Humanitas Indonesian Psychological Journal
Vol. 16, No. 1, February 2019, 54-65

Psychometric properties of the SE-Revised: A rasch model analysis

[1]Siti Juniawati Rosa, [2]Ahyani Radhiani Fitri, [3]Ivan Muhammad Agung

[1,2,3]Faculty of Psychology Universitas Islam Negeri Sultan Syarif Kasim Riau, Indonesia
[1]siti.juniawati.rosa@students.uin-suska.ac.id, [2]ahyani.radhiani.fitri@uin-suska.ac.id,
[3]ivan.agung@uin-suska.ac.id

ARTICLE INFO

Article history
Received 17 August 2018
Revised 29 February 2019
Accepted 7 March 2019

Keywords
Intelligenz Struktur Test (IST)
Satz Ergaenzung (SE)
rasch model

ABSTRACT

Satz Ergaenzung (SE) is the first subtest in the *Intelligenz Struktur Test* (IST) that measures reasoning ability. Previous studies found that SE has suboptimum performance in measuring subject's ability. Therefore, this study intended to revise the SE items that were found to have poor psychometric properties, either by changing the questions and answer choices, the answer choices only, or changing the order of the questions. This study used the rasch model to determine the psychometric property of SE-revised. A total of 159 undergraduate students of a university in Riau participated in this study. The results showed that the revised SE had fulfilled the pre-requisites of unidimensionality. The instrument reliability increased after being revised, where the items of SE-revised correctly measure respondents' abilities. The revised version may be used as an improvement of the original SE, with better psychometric properties. However, there were four items with low difficulty index and two items that have DIF. Even though SE-revised is better than the original SE, further research is needed to revise the six items of SE-revised subtest.

Introduction

Intelligenz Struktur Test (IST) is an intelligence test developed by Rudolf Amthauer in Frankfurt, Germany in 1953 (Adinugroho, 2016; Wiratna, 1993). This intelligence test is classified as a speed test, which prioritizes speed and accuracy of work (Nur'aeni, 2012). In Indonesia, IST is still used quite often, particularly in workplace selection and placement in the workplace and education settings (Bawono, 2008; Hamidah, 2001; Princen, 2011; Rahmawati, 2014). Therefore, IST is expected to have good measurement performance, which can appropriately measure the test-taker's abilities. The IST currently used in Indonesia was adapted from IST-70 by Universitas Padjadjaran in 1973. As time goes by, this instrument needs regular evaluation to ensure that it accurately performs in measuring test takers' abilities. However, the IST currently in use has not undergone any revision, despite the original IST having already undergone two revisions from IST-70 to the latest version which is the IST 2000-Revised (Kipman, Kohlböck, & Weilguny, 2012).

The IST consists of 9 subtests, each of which can stand alone to measure specific abilities in individuals (Wahyuni, Widyastuti, & Fitriyani, 2015). Previous studies have measurement properties. Widianti (2008) tested for convergent validity where the SE subtest was correlated to the RA subtest. The results show that there is a significant

Corresponding Author siti.juniawati.rosa@student.uin-suska.ac.id

correlation between SE and RA, meaning that SE items can measure reasoning abilities. In testing discriminant validity by correlating SE and WU, it shows that SE correlates significantly with WU despite having a relatively weak correlation. This indicates that SE does not only measure one aspect of reasoning but also measures the ability measured by WU. In testing reliability, the results of previous studies show that SE has low internal consistency (Agung & Fitri, 2016; Widianti, 2008).

The results of the item parameter analysis indicate that the level of difficulty of SE items varied, with nine items being difficult (<.3) and three items being easy (>.7) (Agung & Fitri, 2016). Research by Elvira (2011) indicated that eleven items need to be improved, while research by Rahmawati (2014) shows that nine items need to be improved. Also, Suryani (2018) found that item number 20 was biased towards a particular gender, thus also suggested improvement.

Based on previous studies that found poor validity and reliability of SE as well as poor quality of the items, we developed an interest to revise the SE items and conduct psychometric tests on the revised SE. Evaluation of psychometric properties of the SE subtest was done in almost the same way as what has been done by previous researchers (Agung & Fitri, 2016; Elvira, 2011; Rahmawati, 2014; Widianti, 2008). However, previous researchers evaluated data obtained from the original version of the SE subtest, while this study evaluated the revised version of the SE subtest. Two evaluation processes were conducted by the researchers, namely evaluation by using a data bank obtained from the SE test results as the basis for revising items that have poor psychometric properties. After the revision was done, then the data obtained from the revised SE test were re-evaluated.

