IST that we use until now is the result of the adaptation of the first version in 1973. A very influential factor in this is the quality of the items that reflect the characteristics of the measurement that are considerably old or old enough to keep up with the times.

Moreover, the norms used are still the same, besides changes in the characteristics or the abilities of the population may occur over time. The results conducted by Flynn (Princen, 2011) found an increase in IQ scores within decades. Thus, the norms that have never been updated can influence the results of the supposed IQ scores. Based on the observations by researchers, there have never been evaluating or revising in the IST in Eastern Indonesia, especially in Makassar City. Seeing that the use of IST in several cities in Eastern Indonesia, especially in Makassar City, has never been reviewed, it is very reasonable if the items in it are not appropriate and in line with the increasingly dynamic times, this is according to the findings of Rahmawati (2014) conducted at the Center for Research and Community Service (P3M) the University of North Sumatra showed that 54.375% of IST items had poor quality. Moreover, the use and administration of tests that are carried out repeatedly can have a learning effect for test-takers. Not to mention other conditions that are difficult to control, such as question leaks. If the IST from 1973 used has never been revised again, the chances of leaking questions are getting bigger.

Researchers found indications of leakage of questions through the internet that provide psychological test training and selling the psychological test modules, one of which is the IST module. If this happens, the IST which is still used today (IST-70) can be doubted to measure the actual ability of individuals. So that it is not following the measurement rules that must be valid and reliable.

The purpose of a measurement is to measure what it is supposed to measure. The success of a measuring instrument can be seen from how much the measuring tool can provide accurate and exact results (Azwar, 2001). That is if intelligence tests are often used to select employees, of course, it must be related to how appropriate a person is to occupy the intended position following predetermined criteria. If the instrument on a test tool is not right, it will produce an inaccurate score result, so it can be said that the test tool is invalid.

The validity of a test instrument can affect its reliability. Reliability is the extent to which a measurement can provide reliable results. If testing the reliability of the test instrument provides a very high difference in score, then the measurement results cannot be trusted. Likewise, with validity, if the measuring instrument cannot provide accurate and correct measurement functions, then the measuring instrument cannot be trusted. Based on the description above, the researcher assumes that the above problems need to be studied further regarding the psychometric characteristics of the IST itself.

To produce good quality instruments, it is necessary to carry out theoretical analysis (item study) as well as empirical analysis, namely item analysis. Mainly in empirical analysis can be done with two approaches, one of the approaches is the classical test theory. Classical test theory has several advantages, namely: (1) using simple concepts to determine the ability of test-takers, (2) using simple concepts to calculate item parameter values, (3) can be used on small samples (Suwarto, 2011).

said that the classical test theory is a theory that is easy to apply and can provide an overview of errors in measurements that affect the observed score (Lababa, 2008).

Allen & Yen (Azwar, 2009) stated that several assumptions of classical test theory, namely: 1). The first assumption of classical test theory is that there is a relationship between the visible score given the symbol X , the pure score represented by T and the measurement error score (error) represented by E . The high score of the raw score is determined by the sum of the true score and measurement error, which can be represented by $X = T + E$. 2). The second assumption states that the true score (T) is the expected value of the raw score (X) so that the true score is the average score of the results carried out through repeated measurements and each measurement is not with the others. Can be represented by $(X) = T$. 3). The third assumption states that the true score (T) and the measurement score on a test have no relation. In other words, high and low true scores (T) have no positive or negative errors. It can be represented by $(\rho_{TE} = 0)$. 4). The fourth assumption states that the first measurement error and the second measurement error have no relation with each other. This means that the sum of error (E) in one test does not depend on errors in other tests. $(\rho_{E1E2} = 0)$. 5). The fifth assumption is that if two tests measure the same attribute, then the error (E) on the first test does not correlate with the true score (T) on the second test. 6). $\rho_{ET2} = 0$. 7). The assumption which is the classical test theory states that if there are two tests to measure the same attribute, and have the same true score ($T = T'$) and the same error score variance, then the test is said to be a parallel test. 7). The seventh assumption states that the test can be said to be equivalent if the two tests have the same visible score (X and X') and comply assumptions 1 to 5, and each subject population $X_1 = X_2 + C$ where C is a constant number, then the test is said to be parallel. The assumptions of the classical theory mentioned above look good enough to be developed as the basis for developing psychological measurements. Item difficulty index, item discrimination index, reliability, and validity are the most important things in classical test theory.

RESEARCH METHOD

This study aims to examine whether the IST is still running following the initial purpose of the measurement, which is based on the analysis of psychometric characteristics. The data collection method used in this research is the documentation method, namely: the process of searching and tracing data or things in the form of writing (paper), places and people that could be used as sources of information. The documents used in this study came from institutions or psychological test service bureaus in Makassar City. The sources of data involved in this research are people who have been examined by the psychological test agency or service bureaus. In this study, 1112 people were analyzed. Analysis of the data used in this study aims to determine the psychometric characteristics of IST. The analysis conducted includes the analysis of item difficulty index, item discriminatory index, loading factor, reliability test, and construct validity. Data analysis in this study used the Iteman (tm) version 3.0 and M.Plus programs.

