

PAI SD Test Quality Analysis: Validity and Reliability Review

Pipit Pitriani, Enung Nugraha, Fitri Hilmiyanti

UIN Sultan Maulana Hasanuddin Banten, Indonesia

pipitpitrianisyafitri@gmail.com

ABSTRACT

This study aims to analyze the quality standards of test instruments, namely the difficulty level index, discrimination index, validity and reliability of the PAI UAS test made by PAI teachers in Tigaraksa District. This study is a descriptive study with a qualitative and quantitative approach. Data was taken using documentation data, namely PAI UAS question documents student test results. Qualitative data was analyzed logically using Bloom's Taxonomy theory. Quantitative data was analyzed to calculate the difficulty level index, the discrimination index, the validity test used the point biserial correlation technique and the validity test used the Kuder Richardson 20 (KR-20) correlation technique. The study results showed that 47.5% of the questions were at cognitive level C1. In general, the PAI test items developed by PAI teachers in Tigaraksa District, the quality of the Islamic Religious Education (PAI) questions produced is still far from standard. This is based on the fact that there are still many invalid items and the reliability index of the test instrument is categorized as unreliable.

Keywords: PAI Test Quality, Validity and Reliability

ABSTRAK

Kajian ini bertujuan untuk menganalisis standar kualitas instrumen tes yaitu indeks tingkat kesukaran, indeks daya pembeda, validitas dan reliabilitas tes Ujian Akhir Sekolah Pendidikan Agama Islam (PAI) yang dibuat oleh guru PAI di Kecamatan Tigaraksa. Kajian ini merupakan kajian deskriptif dengan pendekatan kualitatif dan kuantitatif. Data diambil menggunakan data dokumentasi yaitu dokumen soal UAS PAI di Kecamatan Tigaraksa Kabupaten Tangerang dan hasil tes peserta didik. Data kualitatif dianalisis secara logika menggunakan teori Taksonomi Bloom. Data kuantitatif dianalisis untuk menghitung indeks tingkat kesukaran, indeks daya pembeda, uji validitas digunakan teknik korelasi *point biserial* dan uji validitas menggunakan teknik korelasi Kuder Richardson 20 (KR-20). Hasil kajian menunjukkan 47.5% soal berada pada level kognitif C1. Secara umum, Butir-butir tes PAI yang dikembangkan oleh guru-guru PAI di Kecamatan Tigaraksa kualitas soal Pendidikan Agama Islam (PAI) yang dibuat masih jauh dari standar. Hal ini didasarkan pada masih banyak item yang tidak valid dan indeks reliabilitas instrumen tes dikategorikan tidak reliabel.

Kata Kunci: Kualitas Tes PAI, Validitas dan Reliabilitas

INTRODUCTION

Reliability and validity are needed to see the quality of an instrument. The quality of the instrument can determine the quality of the results from measurements using the instrument. Reliability refers to the stability of findings, while validity represents the truth of the findings (Nunnally & Bernstein, 1994)

Validity and reliability increase transparency, and reduce the opportunity to introduce tester (test creator) or test user bias in the measurement (Bajpai & Bajpai, 2014). Evidence of validity and reliability is a prerequisite for ensuring the integrity and quality of a measurement instrument (Nunnally & Bernstein, 1994).

For secondary data, detailed assessments of reliability and validity involve assessing the methods used to collect the data. This provides a good link to interpreting scores from psychometric instruments, for example, symptom scales, questionnaires, educational tests, and observer ratings) used in clinical practice, research, education, and administration (Cook & Beckman, 2006).

This study aims to analyze the quality standards of test instruments, namely the difficulty level index, distinguishing power index, validity and reliability of the PAI UAS test made by PAI teachers in Tigaraksa sub-district, Tangerang district. The study is expected to provide PAI teachers with an idea of the quality of the questions they create.

LITERATURE REVIEW

Cognitive Domain

There are six levels of cognitive learning according to the revised version of Bloom's Taxonomy. Each level is conceptually different. The six levels are remembering, understanding, applying, analyzing, evaluating, and creating (Anderson et al., 2001). These levels can help in developing learning outcomes because certain verbs are very appropriate at each level and not appropriate at other levels.