Evaluation is carried out using the Item Response Theory (IRT). IRT is utilized instead of the Classical Test Theory (CTT) which has the weakness of being test-dependent, which means that the ability of individuals is influenced by the characteristics of items in a test (Embretson & Reise, 2000; Fan, 1998). The ability of test-takers changes depending on different occasions when they take the test results in poor test consistency (Magno, 2009). Based on this explanation, the characteristics of items in CTT are influenced by test-takers abilities, and vice-versa, test-takers abilities are influenced by the characteristics of the items.

Unlike the CTT which focuses on the obtained scores, the IRT does not depend on particular sample of items or the person selected in the test (item free and person free), so that the measurements are more precise and the items can also be calibrated (Ariffin, et al., 2010; Sumintono & Widhiarso, 2014). The IRT assumes that in a test condition, the test taker's performance on the test can be predicted by defining the characteristics of the individual's trait or ability, estimating the test-taker's scores based on these traits (ability scores), and using the scores to predict or explain the items and test results (Hambleton & Swaminathan, 1985; Kubinger, Rasch, & Yanagida, 2011; Prieto, Alonso, & Lamarca, 2003).

Psychometric characteristics are quantitative attributes that relate to the strengths or weaknesses of the statistics obtained from tests or measurements, consisting of reliability, validity, and difficulty index (Embretson & Reise, 2000). The psychometric analysis conducted in this study used the rasch model approach, which consists of unidimensionality, reliability, item-fit order, item difficulty index, and differential item functioning (DIF). The rasch model provides various diagnostic information that allows researchers to recognize and diagnose the difficulties of the test and then suggest corrective actions that can improve the nature of test measurements (Curtis & Boman, 2007; Petrillo et al., 2015).

Unidimensionality means that only one attribute or ability is measured by a set of items in the test (Bond & Fox, 2015). Therefore, one instrument must be able to measure a particular ability of the test-taker. This assumption cannot strictly be met because there are always other factors that influence the implementation of tests, such as cognitive factors, personality, motivation, levels of anxiety, the ability to perform in fast-pace, and the tendency to guess answers when in doubt (Hambleton & Swaminathan, 1985; Hambleton, Swaminathan, & Rogers, 1992). However, there are circumstances where it is necessary to think of concepts in unidimensional terms so that comparisons can be made using the differences (Hagell, 2014). The minimum prerequisite of unidimensionality is 40% of the raw variance value, indicating good unidimensionality, while 60% means very good unidimensionality. The variance that cannot be explained by an instrument should ideally not exceed 15% (Sumintono & Widhiarso, 2014).

Reliability indicates the extent to which repeated measurements will produce the same information, meaning that it does not produce significant meaningful differences in information. Differences in information will always exist; therefore, convincing measurements do not have to produce the same information, rather, differences of very little value which can still be tolerated (Azwar, 2009, 2014; Sumintono & Widhiarso, 2014). The reliability of test scores range from 0–1, where r = 0 indicates no reliability, and r = 1 shows absolute reliability (Aiken & Marnat, 2008; Azwar, 2009).

In the rasch model approach, there are two reliability indexes: person reliability and item reliability (Boone, Staver, & Yale, 2014). Person reliability and item reliability are considered weak if reliability <.67, sufficient between .67-.80, good between .81-.90, excellent between .91-.94, and special if > .94 (Sumintono & Widhiarso, 2014, 2015).

Item-fit is a "quality-control mechanism" which explains whether items can measure certain variables according to the unidimensionality construct (Bond & Fox, 2015). The criteria that we used to check whether an item is fit are if the *Outfit Mean Square* (MNSQ) obtained score is .5<MNSQ<1.5; *Outfit Z-Standard* (ZSTD) obtained score is -2.0<ZSTD<+2.0; and *Point Measure Correlation* (*Pt Measure Corr*) obtained score is .4<*Pt Measure Corr*<.85 (Osman et al., 2012; Rashid, et al., 2008; Sumintono & Widhiarso, 2014, 2015).

According to Boone et al. (2014), the scores of MNSQ, ZSTD, and Pt Mean Corr are criteria used to see how the item's suitability measures the variables that should be measured. If the item does not meet the criteria for outfit MNSQ, outfit ZSTD, and the *point measure correlation*, it means that the item is not good enough and needs to be adjusted or replaced. This can be caused by an error of setting the wrong answer key, the many of individuals who are less motivated in working on the questions, and questions with low power difference that reduces the accuracy of the items (Sumintono & Widhiarso, 2015).

