



# Analisis Kualitas Tes dan Butir Soal UTS Pilihan Ganda Bahasa Arab Kelas III di SDIT Mutiara Global Kubang Raya Pekanbaru

Kiki Dewi Octaviani<sup>1</sup>, Hikmah<sup>2</sup>, Masrun<sup>3</sup>  
Universitas Islam Negeri Sultan Syarif Kasim Riau<sup>1,2,3</sup>  
[kikidewi1510@gmail.com](mailto:kikidewi1510@gmail.com)

Received: 18-09-2024

Accepted: 28-11-2024

Published: 30-12-2024

## ABSTRACT

*The quality of the test greatly affects the accuracy of the test itself. Therefore, analysis of test quality is a step that must be taken to determine the degree of quality of a test, both the test as a whole and the items that are part of the test. In the assessment of learning outcomes, the test is expected to describe a sample of behavior and produce an objective and accurate value. The test used by the teacher must have good quality in terms of various aspects. Tests should be prepared in accordance with the principles and procedures for preparing tests, after use it is necessary to know whether the test is of good or poor quality. The type of this research is experimental research with a quantitative approach. To obtain data, of course, a tool or instrument is needed. The instrument used by researchers is a written test in the form of an objective test (multiple choice test) containing 10 questions in which there are questions to be answered by participants. The population in this study were all 3rd grade students of SDIT Mutiara Global Kubang Raya with a total of 52 people. From the results of the analysis using the ITEMAN (SPSS) Version 2.4 program, the items that were grouped well based on the level of difficulty were 2 items. From the results of the analysis using the SPSS version 2.4 program, the items categorized as very good in terms of differentiating power are 7 items, while those categorized as good enough do not need improvement are 2 items, those categorized as less good and need improvement are 0 items, and those categorized as bad are 0 items. Based on the results of the analysis using the ITEMAN program, there is one item that shows the ability of the test takers is low, namely in item 1. Based on the results of calculations using the help of the item output and Test Analysis program SPSS Version 2.4, the reliability test obtained a result of 0.950.*

*Keywords: Quality, Test, Items*

## ABSTRACT

Kualitas tes sangat mempengaruhi akurasi tes itu sendiri. Oleh karenanya analisis kualitas tes merupakan suatu tahap yang harus ditempuh untuk mengetahui derajat kualitas suatu tes, baik tes secara keseluruhan maupun butir soal yang menjadi bagian dari tes. Dalam penilaian hasil belajar, tes diharapkan dapat menggambarkan sampel perilaku dan menghasilkan nilai yang objektif serta akurat. Tes yang digunakan guru harus memiliki kualitas yang baik dilihat dari berbagai segi. Tes hendaknya disusun sesuai dengan prinsip dan prosedur penyusunan tes, setelah digunakan perlu diketahui apakah tes tersebut berkualitas baik atau kurang baik. Adapun jenis penelitian ini adalah penelitian experiment dengan pendekatan kuantitatif. Untuk memperoleh data tentunya diperlukan suatu alat atau instrumen. Instrumen yang digunakan peneliti adalah tes tertulis berupa tes objektif (tes pilihan ganda) yang berisi 10 soal yang dalamnya terdapat pertanyaan-pertanyaan untuk dijawab oleh partisipan. Populasi dalam penelitian ini adalah seluruh siswa kelas 3 SDIT Mutiara Global Kubang Raya dengan jumlah 52 orang. Dari hasil analisis menggunakan program ITEMAN (SPSS) Version 2.4, butir soal yang dikelompokkan baik berdasarkan tingkat kesukaran

adalah butir 2 butir. Dari hasil analisis menggunakan program SPSS version 2.4, butir soal yang dikategorikan sangat bagus dari segi daya pembeda adalah sebanyak 7 butir, sedangkan yang dikategorikan cukup bagus tidak perlu perbaikan adalah sebanyak 2 butir, yang dikategorikan kurang bagus dan perlu perbaikan adalah sebanyak 0 butir, dan yang dikategorikan jelek adalah sebanyak 0 butir. Berdasarkan hasil analisis menggunakan program ITEMAN, ada satu butir soal yang menunjukkan kemampuan peserta tes tersebut rendah, yaitu terdapat pada butir ke 1. Berdasarkan hasil perhitungan dengan menggunakan bantuan output item and Test Analysis program SPSS Version 2.4, uji reliabilitas memperoleh hasil 0,950.