RESULT AND DISSCUSSION

This study uses secondary data, it is the documentation of test results from one of the tests administering bureaus in Makassar City. Research data is the response of the subject when answering the items in each subtest contained in the IST. IST has nine subtests with a total of 176 items. The data in this study is dichotomous, the subject's response by given a score of one (1) for correct answers and zero (0) for incorrect answers, and the unanswered items are left blank and coded (N). There were 1112 IST data to analyze from 742 males and 370 females. Based on the age, there are 30 people under the age of 20 years, 861 people with an age range of 21-30 years, 134 people aged between 31-40 years, 62 people aged 41-50 years, and there are around 22 people who are over the age 50 years.

Tabel 1. Description

No	Variable	Desc	Freq	%
1	Sex	Male	742	66,7
		Female	370	33,3
2	Age	< 21	30	2,7
		21-30	861	77,4
		31-40	134	12,1
		41-50	62	5,6
		>50	22	2,0
		Not Identified	3	0,3

Following are the results of item analysis using the Iteman (tm) Version 3.00 program, it is known that the item difficulty index (p) in each subtest on the IST (Intelligenz Structure Test), subtest & test reliability and factorial validity.

Subtest of SE

The analysis results of the item difficulty index on the SE subtest showed variations in the level of difficulty (p) consisting of easy, medium and difficult items. There are three easy items, 11 items with medium difficulty, and six items that are classified as difficult. Based on the results of the item discrimination power analysis, it shows that five items have good different power, five items that have quite good different power, nine items that have quite poor different power, and one item that has a poor different power. The results of the analysis showed that 13 items had effective distractors, five items that were less effective and two that had ineffective distractors. The research item selection process was based on the item difficulty index, item discrimination index and distractor effectiveness. The criteria for selecting a good item with a satisfactory differentiating power is $d > 0.3$ and an item with a discriminatory power $d < 0.2$ deserves to be aborted. In this analysis process, seven items need to be revised, one item deserves to be aborted and 12 other items are still considered acceptable. Based on the researcher's observations, two items did not reach 0.3 but because they were considered vital content, both items were acceptable for use. The reliability coefficient on the SE subtest was obtained using the Iteman (tm) Version 3.0 program with the KR-20 formula, so this subtest has an internal consistency index (α) of 0.537.

Table 2. Item analysis of SE

Item	P	Rpbis	Loading factor	Distractor				
				a	b	c	d	e
a1	0.54	0.403	0.319	-0.106	-0.193	-0.085	-0.169	0.403
a2	0.949	0.214	0.202	-0.098	-0.106	0.214	-0.11	-0.103
a3	0.842	0.311	0.315	-0.076	-0.219	-0.117	0.331	-0.098
a4	0.900	0.352	0.374	-0.097	-0.273	-0.121	0.352	-0.080
a5	0.266	0.282	0.157	-0.258	0.0430	-0.13	0.282	0.107
a6	0.656	0.292	0.160	-0.116	0.292	-0.206	-0.076	-0.085
a7	0.336	0.220	0.081	-0.122	0.03	0.220	-0.057	-0.116
a8	0.457	0.369	0.229	0.369	-0.134	-0.179	-0.066	-0.163
a9	0.545	0.441	0.423	-0.153	-0.196	-0.106	-0.112	0.441
a10	0.678	0.423	0.378	-0.202	0.423	-0.264	-0.03	-0.126
a11	0.597	0.388	0.317	-0.107	-0.211	0.388	-0.083	-0.081
a12	0.103	0.165	0.053	0.046	-0.046	0.009	0.165	0.034
a13	0.364	0.449	0.408	-0.251	-0.013	-0.077	0.449	-0.050
a14	0.507	0.381	0.295	-0.205	-0.084	-0.091	-0.059	0.381
a15	0.241	0.216	0.071	-0.04	0.001	0.216	-0.011	0.047
a16	0.145	0.242	0.110	0.242	-0.046	0.005	0.054	-0.028
a17	0.241	0.215	0.046	0.003	0.215	-0.025	-0.070	-0.010
a18	0.314	0.234	0.124	0.026	0.234	0.055	0.008	-0.018
a19	0.309	0.420	0.344	-0.032	-0.047	0.420	-0.133	-0.050
a20	0.174	0.283	0.185	-0.013	0.283	0.011	-0.081	0.036

Subtest of WA

The analysis results of the item difficulty index on the WA subtest showed variations in the level of difficulty (p) consisting of easy items to difficult items. The results of the analysis show that there are eight easy items, 11 items with moderate difficulty level and one item that is classified as difficult. However, the variation in the value (p) is not well organized from the highest to the lowest. Based on the results of the item discrimination index analysis (d) it shows that there are items that have different power which is classified as good, quite good and not good. The results of the analysis show that three items have good different power, 10 items that have good different power and six items that have poor different power and one item that has a poor different power. In the WA subtest, it can be seen that the distractors have functioned well as a whole, thus the distractors in each item can outwit subjects who have low abilities. Item selection in this study was based on item difficulty index, item discrimination index and distractor effectiveness. The criteria for selecting a good item with satisfactory discriminatory power is $d > 0.3$, if $d < 0.2$ it deserves to be aborted and considering the effectiveness of the distractor. The analysis process showed that there was one item that removed, six items needed to be revised and 13 other items were still considered acceptable. The reliability coefficient on the SE subtest showed an internal consistency (α) of 0.547. (α) sebesar 0.547.

