JISAE
Online ISSN : 2597-8943
Print ISSN : 2442-4919
Journal of Indonesian Student Assessment and Evaluation

ANALYSIS OF CLASICAL TEST THEORY (CTT) APPROCH ON ACADEMIC ABILITY TEST INSTRUMENT

Dinar Pratama
Fakultas Tarbiyah IAIN Syaikh Abdurrahman Siddik Bangka Belitung
Jalan Raya Petaling KM. 13 Kec. Mendo Barat Kab. Bangka
email: dinarpratama24@gmail.com

ABSTRACT

The purpose of this study to estimate the parameters in Clasical Test Theory (CTT) approach on Academic Ability Test Instrument new students of IAIN Syaikh Abdurrahman Siddik Bangka Belitung academic year 2018/2019. Data was collected through documentation techniques in the form of 425 sheets of student test answers. Based on the results of data analysis, it is known that the index of difficulty level in all problem fields has not shown a balance of comparison of easy, medium and difficult questions. The average questions are distributed in the medium category questions with a percentage of 25% easy questions, 51% medium questions, and 24% difficult questions. Distinguishing power index, in all question areas shows there are 37% of questions that are able to distinguish test takers 'abilities and as many as 18% questions are not able to distinguish test takers' abilities. While the effectiveness of distractors there are 239 or 79.6% functioning and there are 61 or 20.3% of the non-functioning distractors. The results of the validity analysis of 100 questions obtained 54% which have a coefficient of validity more than 1.59 and as many as 46% that have a coefficient value less than 1.59. For the overall reliability coefficient, a value of 0.81 is obtained. This means that 80% of the difference in scores obtained by test takers with others is their pure score difference or is not influenced by other factors as a source of error in the measurement.

Keywords*:* clasical theory test, test instruments.

In the event the selection of new admissions in Universities, the test is one of the instruments used to measure the ability of prospective students. Test instrument is ideally used not only to determine the acquisition of the highest score or the lowest score of the test participants as the basis for the college in determining the threshold to pass or not prospective students. However, there is a lot of information that can actually be seen from the results of measurement of a test instrument. The quality of the test that are less good will not be a lot to give meaningful information. Even the measurement results can be declared invalid. Practically, there is a possibility of in determining the pass or the participants of the test. For example, grain test which has the distinguishing features are less good then the grains are not able to distinguish test-takers who have the ability high or low. Usually the point about the discriminating of less good will produce a negative score. This shows that, the test participants with the ability low be able to answer correctly. While test-takers with high ability answered incorrectly.

On the selection of new admissions in universities, test instruments should have good quality. It is directly related to the quality of prospective students. Each college course has a minimum standard for acceptance of prospective students. If the test that is used to perform the selection of candidates students have the ugly quality, of standards the prospective students desired by the college will not be achieved. Zucker, (2003) as quoted by Azwar, (2008) revealed that, in order for a test to function effectively at least the test has three such quality, reliable, valid, and unbiased.

According to Nitko, (2001) test is defined as an instrument or systematic procedure for observing and describing one or more characteristics of a student using either a numerical scale or a classification schame. Indrakusuma as quoted Daryanto, (2012) defines the test as a tool or a procedure of systematic and objective to obtain data or information desired about a person by the way that can be said quickly and precisely. While Norman, as quoted Djaali dan Pudji Muljono, (2008) suggested that the test is one of the evaluation procedure of comprehensive, systematic, and objective result can be used as the basis of decision making.

So the test can provide a picture of the person's ability, then, the development of the test need to pay attention to the rules that apply in the preparation of the test. According to Djaali dan Pudji Muljono, (2008) development of minimal test follow the steps as; a) goal setting test; b) curriculum analysis; c) analysis of material and source of support; d) develop a lattice; e) draw up the details of the problem; f) trials of the test; g) the analysis of the test results; h) revision of the question; and i) make about the results of the revision.