---

*Kata kunci: Kualitas, Tes, Butir Soal*

---

MUMTAZA (Journal of Arabic Teaching, Arabic Linguistic and Literature)



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

## PENDAHULUAN

Dalam konteks evaluasi pendidikan, kualitas instrumen pengukuran merupakan aspek fundamental yang menentukan validitas dan reliabilitas hasil penilaian. Analisis kualitas tes dan butir soal menjadi prosedur sistematis untuk mengevaluasi karakteristik psikometrik suatu instrumen pengukuran, yang bertujuan menghasilkan alat evaluasi yang handal, tepat guna, dan mampu memberikan informasi akurat tentang kompetensi peserta didik. Menurut Azwar, proses analisis butir soal merupakan tahapan kritis dalam mengonstruksi dan mengembangkan instrumen pengukuran yang berkualitas (Azwae, 2023).

Pentingnya analisis kualitas tes tidak hanya sekadar memenuhi persyaratan metodologis, melainkan juga memberikan kontribusi signifikan terhadap perbaikan berkelanjutan sistem penilaian. Melalui analisis komprehensif, peneliti dapat mengidentifikasi kelemahan instrumen, memperbaiki butir soal yang tidak memenuhi kriteria psikometrik, serta mengoptimalkan kemampuan instrumen dalam mengukur konstruk yang dimaksud. Saifuddin menegaskan bahwa proses ini melibatkan serangkaian prosedur statistik untuk mengevaluasi karakteristik setiap butir soal dan instrumen secara keseluruhan (Saifuddin, 2022).

Beberapa parameter kunci dalam analisis kualitas tes meliputi validitas, reliabilitas, tingkat kesukaran, daya pembeda, dan efektivitas pengecoh pada instrumen pilihan ganda. Parameter tersebut memberikan informasi mendalam tentang kekuatan dan kelemahan instrumen pengukuran. Menurut penelitian Widoyoko, evaluasi komprehensif terhadap parameter-parameter tersebut memungkinkan pengembang tes untuk melakukan rekonstruksi dan penyempurnaan instrumen secara berkelanjutan (Widoyoko, 2023).

Kompleksitas analisis kualitas tes memerlukan pendekatan multidimensional yang memadukan aspek teoritis dan empiris. Proses ini tidak sekadar mengandalkan perhitungan statistik, tetapi juga mempertimbangkan konteks pedagogis dan karakteristik subjek yang diukur. Melalui sintesis pendekatan kuantitatif dan kualitatif, peneliti dapat menghasilkan instrumen pengukuran yang tidak hanya memenuhi kriteria statistik, tetapi juga selaras dengan tujuan

---

pembelajaran dan pengembangan kompetensi.

Untuk melihat kelemahan dan kekurangan tes, dapat dilakukan dengan dua cara yaitu: 1) dengan analisis soal, 2) dengan menghitung validitas serta reliabilitas soal (Firmansyah & Sridadi, 2018). Berdasarkan hal tersebut, maka penting dilakukan uji kualitas tes mata pelajaran bahasa Arab di SDIT Mutiara Global Kubang Raya, Pekanbaru. Adapun tujuan penelitian adalah untuk mengetahui a) tingkat kesukaran butir soal, b) daya pembeda butir soal, c) validitas soal, d) Uji reliabilitas butir soal. Dari serangkaian kegiatan tersebut akan dihasilkan satu paket alat ukur berupa tes yang berkualitas dilihat dari validitas dan reliabilitasnya.