The item difficulty index (symbolized by *b)* is indicated from the logit score in the item measure table, which has been sequenced from the highest to the lowest logit score. High logit scores indicate high item difficulties (Sumintono & Widhiarso, 2015).The item measure provides information on the standard deviation score, which when combined with the logit mean allow grouping of items based on the difficulty (Sumintono & Widhiarso, 2015). For example, .0 logit +1SD is categorized as a difficult item, greater than +1SD is considered a very difficult item, .0 logit -1SD is considered an easy item, and smaller than -1SD is categorized as a very easy item. This means that there are four groups of items based on the level of difficulty.

Differential Item Functioning (DIF) is a crucial technique for analyzing survey data and tests (Boone et al., 2014). DIF serves to detect whether items contain biases based on

the respondent's demographic variables. This occurs when different groups in the sample (e.g., men and women) respond differently to each item (Pallant & Tennant, 2007). The bias in the item can be determined based on the value in the 'Prob.' Identification of the values in this column that are at or below .05 (the threshold used in the statistical analysis) shows that the relative locations of items differ between certain demographic variables, such as between male and female (Boone et al., 2014). This means that there are indications that this item is biased towards a demographic variable.

This study aims to look at the psychometric characteristics of the SE-revised subtest using the rasch model approach. This study is essential as, among the nine IST subtests, the SE subtest has problems in item parameters and reliability. By revising the SE subtest, it is expected there will be an improvement in the quality of its measurements, as well as the reliability of the IST test as a whole. Thus the measurement results obtained can be used as a basis of making decisions, both in the context of selection and placement.

Method

Respondents

Respondents of this study were 159 first-year undergraduate students from a university in Riau, which consisted of 46 men and 113 women. The age of respondents ranged between 17 to 22 years (mean = 19 and SD = .78).

Instrument

The measuring instrument utilized in this study was the SE-revised. Respondents were given 6 minutes to work on 20 questions of the SE-revised. The data was collected in the form of responses to the 20 items of the SE-revised. For each correct answer, a score of 1 is given, while for any wrong answer a score of 0 is given.

Procedure

This study was conducted in 4 stages: (I) Preliminary study, carried out by evaluating the psychometric characteristics of the SE subtest based on the rasch model, using the data bank obtained from IST testing on the 293 first-year students; (II) Revision of the SE subtest based on information of psychometric characteristics obtained from the preliminary stage; (III) Administration of SE-revised to respondents in a situation that imitates the real test condition; (IV) Analysis of respondents' answer to understand the psychometric characteristics of the SE-revised. Data analysis at the final stage analyzed with the rasch model. Stage I and II were the stages of constructing the SE-revised subtest. Stage III was the stage of data collection. Stage IV was the data analysis of psychometric characteristic of the SE revised.

Stage I. Based on the preliminary study that utilized data from an IST testing conducted on new students of a Faculty of Psychology at a university in Riau, the SE had fulfilled the instrument unidimensionality requirements (θ = 38.2%). This means that the SE subtest was able to measure the reasoning construct accordingly. However, results of the reliability test show that SE has very low reliability (α = .41) as a result of the interaction between the respondents' low reliability (α = .36) and the item reliability that is

classified as exceptional (α = .99). This shows that the respondent's ability is lower than the difficulty of the item.

Rasch analysis was conducted to identify the SE items that need to be adjusted. The test results indicated one misfit item (item number 17), five items with a low difficulty index (items number 2, 3, 4, 12, 18), and one item infected with DIF (item number 16). Item number 17 was classified as a misfit (MNSQ=1.54, ZSTD=2.5, *Pt Mean Corr*=.15); indicating that the item is not suitable for measuring the reasoning variable. Whereas item number 16 was found to have been infected with DIF (.0063 <5%) in the male category, meaning that the item was considered more difficult to answer by the male respondent group.

Stage II. Several items that require adjustment were items number 2, 3, 4, 12, 16, 17, and 18. Revisions to the SE subtest were conducted by analyzing item by item qualitatively through focus group discussions (FGD) with psychologists, psychometric professors, and students who were researching the field of psychometry. Changes to the items were based on collective decisions made in the FGD. Items number 12 and 18 were revised by changing the question along with its answer choices. Item number 16 was revised only in one of the answer choices that was mistyped and might create confusion for respondents. In item 17 (misfit), one of the answer choices was changed. There were no changes in items number 2, 3, and 4 because it is assumed that the difficulty index of these items will increase when the SE item order is sorted from easy to difficult.