The above stages is a minimum standard in the preparation of the test. The most important thing from the steps of the preparation of the above tests are pilot tests. That is, the tests which have been compiled based on the lattice should be tested first before use. If there is a grain tests are less functional then the item should be revised. So that the grains of the tests which have good quality are used to measure the ability of a person. The main thing that need to be considered in the preparation of a test instrument is the aspect of the validity of such tests. The validity of the test refers to the quality of the test itself, whether such tests can measure what should be measured. Generally a test to measure the maximum capability of a person. So as to avoid wrong interpretation of the results of the test then a test must meet the criteria as a good test.

According to Surapranata, (2009) to determine the quality of a test can be done in two ways, namely through the analysis of qualitative and quantitative analysis. Qualitative analysis in terms of technical writing, materials, construction, and language. While the quantitative analysis emphasis on the analysis of the internal characteristics of the test through the data obtained empirically. The internal characteristics quantitatively intended to include parameters about the level of difficulty, distinguishing features, and reliability (Surapranata, 2009)

To analyze the test instrument are qualitatively, the test can be seen from how the right test include the purpose or area is measured and the test material in accordance with the lattice developed (Saifuddin Azwar, 2009) While for the analysis of quantitative test should be tested first before the test is used. Based on the theory of classical test analysis about at least include the index of difficulty, index of discrimination, the effectiveness of the rapscallion, and reliability.

Clasical test theory (CTT) is one method that can be used to determine the quality of an instrument. The basic concepts of CTT is formulated with formula X=T+E, where X is the score of the object, T is the score actually, and E is a score measurement error. According to Mistiani, (2016) each test taker will have a test score is actually if there is no measurement error. The following will be described the parameter in the method of CTT, which consists of the difficulty level, the discrimination, the effectiveness of the rapscallion, and reliability.

The difficulty level of the items shows the proportion of students who answer yes in the matter of which is carried out using an objective test (Sukardi, 2010) The difficulty level of the test items are generally shown with the percentage of students who obtained answers the item correctly. According to Surapranata, (2009) the difficulty level can be expressed through several ways including, 1) the proportion answered correctly, 2) the scale of the difficulties linear, 3) index davis, and 4) the scale of the bivariate. The equation used to determine the level of difficulty with the proportions answering correctly were:

$$p = \frac{\sum x}{S_m N}$$

Ket:

p = The proportions answering true or difficulty level

$\sum x$ = The number of test takers who answered correctly

S_m = The maximum score

N = The number of participants test

Index difficulty levels are usually distinguished into three categories; items with p <0.3 in the category of item difficulty, item with p> 0.7 easily enter the category of items, and items with p between 0.3 to 0.7 in the category of matter being. In the test instrument according to Sudjana in Syriac, (2017) level of difficulty of items should have a balance between the item easily, simply, and difficult with a ratio of 3: 4: 3 or 3: 5: 2. For example, if there is a numbered item 50 item comparison easy: simply: difficult is 15:20:15 or 11:28:11. In addition to an index level of difficulty, classical test theory can also estimate the item distinguishing. According to Barnard (1999) as quoted Sukardi, (2010), distinguishing index or coefficient is a number that provides information on distinguishing them individually, including distinguishing between high achievement of students with low achievement of students in a test. Distinguishing index, used mainly in reference norm is to distinguish between who is able and who is not. The amount ranging from -1 to + 1. The meaning of a positive price is that the material master answered correctly, and that do not master answered incorrectly. Vice versa if the score of the index is negative (Mardapi, 2012)

Distinguishing index is calculated based on the division of the group into two parts, namely the top of which is a group of highly capable test takers with a group under the group of low ability. According to Kelley (1939), Crocker, and Algina (1986) as quote Surapranata, (2009) division of the top group and a lower group of the most stable and sensitive as well as the most widely used is to determine the 27% upper group and 27% lower group. Crocker, and Algina (1986) in Azwar, (1993) items having distinguishing good if it has a coefficient greater than r_{bis} = 0.200.