Page | 70

## METODE

Penelitian ini menggunakan pendekatan kuantitatif dengan metode deskriptif. Menurut Suradika (2000), penelitian deskriptif bertujuan untuk menggambarkan fenomena sosial secara terperinci. Penelitian ini dilakukan pada siswa kelas III SDIT Mutiara Global Kubang Raya, yang berjumlah 52 orang, pada tahun 2024. Jenis penelitian ini adalah eksperimen dengan pendekatan kuantitatif menggunakan metode eksperimen semu. Metode eksperimen digunakan untuk menganalisis kualitas tes dan butir soal mata pelajaran Bahasa Arab.

Instrumen penelitian meliputi tes tertulis dan studi dokumentasi. Tes tertulis berupa soal pilihan ganda sebanyak 10 butir yang dijawab oleh partisipan, sementara studi dokumentasi dilakukan dengan menganalisis dokumen soal Ujian Sekolah (US) kelas III SDIT Mutiara Global Kubang Raya (Wirandani, dkk., 2019). Teknik pengumpulan data mencakup beberapa langkah, yaitu menghimpun soal Ujian Sekolah (US) mata pelajaran Bahasa Arab, memberikan kode pada setiap muatan soal yang mencakup tahun pelajaran, kompetensi dan indikator SKL, nomor halaman, serta nomor soal. Setelah itu, dilakukan analisis data dan penyusunan kesimpulan berdasarkan persentase hasil penelitian (Mujib, dkk, 2018).

Teknik analisis data difokuskan pada analisis butir tes pilihan ganda untuk mengetahui kualitas tes mata pelajaran Bahasa Arab. Analisis ini mencakup pengukuran tingkat kesukaran butir tes, daya pembeda, validitas, dan reliabilitas tes. Tingkat kesukaran diukur dengan membandingkan jumlah responden yang menjawab benar terhadap jumlah responden secara keseluruhan. Hasil uji daya pembeda mengelompokkan butir tes ke dalam kategori baik dan unggul. Uji validitas menunjukkan bahwa seluruh butir tes valid, sementara uji reliabilitas menyatakan bahwa tes yang dirancang dapat diandalkan. Dengan demikian, butir tes yang dirancang dianggap tidak memerlukan perbaikan dan layak digunakan untuk mengukur keterampilan berpikir kritis serta pengetahuan prosedural dalam grammar Bahasa Arab (Alfiana, dkk, 2021).

## HASIL DAN PEMBAHASAN

Untuk mencapai tujuan penelitian, maka data yang dikumpulkan berupa lembar jawaban peserta tes, dianalisis secara kuantitatif dan hasilnya sebagai berikut:

### Tingkat Kesukaran Soal

Soal yang baik adalah soal yang tidak terlalu mudah atau tidak terlalu sukar. Soal yang terlalu mudah tidak merangsang siswa untuk mempertinggi usaha memecahkannya. Sebaliknya soal yang terlalu sukar akan menyebabkan siswa menjadi putus asa dan tidak mempunyai semangat untuk mencoba lagi karena di luar jangkauannya (Arikunto, 2021). Adapun bilangan yang menunjukkan sukar dan mudahnya sesuatu soal disebut indeks kesukaran (difficulty index). Besarnya indeks kesukaran antara 0,00 sampai dengan 1,0. Indeks kesukaran ini menunjukkan taraf kesukaran soal. Soal dengan indeks kesukaran 0,0 menunjukkan bahwa soal itu terlalu sukar, sebaliknya indeks 1,0 menunjukkan bahwa soalnya terlalu mudah.