Table 3. Item analysis of WA

Item	P	Rpbis	Loading factor	Distractor				
				a	b	c	d	e
a21	0,524	0,400	0.360	-0,102	0,4	-0,307	-0,063	-0,064
a22	0,904	0,311	0.278	-0,133	0,311	-0,19	-0,104	-0,137
a23	0,693	0,272	0.152	-0,027	-0,203	-0,114	0,272	-0,038
a24	0,809	0,371	0.347	-0,165	-0,114	0,371	-0,138	-0,221
a25	0,794	0,31	0.243	-0,075	-0,193	0,31	-0,201	-0,06
a26	0,787	0,369	0.301	-0,14	-0,048	0,369	-0,114	-0,282
a27	0,831	0,311	0.267	-0,137	-0,236	0,311	-0,083	-0,081
a28	0,733	0,233	0.098	-0,001	-0,115	-0,122	0,233	-0,164
a29	0,338	0,336	0.286	-0,234	0,027	-0,073	0,336	-0,122
a30	0,731	0,468	0.480	0,468	-0,132	-0,317	-0,117	-0,203
a31	0,500	0,383	0.280	-0,150	-0,147	-0,134	-0,137	0,383
a32	0,342	0,440	0.437	0,440	-0,046	-0,136	-0,333	0,022
a33	0,382	0,160	-0.019	0,160	0,046	-0,056	-0,084	-0,079
a34	0,648	0,278	0.159	-0,047	0,278	-0,114	0,092	-0,185
a35	0,463	0,227	0.083	-0,032	-0,092	0,227	-0,059	-0,136
a36	0,228	0,39	0.356	0,390	-0,035	-0,015	-0,102	-0,161
a37	0,567	0,247	0.106	0,013	-0,061	-0,151	0,247	-0,098
a38	0,364	0,367	0.283	0,006	-0,071	-0,134	-0,166	0,367
a39	0,326	0,369	0.279	-0,108	0,369	-0,058	-0,097	-0,067
a40	0,713	0,268	0.098	-0,048	-0,034	0,268	-0,102	-0,114

Subtest of AN

The analysis of the item difficulty index on the AN subtest showed that there were two easy items, 10 items with medium difficulty and eight items with highly difficult. The arrangement of p-values from the highest to the lowest looks quite good, so it almost resembles the preparation of an intelligence test that arranged from the easiest to the most difficult of items. The item discrimination index analysis on the AN subtest looks quite satisfactory, it is indicated by the items having good discriminatory power of seven items, items of the quite good discriminating power of 9 items, and items having poor distinguishing power of four items. The distractor analysis in the subtest shows that 16 items have effective distractors and four items that are still less effective so that the item selection process obtained three items that need to be revised and 17 other items that are still eligible to be accepted. In the AN subtest, there is one vital content of item, so that the item can be accepted to use. The reliability coefficient on the AN subtest shows an internal consistency index (α) of 0.701.

Table 4. Item analysis of AN

Item	P	Rpbis	Loading factor	Distractor				
				a	b	c	d	e
a41	0,935	0,298	0.250	-0,217	-0,012	0,298	-0,2	-0,022
a42	0,612	0,306	0.171	-0,175	-0,082	-0,028	-0,127	0,306

a43	0,761	0,437	0.373	-0,248	-0,077	-0,17	0,437	-0,197
a44	0,68	0,469	0.428	-0,067	-0,266	-0,282	0,469	-0,054
a45	0,531	0,519	0.469	-0,225	-0,179	-0,291	0,519	-0,219
a46	0,665	0,403	-0.095	-0,112	0,403	-0,056	-0,316	-0,1
a47	0,409	0,604	0.592	-0,351	-0,256	-0,065	0,604	-0,053
a48	0,522	0,333	0.220	-0,072	0,333	0,058	-0,307	-0,165
a49	0,304	0,299	0.208	0,104	-0,322	-0,044	-0,057	0,299
a50	0,342	0,433	0.379	-0,09	-0,183	-0,146	0,433	-0,129
a51	0,389	0,584	0.585	-0,275	-0,216	0,584	0,041	-0,272
a52	0,107	0,368	0.336	-0,068	-0,448	0,368	0,021	0,221
a53	0,107	0,245	0.195	-0,2	0,092	0,245	0,021	-0,047
a54	0,272	0,397	0.316	-0,024	-0,219	0,397	-0,077	-0,066
a55	0,128	0,252	0.201	0,041	0,061	-0,086	0,252	-0,141
a56	0,112	0,309	0.269	-0,138	-0,1	0,309	0,137	-0,065
a57	0,165	0,383	0.325	0,127	-0,09	0,383	-0,219	-0,107
a58	0,58	0,346	0.044	-0,193	-0,174	-0,09	-0,047	0,346
a59	0,083	0,315	0.304	-0,026	-0,096	0,068	0,019	0,315
a60	0,239	0,361	0.286	-0,09	-0,111	0,102	-0,142	0,361