Item Difficulty Index (IDI)

The difficulty level of a question item is defined as the chance of answering the question correctly based on certain abilities, which is usually shown in the form of an index. The Aiken difficulty index ranges between 0.00 and 1.00.

0,00  1,00

The greater the difficulty index, the easier the question. Based on the difficulty level index, it should be more appropriate to call the question item's level of ease. Because it has become an agreement among experts, to this day we still use the term item difficulty level. The meaning of the level of difficulty (IDI) = 1, meaning that students answered the question correctly, IDI = 0, meaning that no students answered the question correctly.

The level of test difficulty can be calculated based on the number of participants who answered correctly in the upper and lower groups which is formulated as follows:

$$IDI = \frac{Gu + Gl}{N}$$

IDI = Item Difficulty Index

Gu = Number of correct answers in the upper group (27%)

Gl = Number of correct answers in the lower group (27%)

N = Group size (number of participants in the upper and lower groups)

Item Difficulty Index Criteria (Nana Sudjana):

0.00 - 0.30 = Difficult

0.31 - 0.70 = Medium

0.71 - 1.00 = Easy

Discrimination Index

The differentiating index of test items refers to the ability of the items to differentiate between test takers who have mastered the material and test takers who have not/have not mastered the material in question. Discriminating power is expressed in an index. The discriminating power index ranges from -1.00 to +1.00. The higher the discriminating power index of a question means that the question in question is more capable of differentiating students/test takers who have understood the material from students/test takers who have not understood the material. The higher the differentiating power of a question item, the stronger/better the question item. If the discriminating power index is negative ($ID < 0$), it means that more of the lower group (students/test takers who do not understand the material) answered the question correctly compared to the upper group (students/test takers who understand the material).

In item selection, every item that has an index greater than 0.50 can be immediately considered as an item with good discriminatory index, items that have an index of less than 0.20 can be immediately discarded, while other items can be studied further for revision.

$$DI = \frac{Gu - Gl}{\frac{1}{2}N}$$

Classification (Crocker & Algina, 1986)

0.40 – 1.00 Items accepted/good

0.30 – 0.39 Items are accepted but need improvement

0.20 – 0.29 Items corrected

0.19 – 0.00 Items not used/discarded

Validity

Validity comes from the word validity (saheeh in Arabic) which means the extent to which a measuring instrument is accurate and precise in carrying out its measuring function. A test tool can be said to have high validity if the test tool carries out its measuring function, or provides measuring results that are in accordance with the purpose of the measurement. Meanwhile, tests that have low validity will produce data that is not relevant to the measurement objectives. The notion of validity is also very closely related to the purpose of measurement. Therefore, there is no general validity for all measurement purposes. A measuring instrument is usually only a valid measure for one specific purpose. Thus, the assumption of validity as stated in "this measuring instrument is valid" is

incomplete. The valid statement must be followed by information that points to the objective (i.e. valid for measuring what), as well as valid for which subject group.

Validity is divided into three, namely content validity, construct validity and criterion-related validity (Linn & Grondlund, 2000).

a. Construct Validity

Construct validity is validity that concerns the theoretical building of the variables to be measured. A test is said to have construct validity if the items compiled in the test measure every aspect of thinking of a variable that will be measured through the test. To test the validity of the construct, opinions from experts (Judgment Expert) can be used. Experts were asked for their opinions about the test kit.

b. Content Validity

Content validity is also called curricular validity. Therefore, validity is closely related to the material that will be measured in the test. Of course, the material in question is the material contained in the curriculum. Content validity testing can be done by comparing the content of the test tool with the content or design that has been determined. Content validity reflects the extent to which the items in a test reflect the material presented in the curriculum. A test is said to have content validity if the test items are representative of the content of the material in the curriculum.

c. Criterion Validity

As the name suggests, this validity is based on certain criteria. Thus, evidence of validity is shown by a correlational relationship between the score on the test in question and the score of a criterion. This validity test is empirical, meaning that the test can only be carried out after obtaining data in the field. If, based on the results of the analysis carried out on field observation data, it is proven that the learning outcomes test can measure the learning outcomes that should be expressed accurately, then this means that the test tool has empirical validity. For the purposes of testing this type of validity, it can be done in two ways, namely in terms of its ability to make predictions (predictive validity) and its comparative accuracy (concurrent validity).