Dalam istilah evaluasi, indeks kesukaran ini diberi symbol P (p besar), singkatan dari kata “proporsi”. Dengan demikian maka soal dengan  $P = 0,70$  lebih mudah jika dibandingkan dengan  $P = 0,20$ . Sebaliknya soal dengan  $P = 0,30$  lebih sukar daripada soal dengan  $P = 0,80$ . Menurut ketentuan yang sering diikuti, indeks kesukaran sering diklasifikasikan sebagai berikut (Robert I, Thorndike, dan Elizabeth, 1969): 1) soal dengan = kurang dari 0,30 adalah soal terlalu sukar; 2) soal dengan = 0,30 s/d 0,70 adalah soal cukup (sedang); 3) soal dengan = lebih dari 0,70 adalah soal terlalu mudah.

|      |         | V2  | V3  | V4  | V5  | V6   | V7   | V8  | V9   | V10  |
|------|---------|-----|-----|-----|-----|------|------|-----|------|------|
| N    | Valid   | 53  | 53  | 53  | 53  | 53   | 53   | 53  | 53   | 53   |
|      | Missing | 0   | 0   | 0   | 0   | 0    | 0    | 0   | 0    | 0    |
| Mean |         | .75 | .75 | .96 | .98 | 1.00 | 1.02 | .98 | 1.09 | 1.13 |

Dari hasil analisis menggunakan program ITEMAN (SPSS), butir soal yang dikelompokkan baik berdasarkan tingkat kesukaran 2 butir, sedangkan pengelompokan mudah berdasarkan tingkat kesukaran ada 7, Kemudian peneliti tidak menemukan soal yang berkriteria sulit pada soal ini dan hasil untuk kriteria Sulit.

### Analisis Daya Pembeda

Daya pembeda soal adalah kemampuan sesuatu soal untuk membedakan antara siswa yang pandai (berkemampuan tinggi) dengan siswa yang kurang pintar (berkemampuan rendah) (Arikunto, 2021). Angka yang menunjukkan besarnya daya pembeda disebut indeks diskriminasi, disingkat (d besar). Indeks diskriminasi berkisar antara 0,00 sampai 1,00. Hanya bedanya, indeks kesukaran tidak mengenal tanda negatif (-), tetapi pada indeks diskriminasi ada tanda negatif. Tanda negatif pada indeks diskriminasi digunakan jika sesuatu soal “terbalik” menunjukkan

kualitas testee (Arikunto, 2021).

D = 0,00 ----- 0,20 → Jelek (poor)  
D = 0,21 ----- 0,40 → Cukup (satisfactory)  
D = 0,41 ----- 0,70 → Baik (good)  
D = 0,71 ----- 1,00 → Baik Sekali (exellent)

D = Negatif, semuanya → tidak baik, jadi semua butir soal yang mempunyai nilai D negatif sebaiknya dibuang saja.

Page | 72

|     | Scale Mean if Item Deleted | Scale Variance if Item Deleted | Corrected Item-Total Correlation | Cronbach's Alpha if Item Deleted |
|-----|----------------------------|--------------------------------|----------------------------------|----------------------------------|
| V2  | 7.92                       | 39.802                         | .448 (baik)                      | .962                             |
| V3  | 7.92                       | 38.225                         | .615 (baik)                      | .957                             |
| V4  | 7.72                       | 37.091                         | .851 (baik sekali)               | .950                             |
| V5  | 7.70                       | 35.561                         | .892 (baik sekali)               | .946                             |
| V6  | 7.68                       | 33.876                         | .935 (baik sekali)               | .942                             |
| V7  | 7.66                       | 32.382                         | .945 (baik sekali)               | .941                             |
| V8  | 7.70                       | 30.676                         | .944 (baik sekali)               | .941                             |
| V9  | 7.58                       | 29.671                         | .955 (baik sekali)               | .941                             |
| V10 | 7.55                       | 28.291                         | .962 (baik sekali)               | .943                             |

Dari hasil analisis menggunakan program ITEMAN (SPSS), butir soal yang dikategorikan sangat bagus dari segi daya pembeda adalah sebanyak 7 butir, sedangkan yang dikategorikan bagus adalah sebanyak 2 butir, yang dikategorikan cukup bagus dan perlu perbaikan adalah sebanyak 0 butir, dan yang dikategorikan jelek adalah sebanyak 0 butir.