Subtest of GE

The GE subtest has three forms of scoring, namely 2, 1, and 0. The item analysis process in the GE subtest is that a score of two and one will be considered as a correct answer, and a score of zero is considered a wrong answer so that the score is dichotomous. In the analysis of the item difficulty index on the GE subtest, seven items are classified as easy, seven items that are classified as moderate difficulty and two items that are classified as difficult. The item discrimination index analysis shows that there are 12 items with good discriminatory power, three items that are quite good and one item that is classified as bad. After going through the selection process for item $d > 0.3$, it can be seen that the items contained in it are still eligible to be accepted. The analysis process found one item that is considered vital content so that the item can be accepted. The reliability coefficient on the GE subtest shows an internal consistency index (α) of 0.664.

Table 5. Item analysis of GE

Item	P	Rpbis	Loading factor
a61	0,995	0,199	0.276
a62	0,933	0,414	0.472
a63	0,332	0,426	0.321
a64	0,889	0,307	0.204
a65	0,666	0,477	0.410
a66	0,535	0,460	0.374
a67	0,744	0,408	0.409
a68	0,762	0,508	0.454
a69	0,432	0,403	0.287
a70	0,607	0,477	0.362
a71	0,218	0,411	0.342

a72	0,735	0,452	0.474
a73	0,228	0,365	0.255
a74	0,786	0,469	0.431
a75	0,459	0,352	0.162
a76	0,468	0,454	0.344

Subtest of RA

The analysis of the item difficulty index on the RA subtest showed that there were two items with the easy level of difficulty, two items with a moderate level of difficulty and 16 items in a difficult level. The item analysis shows that the difficult items dominate more than the other items, this is because the variation in the answers of the subjects on average looks empty in the sense that the subjects do not answer some of the lower items. So that it affects the level of difficulty of the item. The item discrimination index analysis shows that there are 14 items with good discriminatory power, three items with fairly good discriminatory power, two items that look less good, and one item with poor discriminatory power. The item selection process based on the discrimination index $d > 0.3$ as a satisfactory item and $d < 0.2$ deserves to be aborted, two items do not reach 0.3 but because it is considered vital content only one item deserves to be accepted and one item deserves to be aborted and 18 items others are considered to still eligible to be accepted. The reliability coefficient on the RA subtest shows an internal consistency index (α) of 0.779.

Table 6. Item analysis of RA

Item	P	Rpbis	Loading factor
a77	0,942	0,24	0.127
a78	0,757	0,399	0.235
a79	0,233	0,404	0.277
a80	0,249	0,545	0.404
a81	0,662	0,547	0.373
a82	0,291	0,585	0.453
a83	0,524	0,484	0.315
a84	0,245	0,603	0.500
a85	0,136	0,534	0.480
a86	0,079	0,442	0.410
a87	0,200	0,564	0.490
a88	0,085	0,521	0.561
a89	0,048	0,488	0.539
a90	0,074	0,527	0.585
a91	0,039	0,454	0.547
a92	0,038	0,418	0.472
a93	0,01	0,229	0.285
a94	0,021	0,328	0.389
a95	0,018	0,354	0.419
a96	0,016	0,161	0.133

Subtest of ZR

The Analysis of the item difficulty index in the ZR subtest showed that there were five items with easy difficulty levels, eight items with moderate difficulty levels and seven items with higher difficulty levels. Variations in difficulty levels move from easy items to difficult items. The item discrimination index analysis shows that there are 17 items with good discriminatory power, three items with fairly good discriminating power. The item selection process based on the discrimination index $d > 0.3$ shows that the overall quality of the items is still worthy of acceptance. The reliability coefficient on the RA subtest shows an internal consistency index (α) of 0.851.

Table 7. Item analysis of ZR

Item	P	Rpbis	Loading factor
a97	0,924	0,381	0.313
a98	0,869	0,429	0.349
a99	0,845	0,457	0.387
a100	0,865	0,452	0.370
a101	0,526	0,431	0.336
a102	0,74	0,503	0.417
a103	0,444	0,564	0.500
a104	0,327	0,592	0.541
a105	0,373	0,492	0.422
a106	0,554	0,567	0.508
a107	0,481	0,656	0.619
a108	0,460	0,585	0.542
a109	0,353	0,613	0.601

Table 7. Item analysis of ZR.