During this revision process, the SE items were also reordered based on the item difficulty index from easy to difficult, in accordance with the provisions of intelligence tests that order its items from the easiest to the most difficult (Murphy & Davidshofer, 2003). This allows the respondents to answer easy items first. The order of items based on the level of difficulty is: 2, 4, 3, 6, 10, 11, 8, 9, 1, 7, 14, 19, 15, 13, 20, 5, 16, 17, 18, 12.

Data analysis

Data analysis was carried out using computerized rasch model through the Winstep 3.73 for Windows application program, which produces results that has been sorted based on difficulty level - from the highest difficulty level to the lowest difficulty level, making it easier to identify which questions are difficult and which questions are easy (Suryani, 2018). The data analysis includes analysis of unidimensionality, reliability, fit items, item difficulty index, and DIF.

Results

The results of the analysis on SE-revised provided various information, both in terms of the instruments and items. Table 1 shows a raw variance of 38.5%, which is not much different from the expected 38.7% - very close to the unidimensionality requirement of 40%. This indicates that the SE-revised is capable of measuring reasoning abilities. This raw variance has undergone a slight increase from 38.2%, prior to the revision.

e-ISSN 2598-6368/ISSN 1693-7236

Table 1
Test Results of the Unidimensionality of SE-Revised

		Empirical		Modeled
Total raw variance in observations =	32.5	100.0%		100.0%
Raw variance explained by measures =	12.5	38.5%		38.7%
Raw variance explained by persons =	5.2	15.8%		15.9%
Raw variance explained by items =	7.4	22.7%		22.8%
Raw unexplained variance (total) =	20.0	61.5%	100.0%	61.3%
Unexplained variance in 1st contrast =	1.8	5.4%	8.9%	
Unexplained variance in 2nd contrast =	1.7	5.1%	8.3%	
Unexplained variance in 3rd contrast =	1.5	4.5%	7.3%	
Unexplained variance in 4th contrast =	1.4	4.2%	6.9%	
Unexplained variance in 5th contrast =	1.3	4.1%	6.7%	

As in Table 1, the value of the raw unexplained variance is 61.5%, indicating the magnitude of other factors that also affect the test, such as cognitive factors, personality, motivation, and anxiety (Hambleton & Swaminathan, 1985; Hambleton et al., 1992). From the unexplained variance, it can be seen that all percentages are below 10%, meaning that the independence of the items in the test is classified as good (Wibisono, 2016).

The results of the reliability analysis contain two outputs, namely person reliability and item reliability, as shown in Table 2. The Cronbach alpha is defined as the reliability of the interaction between the person and item reliability. In Table 2, α=.61 which can be considered as sufficient. Person reliability is .65, meaning that the ability of respondents who worked on the SE-revised test is classified as weak. In addition, the item reliability is .98, meaning that items of the SE-revised are classified as exceptional. The average of the measure value of the person table is -.01 (μ<.00), which indicates that the respondents have lower ability relative to the item difficulty level.

The test reliability indicated a good increase, from .41 which was classified as poor to .61 which falls under the category of sufficient. This increase in reliability occurred in accordance with the increase of person reliability from .36 to .65 after being revised, though both are still considered relatively weak.

Table 2
Test Results of Respondents and Item Reliability of SE Before and After Revision

	Before Revision	After Revision
Cronbach Alpha	.41	.61
Person Reliability	.36	.65
Item Reliability	.99	.98

Table 3 shows that all items of the SE-revised subtest are fit. Most of these items have fulfilled one or more of the suggested criteria. The MNSQ, ZSTD, and *pt mean corr values* of each item, particularly for item number 17. Analysis of item fit indicated an increase in the quality of items after being revised. Item number 17 (or item number 18 in SE-revised) changed in the value of MNSQ Outfit=1.54 (>1.5) to 1.01, Outfit ZSTD=2.5 (>2.0) to .1, and *Pt Mean Corr* = .15 (<.4) to .34 (<.4). *Pt Mean Corr* value of item number 17 that has been revised does not match the required criteria, but the other two criteria (MNSQ Outfit and Outfit ZSTD) have been fulfilled; thus the item can still be used and does not need to be discarded.