### Analisis Validitas

Istilah validitas dipakai dalam tiga hal: (1) validitas penelitian, (2) validitas soal, dan (3) validitas alat ukur (Sudijono, 2003) Validitas penelitian adalah derajat kesesuaian hasil penelitian dengan keadaan sebenarnya, atau sejauh mana hasil penelitian menggambarkan keadaan yang sebenarnya. Validitas soal berkaitan dengan kemampuan soal untuk mengukur apa yang dimaksudkan untuk diukur. Validitas alat ukur adalah ukuran seberapa tepat alat ukur tersebut dalam melakukan fungsi pengukuran. Dan pada penelitian saya kali ini yaitu menggunakan validitas soal. Adapun jenis validitas soal yaitu jenis Validitas Isi (Content Validity): Menilai sejauh mana isi soal mencakup semua aspek dari materi yang diajarkan. Ini biasanya dilakukan dengan memeriksa kesesuaian antara tujuan pembelajaran dan butir soal.

| Correlations |                 | respon | V1  | V2   | V3   | V4   | V5   | V6   | V7   | V8   | V9   | V10  | Tot  |
|--------------|-----------------|--------|-----|------|------|------|------|------|------|------|------|------|------|
|              |                 | den    |     |      |      |      |      |      |      |      |      |      | al   |
| Respon       | Pearson         | 1      | -   | -    | -    | -    | -    | -    | -    | -    | -    | -    | -    |
|              | Correlat        |        | .21 | .662 | .638 | .389 | .428 | .419 | .398 | .455 | .300 | .313 | .884 |
|              | ion             |        | 2   | **   | **   | **   | **   | **   | **   | **   | *    | *    | **   |
|              | Sig. (2-tailed) |        | .13 | .000 | .000 | .004 | .002 | .002 | .004 | .001 | .031 | .024 | .000 |
| V1           | N               | 52     | 52  | 52   | 52   | 52   | 52   | 52   | 52   | 52   | 52   | 52   | 52   |
|              | Pearson         | -.212  | 1   | .046 | .039 | -    | -    | .000 | .098 | -    | .021 | .092 | .168 |