Item	P	Rpbis	Loading factor
a110	0,202	0,600	0.615
a111	0,224	0,603	0.620
a112	0,162	0,574	0.594
a113	0,153	0,479	0.482
a114	0,106	0,378	0.367
a115	0,061	0,393	0.386

Subtest of FA

The analysis of the item difficulty index on the FA subtest showed that there were four easy items, 12 items with moderate difficulty and four items with highly difficult. The arrangement of p-values from the highest to the lowest does not look good enough, so it does not resemble the preparation of intelligence tests which are arranged from the easiest to the most difficult items. The item discrimination index analysis on the FA subtest shows eight items with good discriminatory power, eight items with fairly good discriminatory power, three items with poor distinguishing

power and one item with poor distinguishing power. The distractor analysis on the FA subtest showed 18 items that had an effective distractor and one is a less effective item and one is an ineffective item. The item selection process shows that two items are considered vital content so that 18 items deserve to be accepted and two items that seem to need to be revised. The reliability coefficient on the FA subtest shows an internal consistency index (α) of 0.673.

Table 8. Item Analysis of FA

Item	P	Rpbis	Loading factor	Distractor				
				a	b	c	d	e
a117	0,825	0,193	0.096	0,193	-0,185	-0,042	-9	-0,003
a118	0,633	0,378	0.323	-0,184	-0,211	0,378	-0,082	-0,119
a119	0,49	0,251	0.129	-0,018	0,251	-0,035	-0,239	0,04
a120	0,367	0,354	0.269	0,354	-0,175	-0,073	-0,106	-0,129
a121	0,27	0,386	0.345	118,945	-0,152	0,089	0,386	-0,103
a122	0,633	0,436	0.391	-0,063	0,436	-0,147	-0,189	-0,129
a123	0,464	0,491	0.447	-0,13	-0,162	0,491	-0,169	-0,106
a124	0,451	0,417	0.361	-0,075	-0,062	-0,115	-0,072	0,417
a125	0,714	0,369	0.312	-0,02	-0,033	-0,067	-0,176	0,369
a126	0,449	0,415	0.349	-0,048	-0,087	-0,087	0,415	-0,052
a127	0,254	0,339	0.284	-0,012	0,006	-0,061	-0,126	0,339
a128	0,226	0,298	0.205	0,016	0,298	-0,093	-0,081	-0,016
a129	0,695	0,474	0.439	-0,067	-0,145	-0,073	0,474	-0,172
a130	0,461	0,34	0.259	-0,155	-0,077	0,34	-0,043	-0,07
a131	0,733	0,42	0.365	-0,036	0,42	-0,015	0,096	-0,082
a132	0,800	0,335	0.057	-0,026	-0,122	-0,126	-0,034	0,335
a133	0,575	0,478	0.427	-0,058	0,478	-0,065	-0,121	-0,053
a134	0,522	0,435	0.356	-0,072	-0,091	-0,076	0,435	-0,006
a135	0,212	0,293	-0.029	-0,009	0,293	-0,022	0,011	0,09
a136	0,631	0,314	0.183	0,314	0,293	-0,022	0,011	0,09

Subtest of WU

The analysis of the item difficulty index on the WU subtest showed that there were seven easy items, five items with moderate difficulty and eight items with highly difficult. The arrangement of p-values is not good structured, starting from the easiest item to the most difficult item. So that, this is not following the purpose of preparing tests that move from the highest to the lowest. The item discrimination index analysis on the WU subtest showed that there were 15 items with the good of discriminatory power, five items with fairly good discriminating power. The distractor analysis on the WU subtest shows that the distractors as a whole have functioned effectively. Based on the item selection analysis, it shows that all items contained in it are still feasible to be accepted and used. The reliability coefficient on the WU subtest shows an internal consistency index (α) of 0.790.

Table 9. Item Analysis of WU

Item	P	Rpbis	Loading factor	Distractor				
				a	b	c	d	e
a137	0,791	0,458	0.469	0,458	-0,131	-0,305	0,208	-0,131
a138	0,779	0,432	0.436	-0,234	-0,196	0,432	-0,081	-0,217
a139	0,803	0,466	0.506	-0,234	-0,232	-0,159	0,466	-0,196
a140	0,824	0,503	0.549	-0,142	-0,19	0,208	-0,244	0,503
a141	0,236	0,393	0.252	0,393	-0,028	-0,102	-0,213	-0,075
a142	0,698	0,497	0.486	-0,274	-0,196	0,497	0,249	-0,104
a143	0,826	0,468	0.505	-0,132	-0,21	-0,252	0,468	-0,138
a144	0,3	0,365	0.224	-0,093	-0,114	0,365	-0,01	-0,136
a145	0,76	0,522	0.549	-0,184	-0,203	-0,219	-0,113	0,522
a146	0,615	0,498	0.469	0,498	-0,116	-0,201	-0,150	-0,215
a147	0,75	0,539	0.559	-0,142	0,539	-0,179	0,200	-0,15
a148	0,325	0,466	0.340	-0,045	-0,103	-0,149	0,466	-0,055
a149	0,233	0,438	0.319	0,035	-0,064	-0,06	0,093	0,438
a150	0,654	0,505	0.496	-0,109	0,505	0,066	-0,138	-0,102
a151	0,293	0,478	0.348	0,010	-0,095	0,054	0,478	-0,106
a152	0,246	0,432	0.273	-0,121	0,432	0,118	-0,135	-0,015
a153	0,157	0,393	0.260	0,393	-0,049	0,033	0,024	-0,03
a154	0,172	0,41	0.274	-0,026	-0,031	0,076	0,016	0,41
a155	0,201	0,356	0.217	-0,065	0,356	0,088	-0,091	-0,015
a156	0,174	0,323	0.180	-0,053	-0,057	0,323	0,085	-0,039