The main difference between predictive validity and comparative validity is the availability of comparisons (criteria). In forecast validity, the criteria are obtained in the future after the test whose validity will be measured is carried out. Meanwhile, in comparative validity, the criteria already exist or can be obtained at the same time as the time to obtain data about the test whose validity will be measured without having to wait for the future. Test the validity of multiple choice questions using point biserial correlation, namely the correlation between interval data and dichotomous data.

$$r_{pbis} = \frac{\bar{X}_b - \bar{X}_s}{SD_t} \sqrt{pq}$$

Notes:

$\bar{X}_b$ = the mean score of students/test takers who answered correctly

$\bar{X}_s$ = the mean score of students/test takers who answered incorrectly

SD_t = standard deviation of the total score

p = the proportion of correct answers to all student/test taker answers

q = 1 - p

The item criteria are valid if the rpbis value exceeds 0.3 and this means that the item can be trusted and used. Meanwhile, if the rpbis value is less than 0.3 then the item is invalid.

Reliability

Apart from validity, a good measuring instrument must also be reliable. Therefore, a good measuring instrument is a measuring instrument that is valid and reliable. In theoretical studies, reliability is the extent to which measurements from a trial carried out still have the same results even though they are carried out repeatedly on subjects and under the same conditions. A measuring instrument is considered reliable if it gives consistent results for the same measurement and unreliable if repeated measurements give relatively unequal results. Instrument reliability testing to obtain reliable results can be done using various statistical methods.

A learning outcomes test is said to be steady if the current measurement results show similar results at different times for the same students. For example, today's fifth grade students were tested for Islamic Religious Education (PAI) abilities. The following week the student was tested again. The results of the two tests are relatively similar. However, it is still possible for there to be differences in results for certain matters due to coincidence, time lapse, or changes in students' views on the same question. If that happens, the weakness lies in the test, which does not have certain answers or doubts the student. In other words, the degree of reliability is still low.

The main purpose of calculating test score reliability is to determine the level of precision and consistency of test scores. The reliability index ranges between 0 - 1. The higher the reliability coefficient of a test (closer to 1), the higher the consistency/accuracy.

Tests that have consistently high reliability are accurate, reproducible, and generalized to other testing opportunities and test instruments. In detail, the factors that influence the reliability of test scores include: (1) the greater the number of questions, (2) the more steady the test, (3) the longer the test time, the more steady it is, (4) the narrower the difficulty range of the questions, (5) the greater the constancy, (6) questions that are interconnected will reduce the constancy, (7) the more objective the scoring, the greater the constancy, (8) the inaccuracy of scoring, (9) answering the questions by guessing, (10) the more homogeneous the material, the greater the consistency, (11) the examinee's experience, (12) misinterpretation

of the question items, (13) answering questions in a hurry, (14) the examinee's mental readiness, (15) the presence of disturbances in test implementation, (16) the distance between the first test and the second test, (17) cheating in taking the test, (18) the individual's position in learning, and (19) the physical condition of the examinee.

There are 3 ways you can determine the reliability of test scores:

a. Test-Retest Method

This method is applied to avoid compiling two test series. The technique is that the same test is given twice to the same respondent with a certain time interval. If the results of the first test are in line with the results of the second test, then the test is said to be reliable. Because this test is carried out on a test that is tried twice, it is often referred to as the single-test-double-trial-method. The weakness of this method is that if the test time lag is too short while the test questions reveal many aspects of knowledge, respondents tend to still remember the material being tested, so there is a possibility that the results of the second test will be better than the results of the first test. On the other hand, if the time between the first and second tests is too long, it is feared that many factors and situations and conditions have changed a lot and will affect the results of the second test.