|     | Correlat<br>ion            |         |               |            |            | .018       | .007       |            |            | .004       |            |            |            |
|-----|----------------------------|---------|---------------|------------|------------|------------|------------|------------|------------|------------|------------|------------|------------|
|     | Sig. (2-<br>tailed)        | .132    |               | .745       | .779       | .898       | .958       | 1.00       | .485       | .975       | .883       | .514       | .233       |
|     | N                          | 52      | 53            | 53         | 53         | 53         | 53         | 53         | 53         | 53         | 53         | 53         | 52         |
| V2  | Pearson<br>Correlat<br>ion | -.662** | .04<br>6      | 1          | .645<br>** | .352<br>** | .493<br>** | .425<br>** | .375<br>** | .415<br>** | .365<br>** | .376<br>** | .667<br>** |
|     | Sig. (2-<br>tailed)        | .000    | .74<br>5      |            | .000       | .010       | .000       | .002       | .006       | .002       | .007       | .005       | .000       |
|     | N                          | 52      | 53            | 53         | 53         | 53         | 53         | 53         | 53         | 53         | 53         | 53         | 52         |
| V3  | Pearson<br>Correlat<br>ion | -.638** | .03<br>9      | .645<br>** | 1          | .506<br>** | .645<br>** | .505<br>** | .560<br>** | .525<br>** | .593<br>** | .575<br>** | .567<br>** |
|     | Sig. (2-<br>tailed)        | .000    | .77<br>9      | .000       |            | .000       | .000       | .000       | .000       | .000       | .000       | .000       | .000       |
|     | N                          | 52      | 53            | 53         | 53         | 53         | 53         | 53         | 53         | 53         | 53         | 53         | 52         |
| V4  | Pearson<br>Correlat<br>ion | -.389** | -<br>.01<br>8 | .352<br>** | .506<br>** | 1          | .758<br>** | .832<br>** | .839<br>** | .851<br>** | .828<br>** | .836<br>** | .523<br>** |
|     | Sig. (2-<br>tailed)        | .004    | .89<br>8      | .010       | .000       |            | .000       | .000       | .000       | .000       | .000       | .000       | .000       |
|     | N                          | 52      | 53            | 53         | 53         | 53         | 53         | 53         | 53         | 53         | 53         | 53         | 52         |
| V5  | Pearson<br>Correlat<br>ion | -.428** | -<br>.00<br>7 | .493<br>** | .645<br>** | .758<br>** | 1          | .837<br>** | .819<br>** | .867<br>** | .859<br>** | .871<br>** | .441<br>** |
|     | Sig. (2-<br>tailed)        | .002    | .95<br>8      | .000       | .000       | .000       |            | .000       | .000       | .000       | .000       | .000       | .001       |
|     | N                          | 52      | 53            | 53         | 53         | 53         | 53         | 53         | 53         | 53         | 53         | 53         | 52         |
| V6  | Pearson<br>Correlat<br>ion | -.419** | .00<br>0      | .425<br>** | .505<br>** | .832<br>** | .837<br>** | 1          | .913<br>** | .943<br>** | .918<br>** | .927<br>** | .564<br>** |
|     | Sig. (2-<br>tailed)        | .002    | 1.0<br>00     | .002       | .000       | .000       | .000       |            | .000       | .000       | .000       | .000       | .000       |
|     | N                          | 52      | 53            | 53         | 53         | 53         | 53         | 53         | 53         | 53         | 53         | 53         | 52         |
| V7  | Pearson<br>Correlat<br>ion | -.398** | .09<br>8      | .375<br>** | .560<br>** | .839<br>** | .819<br>** | .913<br>** | 1          | .911<br>** | .956<br>** | .963<br>** | .564<br>** |
|     | Sig. (2-<br>tailed)        | .004    | .48<br>5      | .006       | .000       | .000       | .000       | .000       |            | .000       | .000       | .000       | .000       |
|     | N                          | 52      | 53            | 53         | 53         | 53         | 53         | 53         | 53         | 53         | 53         | 53         | 52         |
| V8  | Pearson<br>Correlat<br>ion | -.455** | -<br>.00<br>4 | .415<br>** | .525<br>** | .851<br>** | .867<br>** | .943<br>** | .911<br>** | 1          | .927<br>** | .938<br>** | .569<br>** |
|     | Sig. (2-<br>tailed)        | .001    | .97<br>5      | .002       | .000       | .000       | .000       | .000       | .000       |            | .000       | .000       | .000       |
|     | N                          | 52      | 53            | 53         | 53         | 53         | 53         | 53         | 53         | 53         | 53         | 53         | 52         |
| V9  | Pearson<br>Correlat<br>ion | -.300*  | .02<br>1      | .365<br>** | .593<br>** | .828<br>** | .859<br>** | .918<br>** | .956<br>** | .927<br>** | 1          | .965<br>** | .376<br>** |
|     | Sig. (2-<br>tailed)        | .031    | .88<br>3      | .007       | .000       | .000       | .000       | .000       | .000       | .000       |            | .000       | .006       |
|     | N                          | 52      | 53            | 53         | 53         | 53         | 53         | 53         | 53         | 53         | 53         | 53         | 52         |
| V10 | Pearson<br>Correlat<br>ion | -.313*  | .09<br>2      | .376<br>** | .575<br>** | .836<br>** | .871<br>** | .927<br>** | .963<br>** | .938<br>** | .965<br>** | 1          | .471<br>** |
|     | Sig. (2-<br>tailed)        | .024    | .51           | .005       | .000       | .000       | .000       | .000       | .000       | .000       | .000       |            | .000       |