Based on the results of the item difficulty estimation, a standard deviation (SD) of 1.98 is obtained in the item measure. Through SD value, items can be grouped based on the level

of difficulty. The value of *b*=>1.98 is classified as very difficult, *b*=0.0-1.98 is classified as difficult, *b*=-1.98–0.0 is relatively easy, *b*=<-1.98 is classified as very easy. Items 1 to 20 has a non-sequential difficulty index. Table 3 shows three items have a logit value smaller than -2, so the items were classified as very easy.

Analysis on the item difficulty index shows some items have better index than the original form. The difficulty index of item number 12 (item number 20 in SE-revised) is better with a value of *b*=1.63 (-2<b<+2), while for item number 18 (item number 19 in SE-revised) with changes in the value of *b*=1.31. In addition, one item that was classified as difficult was found after revision, that is item number 16 (item number 17 in SE-revised) with a value of *b*=2.59.

Bias analysis was carried out in the gender category. Based on Table 3, there were two items that are biased towards gender, that are item 8 (.0026<.05) and item 16 (.0143<.05). The results of the DIF testing show that item 16 (item number 5 in the original SE) easier for male respondents to answer. Item number 8 (item number 9 in the original SE) is more favorable for women, which means that the male respondents find it more difficult to answer this question. That being said, the probability of male respondents answering item number 5 correctly is greater than the probability of female respondents. Meanwhile, for item number 9, the probability of male respondents answering correctly is smaller.

e-ISSN 2598-6368/ISSN 1693-7236

Table 3
Summary of Psychometric Characteristics of Original and Revised SE Item Parameters

Item Number		Outfit				Pt Measure Corr		IKA		Prob.		Conclusion
		MNSQ		ZSTD								
Original	Revised	Original	Revised	Original	Revised	Original	Revised	Original	Revised	Original	Revised	
1	9	.18	.91	-1.0	-1.6	.27	.41	-.12	.26	.8969	.9029	Good
2	1	.31	1.00	-1.5	.1	.45	.35	-4.50	-4.98	.2200	.4249	Needs Review
3	3	.54	1.23	-.6	1.8	.28	.19	-2.07	-3.98	.9120	.3760	Needs Review
4	2	1.20	.53	1.8	-1.0	.27	.22	-3.76	-3.60	.5140	.0630	Needs Review
5	16	1.02	.99	.3	.0	.54	.24	1.37	1.51	.5392	.0143	Needs Review
6	4	1.00	.83	.0	-.4	.42	.23	-1.38	-.60	1.0000	.2780	Good
7	10	1.10	.97	1.1	-.7	.32	.38	-.12	.45	.9222	.9534	Good
8	7	.85	1.08	-1.6	1.5	.51	.26	-.39	-.28	.8854	.5397	Good
9	8	.90	1.01	-.9	.3	.46	.32	-.17	.08	1.0000	.0026	Needs Review
10	5	1.00	1.00	.0	.0	.36	.24	-.60	-.16	.2835	.9336	Good
11	6	.97	1.08	-.3	1.0	.42	.24	-.40	-.84	.7833	.4872	Good
12	20	1.02	.79	.2	-1.6	.34	.40	2.75	1.63	.7248	.4390	Good
13	14	1.18	1.02	1.1	.5	.27	.33	1.16	1.68	.0756	.5116	Good
14	11	1.37	.94	1.5	-1.3	.15	.42	-.10	-.04	.4356	.4841	Good
15	13	.77	1.02	-.8	.2	.31	.24	.95	1.06	.8957	.5806	Good
16	17	.98	1.54	.0	2.5	.28	.15	1.55	2.59	.0063	.2090	Good
17	18	.79	1.04	-.5	.2	.25	.17	1.71	.99	.4404	.3600	Good
18	19	1.01	.92	.1	-.9	.34	.39	2.36	1.31	.7148	.9479	Good
19	12	.96	1.03	-.1	.2	.33	.13	.56	.96	.7058	.5324	Needs Review
20	15	1.03	1.03	.2	.3	.27	.26	1.21	1.96	.5197	.9327	Good

Discussion

The main objective of this study was to examine the psychometric characteristics of the SE-revised using the rasch model, which focused on testing unidimensionality, reliability, item fit, item difficulty, and DIF. When compared to its quality before revision, the SE-revised have a general increase in the quality of its items and instrument, indicated by the unidimensionality and reliability of the instrument as well as the fact that there no more misfit items observed. This examination of unidimensionality is very important to see the extent to which items provide independent information, for the SE is aspects of reasoning (Ireland, Goh, & Ida, 2018; Kubinger, Rasch, & Yanagida, 2011). In any case, where there are signs of other dimensions measured, the rasch model will indicate which items have the potential to contribute to these 'other' dimensions, thus directing researchers to further investigate then replace or maintain the items (Ishak, Osman, Mahaiyadin, Tumiran, & Anas, 2018).