A formula that can be used to calculate the index the following distinguishing features:

$$D = \frac{\sum A}{n_A} - \frac{\sum B}{n_B}$$

Ket:

D = Index tests distinguishing

$\sum A$ = number of participants who answered correctly on the top group

$\sum B$ = number of participants who answered correctly on the lower group

n_A = Number of participants on the top group

n_B = Number of participants on the lower group

The test instrument with the form of multiple choice questions generally has stem and response options. In this case there is only one correct answer and the other answer choices are just as distractors in the form of multiple choice questions, the position is very important distractors humbug. According Surapranata, (2009), distractors serves as a participant identifier that high-ability test. Distractor has function effectively if preferred by the test taker from the lower group. Conversely, when the distractor was selected by the test taker from the above group, the distractor was not working properly. A distractor can function well if at least 5% of participants selected by the test. If the distractor elected by all the participants of the test can be classed as a good distractor. According to Azwar, (1993), a good distractor should be selected by the person taking the test in the low group.

According Nitko (1983) as quoted by Surapranata, (2009), the criteria for determining which items are either very dependent on the intended use of the test itself. Whether for general purpose or specific objective. Overall, a good test instrument criteria indicated by the value of coefficient of reliability. Uno, (2010) emphasis on the notion of reliability as a consistency test. That is, how consistent test scores from one measurement to the next measurement.

Reliability refers to the provision of these tools in assessing what is desired, that is to say the ability of the tool used will give relatively similar results. The test instrument is said to be reliable if the results remain when the measurement were taken repeatedly. If the students given the same test at different times, then each student will remain in the same order or steady in his group (Widoyoko, 2009) According Kirk and Miller (1986) as quoted by Golafshani, (2003) "*identify three types of reliability referred to in quantitative research, which relate to: (1) the degree to which a measurement, given repeatedly, remains the same (2) the stability of a measurement over time; and (3) the similarity of measurements within a given time period*".

In estimating the reliability of the test there are several factors that can affect the reliability of the test, so the test is not reliable. In general, the reliability of a test is influenced by the differences idividu. Sometimes reliability is influenced by factors that permanently or factor that occurs due to temporary factors such as fatigue, conjecture, or the effects of exercise. According Arikunto, (2009) many factors that affect the reliability of the test bit, such as matters relating to the test itself, the test length and quality of the grain of the problem. The test consists of many grains, of course, is more valid than the tests that only consists of a few questions grains. High and low validity indicates the high and low reliability of the test. Thus, the longer the test, then the higher reliability.

To measure the reliability of a test can be using the formula coefficient Alpha Crombach, Kuder-Ricardson (KR-20 or KR-21), and techniques spilt half. Mardapi, (2012), revealed that to determine the coefficient of reliability of the test in the form of multiple choice score dichotomy better use Kuder-Ricardson formula (KR-20 or KR-21) The following KR-20 formula to calculate the score dichotomy.

$$KR-20 = \frac{k}{k-1}\left(\frac{S^2X-\sum_1^J P_i^2(1-P_i}{S^2X}\right)$$

According to Linn in Mansour et al (2009: 24), as quoted Iskandar & Rizal, (2018) suggested that the minimum limit of reliability coefficient value of at least 0.70. Even though it limits the coefficient value does not default, because each different researchers in determining the standard reference reliabillitas instruments.

State Islamic Institute Syaikh Abdurrahman Siddik Bangka Belitung is a public university in the process of recruitment of students using the test. New admissions to the test in two ways comprising, UM-PTKIN lines and Self Exam. For UM-PTKIN test, test instrument made by the central committee in the Ministry of Religion. Because this test will be used to measure prospective students throughout Indonesia, of course, has been through validation test before using. In this case at least there is no guarantee of the Ministry that the test has a good quality. Although in fact there will probably not have a good quality test. As well as to test self test, should be validated before use.

This is to ensure that tests are of good quality. In fact, the self-exam test IAIN Syaikh Abdurrahman Siddik Bangka Belitung had never been analyzed both qualitatively and quantitatively. In addition, scores of new admissions results through independent pathways have also not yet been processed for the sake of improving the quality of student input and output. Based on this analysis new admissions test instrument independent pathways IAIN Syaikh Abdurrahman Siddik Bangka Belitung important to do in order to ensure the validity of the measurement results. This study will test the quality of new student recruitment instruments used by IAIN Syaikh Abdurrahman Siddik Bangka Belitung in the academic year 2018/2019 is based on classical test theory approach.