Subtest of ME

The analysis of the item difficulty index on the MA subtest showed that there was only one item that was classified as easy, the other 19 items were classified as moderate difficulty. this is certainly not following the test preparation model which arranged from the easiest item to the most difficult item. The results of the discrimination analysis show that the items contained in it look good, as well as the distractors contained in each item as a whole function effectively, so that the item selection based on the discrimination index and distractor effectiveness shows that all items on the ME subtest are eligible to be accepted.

Table 10. Item Analysis of ME

Item	P	Rpbis	Loading factor	Distractor				
				a	b	c	d	e
a157	0,618	0,422	0.362	-0,241	-0,125	-0,119	0,442	-0,082
a158	0,532	0,439	0.367	-0,101	-0,156	-0,126	-0,053	0,439
a159	0,571	0,404	0.328	-0,065	0,404	-0,084	-0,154	-0,011
a160	0,664	0,459	0.397	0,459	-0,055	-0,116	-0,174	-0,096
a161	0,638	0,434	0.373	-0,113	-0,075	0,434	-0,092	-0,099
a162	0,785	0,42	0.362	0,42	-0,117	-0,07	-0,16	-0,067
a163	0,633	0,456	0.396	-0,025	-0,194	-0,115	0,465	-0,113
a164	0,694	0,478	0.427	-0,049	-0,154	-0,127	-0,115	0,478
a165	0,519	0,482	0.428	0,013	-0,107	0,482	-0,14	-0,127
a166	0,541	0,466	0.412	-0,048	0,466	-0,095	-0,073	-0,125
a167	0,441	0,406	0.343	-0,028	0,406	-0,099	-0,058	-0,083
a168	0,371	0,539	0.512	0,539	-0,099	-0,072	-0,045	-0,04
a169	0,348	0,517	0.488	-0,123	-0,078	-0,086	-0,043	0,517
a170	0,498	0,509	0.466	-0,086	-0,108	0,509	-0,093	-0,049
a171	0,594	0,511	-0.067	-0,093	-0,158	-0,071	0,511	-0,04
a172	0,585	0,463	0.419	-0,039	0,463	-0,114	-0,052	-0,004
a173	0,39	0,524	0.490	-0,053	-0,081	-0,055	-0,059	0,524
a174	0,392	0,466	0.432	0,466	-0,105	-0,036	-0,065	-0,005
a175	0,323	0,532	0.520	-0,044	-0,066	0,532	-0,065	-0,057
a176	0,344	0,508	0.472	-0,083	-0,015	-0,065	0,508	-0,075

Item analysis in this study aims to see the level of item difficulty, item discrimination power and distractor effectiveness. The item difficulty index is a comparison between the number of subjects who answered the item correctly and the number of subjects (Suwanto. 2007). Azwar (2001) suggests that the p-value moves from 0 to 1. Items that close to 0 are categorized as difficult items, items in the range of 0.5 are categorized as ideal items and items that close to 1 are categorized as easy items. This opinion is explained by Gregory (2000), the items that are classified as easy are in the range of $p > 0.7$, the items with moderate difficulty are in the range of 0.3 to 0.7 and the difficult items are in the range of $d < 0.3$.

The item discriminatory index shows a high and low power difference in the subject (Gregory. 2000). The discrimination index consists of items with good discriminating power are at index $d > 0.4$. the difference is quite good of the range from the index 0.3-0.39. Dissimilarity is not good in the range from 0.2-0.29 index and poor discriminating power of the range from $d < 0.2$ (Ebel: Azwar. 2001). The item selection process is based on discriminatory power $d > 0.3$ as a reference to obtain satisfactory items and items with power discrimination, the $d < 0.2$ deserve to be aborted (Widhiarso. 2010; Ebel in Azwar. 2001), but sometimes items with low discrimination power need to be involved in the selection process. The preparation of the test, considering that it is considered a vital content (Widhiarso. 2010). In addition, the research considers reviewing distractors for each item contained in the

IST subtest, so that in this study the items considered by the researcher as vital content can be accepted.

This study uses the KR-20 reliability coefficient approach to calculate the reliability of each subtest, and the omega coefficient (ω) which is an approach based on confirmatory factor analysis to calculate the estimated reliability of constructs or measuring instruments (Widhiarso. 2009). The analysis results found that the construct reliability index on the IST showed an omega coefficient (ω) is 0.845. This shows that the degree of consistency of the measurement is high and satisfactory.