The increase in the general reliability of the test in accordance with the increase of person reliability after being revised. Two items, that are number 15 and 19, were changed based on the agreement of FGD participants. After revision, items number 15 and 19 were found to have good psychometric characteristics as seen from the results of the conformity test (item fit), difficulty, and DIF.

The overall quality of the SE-revised items is quite good because there are no misfit items observed. This finding is in line with previous studies which found several items in the original SE that were not functioning properly and needed improvement (Elvira, 2011; Rahmawati, 2014; Suryani, 2018). After the revision was carried out, all items of the SE was deemed fit, meaning that there was an increase in the quality of items after being revised.

The evaluation results of the SE-revised shows that the items have sufficient ability to measure the reasoning variable. However, 30% of the items did not meet the item difficulty index and DIF criteria and require a review. Various factors may greatly influence the response of the answers given by respondents, for example, factors related to test administration related, the situation of the testing, and the form of the test equipment used when conducting the trial. Azwar (2016) explains that in carrying out tests for data collection trials, the administrative situation needs to be considered and should be executed like the actual test. Meanwhile, the data collection process in this study did not strictly control the situation and condition of the test - for instance in terms of execution time, room temperature, and sitting position of the test takers.

Other assumptions that may influence respondents' answers are the trial respondents who were already familiar with the given questions. This is in line with Rahmawati (2014) opinion that individuals may be familiar due to the fact that it has been 40 years since the test was first adapted in Indonesia. This also indicates allegations that this test has been leaked in the general community (Rahmawati, 2014), for example, we can easily find problem examples along with an explanation on how to answer it on the internet.

As an implication of this research, the SE-Revised can be used in testing. Due to the improvement of its psychometric characteristics, SE-revised can be used as a replacement of the old version of SE, specifically in the context of testing new students at one of the universities in Riau, consequently leading to the attainment of more accurate results. The results of this study need to be improved by further research to obtain better results.

This study has several disadvantages including; First, research respondents were less representative of the population intended, especially in terms of age and education level. Therefore, the results of the study cannot be generalized, and the SE-revised can only be

used for the characteristics of the test participants in accordance with the respondents of this study. Second, lack of respondents to obtain quality results of a measuring instrument, so that even though the reliability of the SE-revised has increased compared to the old SE, the reliability score obtained is not ideal. Third, convergent and discriminant validity testing was not conducted.

Based on the results of this study, the following are suggestions for future research. First, to revise the six items that require review, to improve its psychometric characteristics. Second, revise items of the SE subtest that has low quality by replacing the questions along with its answer choices, including replacing words or terms that are rarely used. Third, research on SE-revised should be performed repeatedly to produce tests and items with good psychometric characteristics. Fourth, the development of similar test tools that measure reasoning construct and then evaluate its psychometric property with item response theory (IRT).

Conclusion

Evaluation of the SE-revised using rasch model indicated better psychometric quality compared to the original SE. This is indicated by an increase in unidimensionality and reliability. However, the quality of some SE-revised items need to be improved. Therefore, this test is the first step to improving the quality of SE items; hence further research is needed to be able to obtain better psychometric measures of SE.

References

Adinugroho, I. (2016). Pengujian properti psikometrik Intelligenz Struktur Test subtes kemampuan spasial dua dimensi (form AUSWAHL): Studi pada dua SMA swasta di Jakarta (Psychometric properties of Intelligenz Structure Test two dimensions spatial ability subtest (form AUSWAHL): Study in two senior high schools in Jakarta*). Jurnal Ilmiah Psikologi MANASA, 5*(2), 165–180.

Agung, I. M., & Fitri, A. R. (2016). Analisis Psikometri dan Standardisasi Norma pada Tes Inteligensi Struktur Test (IST) pada Mahasiswa UIN Sultan Syarif Kasim Riau (Psychometric analysis and norm standardization on Intelligent Structure Test (IST) aming UIN Sultan Syarif Kasim undergraduate students). *Research report.* Pekanbaru: LP2M UIN Suska Riau.

Aiken, L. R., & Marnat, G. G. (2008). *Pengetesan dan pemeriksaan psikologi (Psychological testing and assessment)*. (B. Sarwiji, Ed.) (12th ed.). Jakarta: Indeks.

Anastasi, A., Urbina, S. (2007). *Tes psikologi (Psychological testing)*. (R. H. S. Imam, Ed.) (7th ed.). Jakarta: Indeks.