METHOD

The research is quantitative approach included the category of research Ex Post Facto, where the researcher does not manipulate the variables or characteristics of the sample due to the existence of these variables has occurred (Simon & Goes, 2013) The study was conducted in response Ability Test Academic new students IAIN Syaikh Abdurrahman Siddik Bangka Belitung academic year 2018/2019, amounting to 425 sheets. Academic Ability Test with a multiple choice item number as many as 100 questions. Data collected through technical documentation of manuscript student test questions and answers.

The data were analyzed quantitatively by classical test theory test the parameters of the form, an index level of difficulty, distinguishing, distractor effectiveness, and reliability through application ANATES Version 4.0.2. The criteria that are used to determine the level of difficulty refers to the balance between the items easily, simply, and difficult with a ratio of 3: 4: 3. Distinguishing item item refers to > 0.3 is accepted, the revised 0.29 - 0.10, <0.10 was rejected. For the effectiveness of the distractors at least chosen by 5% of the test participants. While reliability coefficient minimum value of 0.70.

RESULT

Academic Ability Test new admissions IAIN Syaikh Abdurrahman Siddik Bangka Belitung totaled 100 questions comprising the field of Public Knowledge, Basic Mathematics, General Intelligence, English, and Arabic. Each of these fields amounted to 15 items except Arabic numbering 25 items. Quantitative analysis is performed using an application ANATES Version 4.0.2.

The difficulty level of the index

Index about the overall difficulty level has not demonstrated a balance about the comparison easy, medium and difficult. On average about distributed in the medium category with a percentage of 25% easy matter, about 51% moderate, and 24% about difficult.

Distinguishing Index

The results of the analysis of the index overall distinguishing items in all fields about 37% showed no matter who is able to distinguish the ability of test takers and 18% are not able to distinguish the ability of the test taker. A total of 18 questions on general knowledge about the field, General Intelligence, English, Indonesian, and Arabic is not able to distinguish the ability of the test taker. Since there are 8 questions were answered correctly by a lower group and answered one of the above groups. So that such questions better not be used to measure the ability of prospective students.

The Effectiveness of the Distractor

Based on the analysis, the percentage of distractor effectiveness has largely been functioning as a swindler. From there distractor 300 239 or 79.6% were working and there are 61 or 20.3% do not work. However, the data is only a general description of the functioning of the distractor. If visited by functioning distractors field Basic Math kindest matter where, amounting to 91.1% distractor function and only 6.7% are not functioning distractors. As for the distractor most do not work there on the field a matter of common knowledge, where there is no functioning distractors 35.6%.

Validity and Reliability

Analysis of the validity of items consisting of 100 questions obtained 54% of which have validity coefficient values > 1.59, as much as 46% which has a value of coefficient of <1:59. The figure shows that almost half of the number of questions that do not meet the validity coefficient. As for the coefficient of reliability tests overall foreign workers obtained coefficient value of 0.81. The purpose of the analysis was to determine the item any items that have the characteristics of a good question based approach CTT as listed in the table below.

Table. 1. Recapitulation

Subject	**Parameter of Analysis**	**Recomendation**		
		Accepted	**Revision**	**Rejected**
Common Knowledge	The Difficulty Level	6,8,10	1,2,3,4,5,7 9,12,13,14	11,15
	Item Distinguishing	1,10,12	3,5,6,8,9,	2,4,7,11,13, 14,15
	Distractor	1,4,5,8,10,12,13,14		2,3,6,7,9,11,15
Basic Mathematic	The Difficulty Level	16,17,19,21,22,25 26,27,28	18,20,23,24 29,30	-
	Item Distinguishing	16,17,19,20, 21,22,24,25, 26,27,28,29	18,23,30	-
	Distractor	16,17,19,21,22,25 26,27,28	18,20,23,24 29,30	-
General Intelegence	The Difficulty Level	32,33,34,35,36,40 43,45	31,37,38,39 41,42,44,	-
	Item Distinguishing	34,39	33,35,36,37, 38,42,43,44, 45	31,32,40,41
	Distractor	33,36,39,40, 43,44, 45		31,32,34,35, 37,38,41,42
English	The Difficulty Level	49,50,53,56,57,57 59,60	46,48,51,54 55	47,52
	Item Distinguishing	57,58,59	46,47,48,51, 52,53,54,55, 60	49,50,56
	Distractor	49,50,53,54,55,56, 57,58,59,60		47,48,51,52
Indonesian	The Difficulty Level	61,62,63, 65,67,69 70,71	64,66,68,72 73,74,75	-
	Item Distinguishing	61,62,63,64, 69,70,	65,67,68,71, 72,73,74,75	66