|                                                              |          |         |     |      |      |      |      |      |      |      |      |      |    |
|--------------------------------------------------------------|----------|---------|-----|------|------|------|------|------|------|------|------|------|----|
|                                                              | tailed)  |         | 4   |      |      |      |      |      |      |      |      |      |    |
|                                                              | N        | 52      | 53  | 53   | 53   | 53   | 53   | 53   | 53   | 53   | 53   | 53   | 52 |
| Total                                                        | Pearson  | -.884** | .16 | .667 | .567 | .523 | .441 | .564 | .564 | .569 | .376 | .471 | 1  |
|                                                              | Correlat |         | 8   | **   | **   | **   | **   | **   | **   | **   | **   | **   |    |
|                                                              | ion      |         |     |      |      |      |      |      |      |      |      |      |    |
|                                                              | Sig. (2- | .000    | .23 | .000 | .000 | .000 | .001 | .000 | .000 | .000 | .006 | .000 |    |
|                                                              | tailed)  |         | 3   |      |      |      |      |      |      |      |      |      |    |
|                                                              | N        | 52      | 52  | 52   | 52   | 52   | 52   | 52   | 52   | 52   | 52   | 52   | 52 |
| **. Correlation is significant at the 0.01 level (2-tailed). |          |         |     |      |      |      |      |      |      |      |      |      |    |
| *. Correlation is significant at the 0.05 level (2-tailed).  |          |         |     |      |      |      |      |      |      |      |      |      |    |

Berdasarkan hasil analisis menggunakan program (SPSS) Verson 2.4, ada satu butir soal yang menunjukkan kemampuan peserta tes tersebut rendah, yaitu terdapat pada butir ke 1.

### Uji Reliabilitas

Reliabilitas merupakan salah satu karakteristik penting dalam pengukuran kualitas instrumen tes. Secara umum, reliabilitas merujuk pada konsistensi atau stabilitas hasil pengukuran yang dilakukan. Dalam konteks tes atau instrumen pengukuran, reliabilitas menunjukkan sejauh mana suatu alat ukur dapat memberikan hasil yang relatif sama atau konsisten bila dilakukan pengukuran berulang pada subjek yang sama.

Penelitian ini menggunakan uji validitas pendekatan belah dua yang populer adalah formula Spearman-Brown. Formula ini mengasumsikan kedua belahan tes adalah paralel (mean dan varian setara). Keuntungan formula ini dapat dipakai baik untuk skor dikotomi maupun nondikotomi.

|                                        |                  |            |                |
|----------------------------------------|------------------|------------|----------------|
| Cronbach's Alpha                       | Part 1           | Value      | .881           |
|                                        |                  | N of Items | 5 <sup>a</sup> |
|                                        | Part 2           | Value      | .980           |
|                                        |                  | N of Items | 4 <sup>b</sup> |
|                                        | Total N of Items |            | 9              |
| Correlation Between Forms              |                  |            | .905           |
| Spearman-Brown Coefficient             | Equal Length     |            | .950           |
|                                        | Unequal Length   |            | .951           |
| Guttman Split-Half Coefficient         |                  |            | .880           |
| a. The items are: V2, V3, V4, V5, V6.  |                  |            |                |
| b. The items are: V6, V7, V8, V9, V10. |                  |            |                |

Berdasarkan hasil perhitungan dengan menggunakan bantuan output item program ITEMAN SPSS Version 2.4, uji reliabilitas memperoleh hasil 0,950.