b. Parallel Test Method (Equivalent)

This method requires two series of questions that have the same objective, question weight, level of difficulty, question structure, but the question items are different. In other words, the two tests used must be parallel (parallel, equivalent). The reliability coefficient is obtained by correlating the results of the first test with the results of the second test. Of course, this method will add to the hassle. This is the weakness of this method. The advantage of this method is that it can correct the weakness of the first method, namely avoiding the condition of "students still remembering the first test material". The memory and memorization aspects of taking the first test do not carry over when taking the second test.

c. Split Two Method (Split – Half)

This single test method is more practical than the previous two methods. This method only carries out one test on a group of subjects. Thus, there is no need to wait time or have to have data from similar tests to be able to determine its reliability. Reliability is measured by the results of measuring the first and second hemispheres from the same measuring instrument.

To determine the reliability coefficient of the multiple choice test, the Kuder Richardson 20 (KR-20) formula is used as follows.

$$KR - 20 = \frac{k}{k-1} \left(1 - \frac{\sum pq}{SD_t^2} \right)$$

Notes:

k = number of questions

SD_t^2 = total score variance

P = proportion of students who answered correctly

q = 1 - p

To determine the reliability of measuring instruments, the Kaplan criteria are used (Kaplan & Sacuzzo, 2005), is

KR-20 $\geq$ 0,70 = measuring instrument is reliable (less reliable)

KR-20 < 0,70 = measuring instruments are less reliable (reliable)

RESEARCH METHODS

This study is a descriptive study with a qualitative and quantitative approach. The data in the study used documentation data, namely data in the form of UAS questions created by PAI teachers and data on school exam test results for Islamic Religious Education Subjects at the KKG PAI Rayon Elementary School, Tigaraksa District, Tangerang Regency.

There were 31 student test result data analyzed in this study. Qualitative data was analyzed logically using Bloom's Taxonomy theory. Quantitative data was analyzed to calculate the level of difficulty index, the index of discrimination, the validity test used the point biserial correlation technique and the validity test used the Kuder Richardson 20 (KR-20) correlation technique.

RESULTS AND DISCUSSION

Bloom's Taxonomy Cognitive Level

Table 1 shows the item levels based on Bloom's taxonomy levels for Islamic Religious Education (PAI) subject questions created by the Tigaraksa PAI Cluster, Tigaraksa, Tangerang.

Table 1. Description of Bloom's Taxonomy Item Levels

Item	Level	Frequency	%
1,2,3,4,5,9,10,11,12,13,14,17,18,22,23,26,35,38,39	C1	19	47.5
6,8,15,16,19,20,21,24, 25,27,28,29,30,32,34,40	C2	16	40
7, 31,33,36,37	C3	5	12.5

Table 1 shows that the cognitive domain level of PAI questions made by teachers is dominated by level C1 (remembering) at 47.5% and C2 (understanding) at 40%. Levels C1 and C2 are included in the lower level category (L2). There are only 12.5% of questions categorized as C3 (Applying). Level C1 (remembering), namely retrieving relevant knowledge from long-term memory (Hidayat, 2012; Hidayat & Lawahid, 2020).The level at this level are recognizing and recalling. Meanwhile, level C2 (understanding), namely determining the meaning of messages in lessons including oral, written or graphic (Anderson et al., 2001).

Difficulty Level Index

PAI item difficulty level index to see whether the question is easy, medium or difficult (Sudijono, 2005). Table 2 shows the difficulty level index for PAI items.

Table 2. Description of PAI Item Difficulty Level Index

Interval	Difficulty Level	Item	F	%
0,00 – 0,30	Easy	2,4,6,7,8,9,10,11,12,13,14,15,16,17,18,19,20, 21,22,23,24,25,26,29,32,33,35,	27	67.5
0,31 – 0,61	Moderate	1,3,5,27,28,30,31,34,36,37,38,40	12	30
0,62 – 1,00	Difficult	39	1	2.5

Based on table 2, it shows that 67.5% of the questions were categorized as easy, 30% of the questions were categorized as medium and only 2.5% of the questions were categorized as difficult. Therefore, the questions created by PAI teacher Gugus Tigaraksa can generally be categorized as easy

Discriminating Index

The level of difficulty describes the ability of a question item to identify or differentiate the ability of a group of test takers who already understand or do not understand the PAI material being tested (Syarifuddin, 2003). Table 2 shows the difficulty level index for PAI items.