Ariffin, S. R., Omar, B., Isa, A., & Sharif, S. (2010). Validity and reliability Multiple Intelligent item using Rasch measurement model. *Procedia - Social and Behavioral Sciences*, *9*, 729–733. https://doi.org/10.1016/j.sbspro.2010.12.225

Azwar, S. (2009). *Reliabilitas dan validitas (Reliability and validity)*. Yogyakarta: Pustaka Pelajar.

Azwar, S. (2014). *Dasar-dasar psikometri (Basic psychometry)*. Yogyakarta: Pustaka Pelajar.

Azwar, S. (2016). *Konstruksi tes kemampuan kognitif (Construction of cognitive ability test)*. Yogyakarta: Pustaka Pelajar.

Bawono, B. A. (2008). *Uji Aspek-Aspek Psikometrik Subtes Merkaufgaben dari Baterai*

Intelligenz Struktur Test (Psychometric analysis of Intelligent Structure Test Merkaufgaben Subtest). (Master Thesis), Universitas Katolik Atma Jaya, Jakarta.

Bond, T. G., & Fox, C. M. (2015). *Applying the Rasch Model:Fundamental Measurement in the Human Sciences*. *International Journal of Testing*. https://doi.org/10.1207/S15327574IJT013&4_10

Boone, W. J., Staver, J. R., & Yale, M. S. (2014). *Rasch analysis in the human sciences*. London: Springer. https://doi.org/10.1007/978-94-007-6857-4

Curtis, D., & Boman, P. (2007). X-ray your data with rasch. *International Education Journal*, *8*(2), 249–259. https://doi.org/10.1080/00986440902900295

Elvira, R. (2011). *Karakteristik Psikometri Subtes Santzerganzung (SE) Pada Intelligenz Struktur Test (IST) (Psychometric characteristics of Intelligenz Structure Test (IST))*. (Undergraduate Thesis), Universitas Sumatera Utara, Medan.

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for Psychologists multivariate applications book series*. New Jersey: Lawrence Erlbaum Associates, Inc. https://doi.org/10.1016/j.ica.2011.03.035

Fan, X. (1998). Item response theory and classical test theory: An empirical comparison of their item/person statistics. *Educational and Psychological Measurement*, *58*(3), 357–374.

Hagell, P. (2014). Testing rating scale unidimensionality using the principal component analysis (PCA)/t-test protocol with the rasch model: The primacy of theory over statistics. *Open Journal of Statistics*, *4*(6), 456–465. https://doi.org/10.4236/ojs.2014.46044

Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory principles and applications*. (G. F. Madaus & D. L. Stufflebeam, Eds.) (1st ed.). New York: Springer Science+Business Media, LLC. https://doi.org/10.1007/978-94-017-1988-9

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1992). *Fundamentals of item response theory. Contemporary Sociology (Vol. 21)*. *Contemporary Sociology* (Vol. 21). London: Sage Publications. https://doi.org/10.2307/2075521

Hamidah. (2001). Uji Validitas dan Reliabilitas Item Tes IST (Intelligenz Structure Test) (Validity and reliability of IST). *Research Report.* Surabaya: Universitas Airlangga.

Ireland, M. J., Goh, H. E., & Ida, M. (2018). A rasch model analysis of the emotion regulation questionnaire. *Journal of Applied Measurement*, *19*(3), 258–270.

Ishak, A. H., Osman, M. R., Mahaiyadin, M. H., Tumiran, M. A., & Anas, N. (2018). Examining unidimensionality of psychometric properties via rasch model. *International Journal of Civil Engineering and Technology (IJCIET)*, *9*(9), 1462–1467.

Kipman, U., Kohlböck, G., & Weilguny, W. (2012). *Psychologische testverfahrenzur messung intellektueller begabung*. (C. Rech & A. Fritz, Eds.). Salzburg: Österreichisches Zentrum für Begabtenförderung und Begabungsforschung (ÖZBF).

Kubinger, K. D., Rasch, D., & Yanagida, T. (2011). A New Approach for Testing the Rasch Model. *Educational Reserach and Evaluation*, *17*(5), 321–333. https://doi.org/10.1080/13803611.2011.630529

Magno, C. (2009). Demonstrating the difference between classical test theory and item response theory using derived test data. *The International Journal of Educational and Psychological Assessment, 1*(1), 1–11.