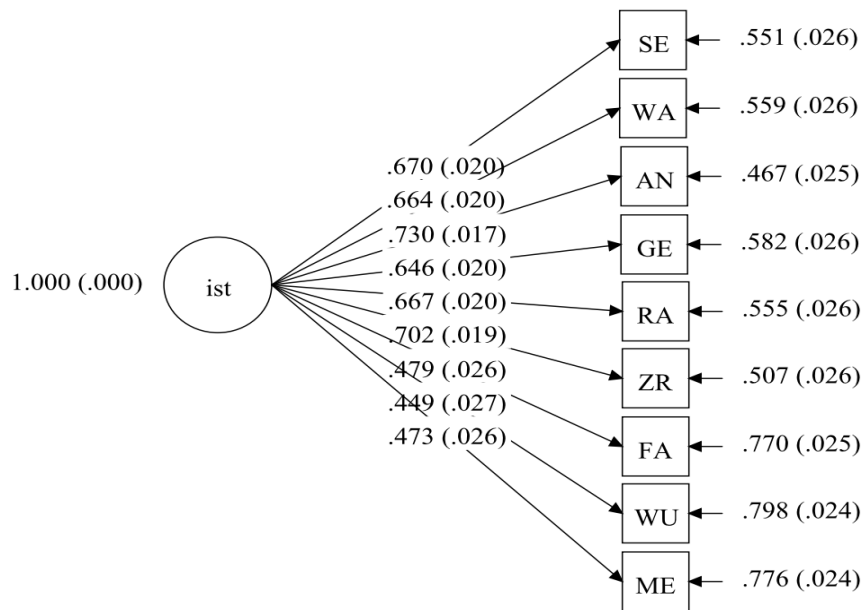


Figure 1. Confirmatory factor analysis of IST Subtest

Table 11. Goodness of fit index

Goodness of fit index	Cut of Value	Result Model
Chi Square	>0.05	241 ($p < 0.001$)
RMSEA	≤ 0.08	0.084
TLI	>0.90	0.905
CFI	>0.90	0.929

This study uses confirmatory factor analysis to test whether the model constructed on the IST is classified as FIT or not. The analysis results by using M. Plus version 7 shows that the FIT model is based on three goodness of fit index criteria, namely; RMSEA is 0.084, TLI is 0.905, and CFI is 0.929. The loading factor of each aspect moves from 0.449 to 0.730 and all factors are significant ($p < 0.05$). Thus, it can be concluded that the IST test instrument which consists of nine dimensions has a valid construct model to measure intelligence.

Table 12. Loading factor of Subtest

Subtes	KR-20 (α)	Loading faktor	SE	Z
SE	0.537	0.670	0.0776	23.6
WA	0.547	0.664	0.0826	23.4
AN	0.701	0.730	0.0830	26.5
GE	0.664	0.646	0.1337	22.6
RA	0.779	0.667	0.0819	23.4
ZR	0.851	0.702	0.1181	25.0
FA	0.673	0.479	0.1000	15.8
WU	0.790	0.449	0.1162	14.6
ME	0.816	0.473	0.1313	15.6

The correlation between subtests on the IST Test moves from 0.145 to 0.586. The subtests that have the highest correlation are RA and ZR of 0.586, SE and An of 0.534, thus it can be seen that the correlation of all dimensions has a low to moderate relationship. This shows that each subtest has its measuring function in measuring general intelligence.

Table 7. Intercorrelation between subtest

Subtes	Mean	SD	Interkorelasi sub tes							
			SE	WA	AN	GE	RA	ZR	FA	WU
SE	8.98	2.73								
WA	11.7	2.91	0.482							
AN	6.97	3.01	0.534	0.518						
GE	15.1	4.67	0.487	0.476	0.487					
RA	4.68	2.87	0.441	0.377	0.459	0.378				
ZR	8.75	4.20	0.388	0.438	0.483	0.416	0.586			
FA	10.0	3.29	0.298	0.286	0.330	0.270	0.348	0.359		
WU	9.86	3.79	0.252	0.268	0.314	0.238	0.317	0.388	0.363	
ME	9.92	4.32	0.299	0.323	0.317	0.348	0.310	0.375	0.216	0.145

CONCLUSION

The purpose of the study is to review the characteristics of psychometric properties on the IST, the conclusions which be drawn: the analysis of the item difficulty index shows that there are four subtests, namely: SE, AN, ZR, FA has a fairly satisfactory level of difficulty, three subtests, namely: WA, GE, WU which emphasizes the easy items more so that the distribution of the difficulty level is not very effective, and two subtests, namely: RA and ME each highlight difficult, and medium difficulty levels, so that the distribution also looks less effective. The item selection process based on item discrimination power and distractor effectiveness showed that in the SE subtest, 12 items were accepted and 8 items needed to be revised. In the WA subtest, there are 13 items received and seven items revised, in the AN subtest there are 17 items received and three items needing revision, in the GE, ZR, WU and ME subtests overall acceptable, in the RA subtest 19 items are accepted and one item

needs to be revised and on the FA subtest there are 18 acceptable items and two items need to be revised. The reliability of the test and the validity of the construct look satisfactory but still need to improve the reliability of each subtest. Thus, it can be seen that the IST 70 used still needs to be reviewed and some items that are no longer relevant to current conditions need to be revised again. In this study, the researchers used a classical theory approach to review psychometric properties, we hope that future researchers will review using the Item Response Theory 1 PL, 2 PL and 3 PL approaches to produce a more detailed conclusion.