Subject	Parameter of Analysis	Recomendation		
		Accepted	Revision	Rejected
	Distractor	63,64,65,66, 67,68,69,70, 72,73,75		61,62,71,74
Arabic	The Difficulty Level	78,79,80,81, 82,83,84,85,86, 89,91,92,93,94, 95,96,97,98,100	76,77,87,88 90,99	-
	Item Distinguishing	77,79,80,82, 85,86, 90,95,97,98,100	76,81,83,84, 87,88,91,92, 93,94,96	78,89,99
	Distractor	78,80,81,82,84, 85,86,87,88,89, 90,91,92,93,94, 95,96,97,98,99,100		76,77,79,83,

CONCLUSION

The New admissions test should ideally not only be used as instruments to determine whether prospective graduate and students. However, it can provide much information about the ability of prospective students. Based on classical test theory approach, parameter test that can provide information about the ability of the test taker can be determined by testing the level of difficulty index, the index of distinguishing, distractor effectiveness, validity, and reliability.

Academic Ability Test new admissions IAIN Syaikh Abdurrahman Siddik Bangka Belitung, academic year 2018/2019 amounted to 100 questions comprising the field of General Knowledge, Basic Mathematics, General Intelligence, English, and Arabic. Related to the level of difficulty of questions, Sudjana in Suryani, (2017) suggest that, level of difficulty item should have a balance between the matter of easy, medium and difficult with a ratio of 3: 4: 3. When referring to this provision, the comparison about the easy, medium, and hard on the problem of foreign workers matter consists of 30% easy, 40% about the average and 30% about difficult.

The analysis showed that only about a subject of Indonesian approaching the ideal ratio spread about the difficulty level. Items were distributed in an easy category as much as 26.7%, while 46.7% and 26.7% difficult. As for the item subject with high inequality of distribution contained in the item of General Knowledge field with a ratio of 60% easy matter, 13.3% moderate, and 26.7% is difficult. General knowledge about the field too much spread about the matter category easily.

In contrast to Sudjana, according to Thomas and Dawson (1972), quoted by Kartowagiran, (2012) explained that the question of who has the level of difficulty of 0.25 - 0.75 already includes a good question. In addition, as disclosed Kadir, (2015) chose a good test items based on the level of difficulty based on the purpose of the test itself. If the test is only used for the purposes of semester exams, then the matter with the level of difficulty was gaining more serving. For diagnostic purposes, it is used about the level of difficulty is low. Whereas for the purposes of selection then been a matter of relatively difficult.

Items are either based on the level of difficulty as expressed Sudjana substantially less suited its purpose as an instrument of foreign workers test new student selection. Where the distribution of matter at the level of difficulty was more than a matter of easy and difficult. Opinion of Thomas and Dawson (1972) also basically still in line with the opinion that more Sudjana choose easy matter and as a matter of good being based on level of difficulty. From some of these opinions, in terms of determining which items are eligible to be used in tests opinions expressed Kadir more appropriate because the goal is for the purpose of selection of candidates for new student. Composition distribution more difficult problem will provide certainty in the ability of students based on each field questions. If we refer to Kadir opinion, test instrument based on the analysis needs to be revised because it is still dominated by problems with category with a percentage of 51%.

In addition to the level of difficulty, test questions also ideally be able to distinguish the ability of the test taker. Based on the ability of CTT approach is known as the index of distinguishing. Distinguishing index is shown with values ranging from -1 to + 1. The meaning of a positive price is that the material master answered correctly, and that do not master answered incorrectly. Otherwise if the score of the index is negative Mardapi, (2012). Further according to Saifuddin Azwar, (2010) in practice, the parameters with negative values requires that the question is not used. As revealed Crocker and Algina (1986) in Azwar, (1993), item having distinguishing good if it has a coefficient greater than r_{bis} = 0.200. While Nitko (1983) as quoted Surapranata, (2009) states that, the value of coefficient of distinguishing at least the lowest at 0:30.