Maseko, J. (2008). *Applying the rasch model to validate a test instrument of students' rational number cognition.* Mpumalanga: Paper presented in the 9th Annual Unisa Iste Conference On Mathematics, Science And Technology Education.

Murphy, K. R., & Davidshofer, C. O. (2003). *Psychological testing: Principles and application*. New Jersey: Prentice-Hall, Inc.

Nur'aeni. (2012). *Tes psikologi: Tes inteligensi dan tes bakat (Psychological test: Intelligence and aptitude test)*. (T. Trianton, Ed.). Yogyakarta: Universitas Muhammadiyah (UM) Purwokerto Press.

Osman, S. A., Naam, S. I. N., Jaafar, O., Badaruzzaman, W. H. W., & Rahmat, R. A. A. O. K. (2012). Application of Rasch Model in Measuring Students' Performance in Civil Engineering Design II Course. *Procedia - Social and Behavioral Sciences, 56*(Ictlhe), 59–66. https://doi.org/10.1016/j.sbspro.2012.09.632

Pallant, J. F., & Tennant, A. (2007). An introduction to the Rasch measurement model: An example using the Hospital Anxiety and Depression Scale (HADS). *British Journal of Clinical Psychology*, *46*(1), 1–18. https://doi.org/10.1348/014466506X96931

Petrillo, J., Cano, S. J., McLeod, L. D., & Coon, C. D. (2015). Using Classical Test Theory, Item Response Theory, and Rasch Measurement Theory to Evaluate Patient-Reported Outcome Measures: A Comparison of Worked Examples. *Value in Health*, *18*(1), 25–34. https://doi.org/10.1016/j.jval.2014.10.005

Prieto, L., Alonso, J., & Lamarca, R. (2003). Classical Test Theory Versus Rasch Analysis for Quality of Life Questionnaire Reduction. *Health and Quality of Life Outcomes*, *1*(27), 1–13.

Princen. (2011). *Karakteristik Psikometri Subtes Zahlenreihen (ZR) pada Intelligenz Struktur Test (IST) (Psychometric characteristics of the Intelligenz Structure Test Zehlenreihen Subtest)*. (Undergraduate Thesis), Universitas Sumatera Utara, Medan.

Rahmawati, E. (2014). *Evaluasi karakteristik psikometri intelligenz struktur test (IST) Psychometric characteristics evaluation of the Intelligenz Structure Test).* Surakarta: Paper presented in the Seminar Nasional Psikometri (National psychometric conference).

Rashid, R. A., Abdullah, R., Ghulman, H. A., & Masodi, M. S. (2008). Application of rasch-based ESPEGS model in measuring generic skills of engineering students : A new paradigm. *Engineering Education*, *5*(8), 591–602. https://doi.org/10.1016/j.jclinepi.2012.08.003

Sumintono, B., & Widhiarso, W. (2014). *Aplikasi model rasch untuk penelitian ilmu-ilmu sosial (Application of rasch model in social sciences research)*. Cimahi: Trim Komunikata.

Sumintono, B., & Widhiarso, W. (2015). *Aplikasi pemodelan rasch pada assessment pendidikan (Rasch model application in educational assessment)*. Cimahi: Trim Komunikata.

Suryani, Y. E. (2018). Aplikasi rasch model dalam mengevaluasi Intelligenz Structure Test (IST). *Psikohumaniora: Jurnal Penelitian Psikologi*, *3*(1), 73–100. https://doi.org/10.21580/pjpp.v3i1.2052

Wahyuni, S., Widyastuti, A., & Fitriyani, E. (2015). *Metode pengukuran bakat dan inteligensi (Method of measuring aptitude and intelligence)*. Pekanbaru: Al-Mujtahadah Press.

Wibisono, S. (2016). Aplikasi model rasch untuk validasi instrumen pengukuran fundamentalisme agama bagi responden muslim (Rasch model applicaton for validating a measurement on religious fundamentalism among Muslims). *Jurnal Pengukuran Psikologi Dan Pendidikan Indonesia*, *V*(1), 1–29.

Widianti, P. (2008). *Pengujian aspek-aspek psikometris subtes Satzergänzung (SE) dari Baterai Tes Intelligenz-Struktur-Test (IST) (Psychometric aspects of Intelligenz Structure Test Satzergänzung (SE) subtest)*. (Master Thesis), Universitas Katolik Indonesia Atma Jaya, Jakarta.

Wiratna, A. (1993). *Intelligenz Struktur Tes (Intelligenz Structure Test)*. Surabaya: P.T. Locita Mandayaguna.