Table 2. Description of the Differentiating Index for PAI Items

Interval	Discrimination Index	Item	F	%
$\leq 0,00$	Very low			
$0,00 < D \leq 0,20$	Low	2,3,6,7,8,9,10,11,12,15,16,1 8,20,21,22,23,24,25,26,27,2 8,31,32,35,38,39,40	27	67,5
$0,20 < D \leq 0,40$	Moderate	1,4,5,13,14,17,19,29,33,34	10	25
$0,40 < D \leq 0,70$	High	30,36,37	3	7,5
$0,70 < D \leq 1,00$	Very High	-	0	0

Table 2 shows that 67.5% of the differentiating power index of PAI items is categorized as low, 25% is categorized as medium, and only 7.5% is categorized as high.

Validity

Validity refers to the extent to which a test measures what it is supposed to measure. Table 3 shows item descriptions based on calculations using the point biserial correlation formula.

Table 3. Description of Item Validity

Indeks r_{pbis}	level	Item	F	%
< 0,00	Invalid	8,9,10,11,12,15,18,21,23,24,32,40	36	90
< 0,30	Invalid	1,3,4,5,6,7,13,14,16,17,19,20,22,26,27,28,29,30,31,33,34,35,38,39		
$\geq 0,30$	Valid	2,25,36,37	4	10

The table shows that 90% of the items are categorized as invalid because the index or validity score of the items is below 0.30. This means that the items cannot measure what they are supposed to measure. There are only 10% of questions that are categorized as valid. Therefore, the questions need to be corrected.

Reliability

Based on calculations using the Kuder Richardson 20 (KR-20) formula, it shows that the reliability of the PAI test instrument created by the teacher is 0.601. Based on the Kaplan criteria, a KR-20 value of less than 0.70 is categorized as unreliable (Kaplan & Saccuzzo, 2005).

CONCLUSION

In general, the quality of the Islamic Religious Education (PAI) test produced is still far from standard. This is based on the fact that there are still many invalid items and the reliability index of the test instrument is categorized as unreliable. Therefore, it is necessary to improve the question items if they are to be used in the next test. Apart from that, if the questions created are to be used more widely, they should be tested for face validity and construct validity by involving relevant experts in the PAI field.

REFERENCES

- Anderson, L. W., Krathwohl, D. R., Airasian, P. W., Cruikshank, K. A., Mayer, R. E., & Pintrich, P. R. (2001). *A Taxonomy for Learning, Teaching, and Assissing: A Revision of Bloom's Taxonomy of Educational Objectives*. Longman.
- Bajpai, S. R., & Bajpai, R. C. (2014). Goodness of Measurement: Reliability and Validity. *International Journal of Medical Science and Public Health*, 3(2), 112–115.
- Cook, D. A., & Beckman, T. J. (2006). Current Concepts in Validity and Reliability for Psychometric Instruments: Theory and Application. *The American Journal of Medicine*, 119(2), 166.e7-166.e16.
- Crocker, L., & Algina, J. (1986). *Introduction to Classical and Modern Test Theory*. Harcourt.

- Hidayat, W. (2012). *Evaluasi Pembelajaran Pendidikan Agama Islam*. Gre Publishing.
- Hidayat, W., & Lawahid, N. A. (2020). *Metode Fuzzy Delphi Untuk Penelitian Sosial*. Alfabeta.
- Kaplan, R. ., & Saccuzzo, D. . (2005). *Psychological Testing: Principles, Application and Issue* (6th ed.). Wadsworth.
- Linn, R. ., & Grondlund, N. . (2000). *Measurement and Assessment In Teaching* (8th ed.). Merrill an imprint of Prentice Hall.
- Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric Theory* (3rd ed.). Mcgraw-Hill.
- Sudijono, A. (2005). *Pengantar Evaluasi Pendidikan*. Rajawali Press.
- Syarifuddin, A. (2003). *Tes Prestasi: Fungsi dan Pengembangan Pengukuran Prestasi Belajar* (4th ed.). Pustaka Pelajar.