Results of the analysis showed that there were distinguishing features about 37% of participants were able to distinguish the ability of the test, 45% poor and 18% are not able to distinguish the ability of the test taker. Problem with power coefficient negative discrimination are still found. Problem having distinguishing negative means the question is answered correctly by many groups with low capacity and many answered incorrectly in the group with high ability. At about Academic Ability Test found problems with negative differentiated power contained in the field of General Knowledge about the number 14 with a coefficient of -0.04.

The parameters contained in the CTT in addition to an index level of difficulty and distinguishing features is the effectiveness of distractors. A dictractor can function well if at least 5% of participants selected by the test. If humbug been evenly, then including posing very good Surapranata, (2009) distractors in this case is one of four possible answers in the answer choices Academic Ability Test. The analysis showed as much as 79.7% or 239 distractors functioned well and 20.3% or 61 distractor does not work or have <5% of the 300 participants who answered the test. Figures distractor malfunction is quite high, amounting to 20.3%. Effectiveness distractor is not only seen the percentage of participants who memili distractor tests only. Because it can be, a distractor may have more than 5% of test participants are even more of an answer key. This happens because the person taking the test is still regarded as a key answer distractor.

Based on the analysis contained at least 6.6% or 3 distractor on the field a matter of common knowledge that is considered as a key response by the test taker. Furthermore, in the field of Mathematics grounds contained about 2.2% or 1 distractors. In the field of General Intelligence contained about 20% or 20 distractors, English field contained about 17.8% or 8 distractors, Indonesian field contained about 8.8% or 4 distractors, and the field of Arabic contained about 6.6% or 5 distractor. General intelligence about the field has more distractors are not functioning that is as much as 20%. It can also be influenced by distinguishing a matter which is also low. There are only about 13.3%, having distinguishing features. Field about English is also only 20% about having distinguishing features, as well as the field about the General Intelligence. Are changes in the effectiveness of this distractor strongly influenced by distinguishing about the need to do further testing.

In addition, to determine whether or not the item can be known through the analysis of the validity. Validity in this case refers to the validity of the item itself and not the validity of the test instrument. According Sudijono in Surapranata, (2009) Validity of the items is a degree of correspondence between an item with a score of device items (item total) So it can be understood that if each item has a correlation with the item (item total) means any items these items measure the dimensions the same one. Based on the analysis of validity to the whole items with a number of 100 items, a significant 54% and 46% or valid or invalid insignificant. Analysis shows there is a 46% validity matter or items that do not measure the same dimension. Or in another understanding that, there are 46 items that do not matter to measure the dimensions of each field problems. There are at least two areas of questions that have significant percentage of low validity, namely General knowledge about the field by 33% and the General Intelligence field by 33%. The second field is the question if the note does have difficulty index, distinguishing, and distractor unfavorable. There is a possibility of a third validity of the items affected by these parameters.

Parameter to estimate the overall quality of the test instrument can be known through the analysis of reliability. Similarly, the validity, reliability is also seen through the coefficient value which starts from -1 to +1. According to Linn in Mansour et al (2009: 24), as quoted I Iskandar & Rizal, (2018) argues that minimum limit of reliability coefficient value of at least 0.70. According Suryabrata as quoted Solichin, (2017), the reliability of the test instrument refers to the extent to which the degree of consistency score two devices are expressed in terms of the correlation coefficient. The smaller the variability score matter or items, the more shows the value of the consistency of a test instrument. Consistency or kejegan very important in the measurement. Instrument tests that have high consistency value necessarily produce reliable measuring or trustworthy, and vice versa.