ACKNOWLEDGEMENTS

Acknowledgements are addressed to the Directorate General of Higher Education for the PNPB grant No. 506/UN36/HK/2021 which has been given to support the smooth running of this research. Furthermore, thanks to the Chancellor of UNM and the Dean of the Faculty of Psychology UNM for their direction and guidance during the process of research activities. Likewise, I would like to thank the director of LPPT Widya Prasthya for his cooperation in this research.

REFERENCES

- Amthaeur, R. Tanpa tahun. Suatu pengenalan terhadap tes ist: intelligenz struktur test (IST-70). Terjemahan oleh Polhaupessy. 2008. Bandung: Fakultas Psikologi UNPAD.
- Anastasi, A. & Urbina, S. (2006). Tes psikologi (Edisi Ketujuh). Jakarta: PT. Indeks.
- Arikunto, S. (1991). Prosedur penelitian: suatu pendekatan praktek. Jakarta: Rineka Cipta.
- Azwar, S., Rustam, A. & Harjito, P. (1989). Laporan penelitian analisis item, pengujian reliabilitas, dan konsistensi internal subtes informasi dan subtes hitungan pada tes inteligensi WAIS adaptasi UGM. Yogyakarta: Fakultas Psikologi UGM
- Azwar, S. (2001). Tes prestasi: Fungsi dan pengembangan pengukuran prestasi belajar. Yogyakarta: Pustaka Pelajar.
- Azwar, S. (2008). Kualitas tes potensi akademik versi 07A. *Jurnal Penelitian dan Evaluasi Pendidikan*, No. 2 ISSN 1410-4725. Yogyakarta: Universitas Gadjah Mada.
- Cronbach. (1970). *Essential of psychological testing*. Newyork. Harper and Row Publisher.
- Gregory, R.J. (2000). *Psychological testing, history, principles, and aplication*. Massachussets: Allyn Bacon.
- Kaplan, R. M. & Saccuzzo, D. P. (2005). *Psychological testing*. USA : Thomson Learning, Inc.
- Lababa, D. (2008). Analisis butir soal dengan teori tes klasik: Sebuah Pengantar, Vol. 5 Januari – Juni 2008.
- Princen. (2011). Karakteristik psikometri subtes zahlenreihen (zr) pada intelligenz struktur test (IST). *Skripsi* (tidak diterbitkan). Medan: Fakultas Psikologi Universitas Sumatera Utara.
- Rahmawati, E.(2014). Evaluasi Karakteristik psikometri intelligenz struktur test. *Proceeding Seminar Nasional Psikometri*, hal 270-282

- Suryabrata, S. (2005). Pengembangan alat ukur psikologis. Yogyakarta: C.V. Andi Offset.
- Suwarto. (2007). Tingkat kesulitan, daya beda, reliabilitas tes menurut teori tes klasik. Program Studi Pendidikan Biologi-FKIP. *Jurnal Pendidikan Jilid 16, No. 2*. Sukoharjo: Universitas Veteran Bangun Nusantara Sukoharjo.
- Suwarto (2011). Teori tes klasik dan teori tes modern. Program Studi Pendidikan Biologi-FKIP Universitas veteran Bangun Nusantara Sukoharjo. (Online), Vol. 20 No.1 (<http://veteranbantara.ac.id/ejurnal/index.php/.../14/>, diakses 04 September 2021).
- Sukadji, S. (1998). Perkembangan konsep, teori dan pengukuran inteligensi. Depok: Universitas Indonesia.
- Srivastava, A.K. & Misra, G. (2007). Rethinking intelligence: conceptualizing human competence in cultural context. Delhi: University of Delhi.
- Widhiarso, W. (2010). Analisis butir dalam pengembangan pengukuran Psikologi (Online), (<http://blog.ugm.ac.id/author/wahyupsy/> , diakses 04 September 2021).
- Widhiarso, W. (2009). Koefisien reliabilitas pada pengukuran kepribadian yang bersifat multidimensi. *Psikobuana* Volume. 1, No. 1 39-48. ISSN 2085-4242. Yogyakarta: Fakultas Psikologi Universitas Gadjah Mada.
- Widhiarso, W. (2004). Evaluasi faktor dalam big five, pendekatan analisis faktor konfirmatori (Online), (http://widhiarso.staff.ugm.ac.id/analisis_faktor_konfirmatori/, diakses 05 September 2021).
- Widodo, P.M. (2006). Reliabilitas dan validitas konstruk skala konsep diri untuk mahasiswa indonesia. *Jurnal psikologi universitas diponegoro vol.3 No. 1*. Semarang: Program Studi Psikologi FK. Universitas Diponegoro.
- Wibowo, A. (2006). Pengantar analisis faktor eksploratori dan analisis faktor konfirmatori. Materi Pelatihan Structural Equation Modeling (Pemodelan Persamaan Struktural) *Angkatan XII*. Surabaya: Lembaga Penelitian dan Pengabdian Kepada Masyarakat Universitas Airlangga.