APPENDIX

A. Question Items

No	Item	Level
1	Tangannya bercahaya merupakan mukjizat dari nabi...	C1
2	Mukjizat nabi Yusuf a.s adalah, <i>kecuali....</i>	C1
3	Nabi yang dilahirkan tanpa ayah dan bisa berbicara saat bayi adalah nabi...	C1
4	Nabi yang mempunyai anak yang durhaka (Kan'an) dan memilih untuk berlindung di atas bukit lantaran banjir yang semakin meluap dari pada ikut Bersama ayahnya naik ke bahtera. Nabi yang dimaksud adalah nabi....	C1
5	Mukjizat nabi Muhammad saw yang terbesar yang berlaku hingga akhir zaman adalah....	C1
6	Rasul yang mempunyai ketabahan dan kesabaran yang luar biasa disebut dengan....	C2
7	Semua manusia akan merugi kecuali orang-orang beriman dan beramal soleh. Diantara amal soleh terhadap diri sendiri adalah....	C3
8	Suatu amal soleh akan syah jika memenuhi syarat sebagai berikut, <i>kecuali...</i>	C2
9	وَالْعَصْرِ Artinya....	C1
10	إِنَّ الْإِنْسَانَ لِفِي خُسْرٍ Artinya	C1
11	Kita sebagai kaum Muslim diperintahkan untuk senantiasa bersikap baik sangka kepada semua orang terlebih lagi baik sangka kepada atas apa yang menimpa kita. Baik sangka disebut juga dengan...	C1
12	Baik sangka merupakan perilaku terpuji yang harus dimiliki seorang Muslim. Lawan dari baik sangka adalah...	C1
13	Puasa yang dilaksanakan di luar bulan Ramadhan sebagai ganti puasa Ramadhan yang telah ditinggalkan disebut puasa...	C1
14	Hari yang dilarang untuk melaksanakan puasa sunah adalah....	C1
15	Diantara para nabi dan rasul yang wajib diketahui ada 25 salah satunya ada nabi Sulaiman A.S. mukjizat nabi Sulaiman a.s adalah, <i>kecuali....</i>	C2
16	Nabi yang memiliki wajah yang tampan dan bisa mentafsirkan mimpi adalah nabi....	C2
17	Puasa yang apabila dilakukan mendapat pahala, apabila ditinggalkan mendapat dosa merupakan dari pengertian dari puasa...	C1
18	Puasa merupakan rukun Islam yang ke 4. Puasa menurut istilah artinya....	C1