### KESIMPULAN

Dari data yang terkumpul melalui lembar jawaban peserta ujian, disekor, dihitung banyak peserta menjawab betul atau banyak penjawab pada masing-masing alternatif jawaban, kemudian dimasukkan kedalam SPSS sehingga diperoleh nilai P dan niali B. Nilai P dan B untuk setiap butir soal ditentukan kualifikasi butir soal berdasarkan criteria yang telah dikemukakan di atas sehingga

dapat disimpulkan: 1) Butir soal yang dikelompokkan baik berdasarkan tingkat kesukaran adalah butir nomer 2, dan 3. sedangkan pengelompokan mudah berdasarkan tingkat kesukaran ada 7, Kemudian peneliti tidak menemukan soal yang berkriteria sulit pada soal ini dan hasil untuk kriteria Sulit. Banyak butir soal yang dikelompokkan soal baik 2 butir dan mudah 7 butir ditinjau dari segi tingkat kesukaran soal adalah 9 butir; 2) Butir soal yang dikategorikan baik sekali dari segi daya pembeda soal adalah 4, 5, 6, 7, 8, 9, dan 10, sedangkan butir dikategorikan baik (cukup memuaskan) adalah butir 2, dan 3; 3) Dari penjelasan di atas tingkat kesukaran soal (baik) ditemukan nol (tidak ada), daya pembeda soal (baik sekali, cukup memuaskan) maka ditemukan 9 butir soal masih ditolerir untuk digunakan yaitu butir soal 2, 3, 4, 5, 6, 7, 8, 9, dan 10; 4) Berdasarkan kesimpulan butir soal yang kurang baik dari segi tingkat kesukaran soal, daya pembeda butir soal sebanyak 9 butir, kemudian dianalisis efektifitas distraktor ternyata hanya option yang tidak efektif hanya terdapat pada butir ke 1 option A.

## DAFTAR RUJUKAN

- Anastasi, A., & Urbina, S. (1997). *Psychological Testing*. Upper Saddle River, NJ: Prentice Hall.
- Arikunto, S. (2021). *Dasar-dasar evaluasi pendidikan edisi 3*. Bumi aksara.
- Arikunto, S. (2022). *Prosedur Penelitian: Suatu Pendekatan Praktik*. Jakarta: Rineka Cipta.
- Azwar, S. (2023). *Konstruksi Tes Kemampuan: Konsep dan Teknik Penyusunan*. Pustaka Pelajar.
- Crocker, L., & Algina, J. (1986). *Introduction to Classical and Modern Test Theory*. New York: Holt, Rinehart and Winston.
- Ebel, R. L., & Frisbie, D. A. (1991). *Essentials of Educational Measurement*. Englewood Cliffs, NJ: Prentice Hall.
- Majid, A. (2014). *Pembelajaran Tematik Terpadu*. PT Remaja Rosdakarya.
- Majid, A., & Firdaus, A. S. (2014). *Penilaian Autentik Proses dan Hasil Belajar*. Interes Media.
- Muna, W. (2023). *Evaluasi Instrumen Tes Bahasa Arab: Pendekatan Multidimensional*. Prenadamedia Group.
- Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric Theory*. New York: McGraw-Hill.
- Nurgiyantoro, B. (2010). *Penilaian Pembelajaran Bahasa*. BPFE.
- Nurgiyantoro, Burhan, *Penilaian Pembelajaran Bahasa Berbasis Kompetensi*, Yogyakarta: BPEF-Yogyakarta, 2013
- Saifuddin, A. (2022). *Evaluasi Instrumen Pengukuran Pendidikan*. Raja Grafindo Persada.
- Sudijono, Anas. *Pengantar Evaluasi Pendidikan*. Jakarta: RajaGrafindo, 2003.
- Thorndike, Robert L., and Elizabeth Hagen. (1969). "Measurement and evaluation in psychology and education".
- Widoyoko, E. P. (2023). "Analisis Butir Soal dalam Pengembangan Instrumen Penilaian". *Jurnal Penelitian Pendidikan Indonesia*, 9(2), 145-160.
- Wirandani, T., Ayu Cendra Kasih,, Latifah. ANALISIS BUTIR SOAL HOTS (HIGH ORDER THINKING SKILL) PADA SOAL UJIAN SEKOLAH KELAS XII MATA PELAJARAN BAHASA INDONESIA DI SMK ANNAHL. Parole (*Jurnal Pendidikan Bahasa dan Sastra Indonesia*), Volume 2 Nomor 4, Juli 2019.