Meanwhile, according to Azwar, (2010) The reliability of the test is the proportion of variability of test scores caused by the actual difference between the test taker. While the tests unreliable is the proportion of variability of test scores caused by error measurement. More Azwar, (2010) explains, the smaller the coefficient of reliability or farther from 1, the greater the variation of errors measuremen that occur. Results of reliability analysis generates coefficient value of 0.81. When referring to the opinion Linn above, the value of the coefficient of reliability Academic Ability Test new students of IAIN Syaikh Abdurrahman Siddik Bangka Belitung more than 0.70 so as to qualify the test reliable.

In addition, by knowing the value of reliability coefficient can also be known how large the error of measurement that occur as described by Azwar. In this case, the results of the analysis of reliability of 0.81, which means that 81% of the variance scores seemed a variant of pure score. Thus, it is understood that, by 80% difference in scores obtained with other test takers is the difference in their pure score or not influenced by other factors as sources of error in the measurement. In reliability, the new applicant Academic Ability Test instrument IAIN Syaikh Abdurrahman Siddik Bangka Belitung quite good.

REFERENCES

Anthony J. Nitko. (2001). *Educational Assessment of Student* (3rd ed.). New Jersey: Prentice Hall Inc.

Arikunto, S. (2009). *Dasar-Dasar Evaluasi Pendidikan* (10th ed.). Jakarta: Bumi Aksara.

Azwar, S. (2008). The Quality Of The Tes Potensi Akademik (TPA) 07A. *Jurnal Penelitian Dan Evaluasi Pendidikan, Nomor, 2*.

Azwar, Saifuddin. (1993). Berkenalan dengan teori respons aitem. *Buletin Psikologi*, *1*(1), 9–16.

Azwar, Saifuddin. (2009). *Tes Prestasi*. Yogyakarta: Pustaka Pelajar.

Daryanto. (2012). *Evaluasi Pendidikan*. Jakarta: Rineka Cipta.

Djaali dan Pudji Muljono. (2008). *Pengukuran Dalam Bidang Pendidikan*. Jakarta: Grasindo.

Golafshani, N. (2003). Understanding reliability and validity in qualitative research. *The Qualitative Report*, *8*(4), 597–606.

Hamzah, B. U. (2010). *Pengembangan Instrumen Untuk Penelitian*. Jakarta: Delima Press.

Iskandar, A., & Rizal, M. (2018). Analisis kualitas soal di perguruan tinggi berbasis aplikasi TAP. *Jurnal Penelitian Dan Evaluasi Pendidikan*, *22*(1), 12–23.

Kadir, A. (2015). Menyusun dan Menganalisis Tes Hasil Belajar. *Al-Ta'dib*, *8*(2), 70–81.

Kartowagiran, B. (2012). Penulisan butir soal. *Yogyakarta: Universitas Negeri Yogyakarta*.

Mardapi, D. (2012). *Pengukuran Penilaian dan Evaluasi Pendidikan*. Yogyakarta: Yuha Medika.

Saifuddin Azwar. (2010). *Sikap Manusia: Teori dan Pengukurannya* (Cetakan XI). Yogyakarta: Pustaka Pelajar.

Simon, M. K., & Goes, J. (2013). Ex post facto research. *Retrieved From*.

Solichin, M. (2017). Analisis Daya Beda Soal, Taraf Kesukaran, Validitas Butir Tes, Interpretasi Hasil Tes dan Validitas Ramalan dalam Evaluasi Pendidikan. *Dirāsāt: Jurnal Manajemen Dan Pendidikan Islam*, *2*(2), 192–213.

Sukardi. (2010). *Evaluasi Pendidikan: Prinsip dan Operasionalnya*. Jakarta: Bumi Aksara.

Surapranata, S. (2009). *Analisis Validitas, Reliabilitas, dan Interpretasi Hasil Tes: Implementasi Kurikulum 2004* (4th ed.). Bandung: Remaja Rosdakarya.

Suryani, Y. E. (2017). Pemetaan kualitas empirik soal ujian akhir semester pada mata pelajaran Bahasa Indonesia SMA di Kabupaten Klaten. *Jurnal Penelitian Dan Evaluasi Pendidikan*, *21*(2), 142–152.

Widoyoko, E. P. (2009). *Evaluasi program pembelajaran*. Yogyakarta: Pustaka Pelajar.