Mimbar Kampus: Jurnal Pendidikan dan Agama Islam

Vol 23 No 2 (2024) 651 - 663 P-ISSN 1411-7673 E-ISSN 2776-5571

DOI: 10.17467/mk.v23i2.1372

No	Item	Level
19	Syarat batal puasa diantaranya, <i>kecuali....</i>	C2
20	Perumpamaan orang yang gibah atau menceritakan keburukan orang lain diibaratkan seperti....	C2
21	Perhatikan potongan ayat berikut! وَأَعْبُدُوا اللَّهَ وَلَا تُشْرِكُوا بِهِ شَيْئًا ۚ وَبِالْوَالِدَيْنِ إِحْسَنًا وَبِذِي الْقُرْبَىٰ وَالْيَتَامَىٰ وَالْمَسْكِينِ وَالْجَارِ	C2
22	Nabi Ibrahim a.s merupakan salah satu yang termasuk kedalam Rasul Ulul Azmi lantaran mempunyai ketabahan dan kesabaran yang luar biasa. Gelar nabi Ibrahim adalah....	C1
23	Nabi yang mempunyai pasukan Jin, hewan, dan manusia adalah mukjizat nabi...	C1
24	Mukjizat yang bisa mengetahui nama-nama benda yang ada di langit dan di bumi adalah mukjizat nabi....	C2
25	Sebagai umat Islam rasa cintanya kepada rasul dapat dilakukan dengan cara....	C2
26	Mengganti puasa di hari-hari setelah hari raya idul fitri karena meninggalkannya pada waktu puasa ramadhan disebut dengan puasa....	C1
27	Sebagai seorang Muslim, kita diperintahkan untuk memakan makanan yang halalan tayyiban, maksudnya adalah....	C2
28	Di dalam Surat Al Asr menjelaskan bahwa semua manusia itu akan merugi <i>kecuali....</i>	C2
29	Perhatikan ayat berikut! وَتَوَاصَوْا بِالْحَقِّ Arti dari ayat berikut adalah...	C2
30	Jika mad tabi'i bertemu dengan alif dalam satu kalimat. Maka hukum tersebut dinamakan dengan....	C2
31	Perhatikan potongan ayat di bawah! وَأَعْبُدُوا اللَّهَ وَلَا تُشْرِكُوا بِهِ شَيْئًا Kalimat yang bergaris bawah terdapat hukum bacaan....	C3
32	Dalam Q.S An Nisa ayat 36 Allah swt memerintahkan kita untuk berbuat baik kepada kedua orang tua, karib kerabat, anak yatim dan orang miskin. Kalimat yang berarti anak yatim adalah....	C2
33	Perhatikan ayat berikut ! فَكُلُوا مِمَّا رَزَقَكُمُ اللَّهُ حَلَلًا طَيِّبًا	C3

No	Item	Level
34	Dalam ilmu tajwid kita mengenal ada Mad far'I. di antara mad far'I ada hukum yaitu mad silah qosiroh bertemu dengan alif maka bacaan tersebut dinamakan dengan....	C2
35	Hikmah mengonsumsi makanan dan minuman halal adalah....	C1
36	Contoh bacaan mad iwad adalah....	C3
37	Contoh hukum bacaan Lam Tafkhim (tebal) adalah.....	C3
38	Lam Tarqiq dibaca....	C1
39	Hukum mim mati bertemu dengan ba disebut dengan....	C1
40	Diantara hukum mim mati terdapat hukum bacaan Idgham mutamasilain atau disebut dengan idgham Mimi. Idgham Mimi artinya...	C2

B. Examples of Calculations in Quality Analysis of PAI Questions

1. Tingkat Kesukaran

$$TK = \frac{BA + BB}{N}$$

$$TK = \frac{7 + 4}{18} = 0,61 \text{ (Tingkat Kesukaran Kategori Sedang)}$$

2. Daya Pembeda

$$DP = \frac{BA - BB}{N}$$

$$DP \text{ No.1} = \frac{7 - 4}{\frac{1}{2}(18)} = \frac{3}{9} = 0,33 \text{ (Daya Pembeda Sedang)}$$

3. Validitas

$$r_{pbis} = \frac{\bar{X}_b - \bar{X}_s}{SD_t} \sqrt{pq}$$

$$r_{pbis} \text{ no 1} = \frac{31 - 28,1}{6,93} \sqrt{0,53 \times 0,47}$$

$$r_{pbis} \text{ no 1} = \frac{2,9}{6,93} \sqrt{0,25}$$

$$r_{pbis} \text{ no 1} = \frac{2,9}{6,93} \sqrt{0,25}$$

$$r_{pbis} \text{ no 1} = 0,42 \times 0,5 = 0,21 \text{ (Tidak Valid)}$$

4. Reliabilitas

$$KR - 20 = \frac{k}{k-1} \left(1 - \frac{\sum pq}{SD_t^2} \right)$$

$$KR - 20 = \frac{40}{40-1} \left(1 - \frac{11,28}{47,1} \right)$$

Mimbar Kampus: Jurnal Pendidikan dan Agama Islam

Vol 23 No 2 (2024) 651 - 663 P-ISSN 1411-7673 E-ISSN 2776-5571

DOI: 10.17467/mk.v23i2.1372

KR – 20 = 0,601 (TIDAK RELIABEL)