Cerdika: Jurnal Ilmiah Indonesia, April 2025, 5 (4), 1530-1542
p-ISSN: 2774-6291 e-ISSN: 2774-6534

Available online at http://cerdika.publikasiindonesia.id/index.php/cerdika/index

Studi Validasi dan Reliabilitas Butir Soal Tes Sumatif dalam Pendidikan Bahasa Arab

M. Baihaqi[1], Mahmudah[2], Ahmad Jundi Al Mubarak[3], Nabilah Rabbaniyah[4], Dwi Susanto[5]
Universitas Sunan Ample, Indonesia[1,2,5]
Universitas Telkom Surabaya, Indonesia[3,4]
Email: baihaqi@uinsa.ac.id[1], mahmudahsg129@gmail.com[2], ahmadjundialmubarak@gmail.com[3], nabilahrobbaniyah@gmail.com[4], dwisusanto@uinsa.ac.id[5]

Abstrak
Penilaian sumatif sangat penting dalam mengevaluasi hasil belajar siswa, namun efektivitasnya bergantung pada validitas dan reliabilitas butir soal. Penelitian ini bertujuan untuk menganalisis validitas dan reliabilitas butir soal sumatif dalam pendidikan Bahasa Arab untuk siswa kelas XII di SMK Maarif NU Sunan Giri Driyorejo, mengatasi kesenjangan dalam analisis sistematis butir soal yang sering diabaikan oleh pendidik. Dengan pendekatan kuantitatif dan metode *ex post facto*, data dikumpulkan dari 30 soal pilihan ganda yang diujikan kepada 60 siswa dan dianalisis menggunakan perangkat lunak SPSS. Hasil menunjukkan bahwa 28 butir soal valid, sedangkan 2 tidak valid, dengan koefisien reliabilitas 0,742, menandakan instrumen yang sangat konsisten. Temuan ini menegaskan pentingnya analisis rutin butir soal untuk memastikan keselarasan dengan tujuan pembelajaran dan standar kurikulum. Penelitian ini memberikan panduan praktis bagi guru untuk menyempurnakan alat penilaian, sehingga meningkatkan akurasi dan kualitas evaluasi pendidikan.

Kata kunci: Reliabilitas, Soal, Sumatif, Validitas

Abstract
Summative assessments are critical in evaluating student learning outcomes, yet their effectiveness hinges on the validity and reliability of test items. This study aims to analyze the validity and reliability of summative test items in Arabic language education for twelfth-grade students at SMK Maarif NU Sunan Giri Driyorejo, addressing the gap in systematic test item analysis often overlooked by educators. Using a quantitative approach with ex post facto methods, data were collected from 30 multiple-choice questions administered to 60 students and analyzed using SPSS software. Results revealed that 28 items were valid, while 2 were invalid, with a reliability coefficient of 0.742, indicating a highly consistent instrument. The findings underscore the importance of routine test item analysis to ensure alignment with learning objectives and curriculum standards. This study provides practical guidance for teachers to refine assessment tools, thereby enhancing the accuracy and quality of educational evaluations.

***Keywords**: Reliability, Summative, Test Items, Validity*

Article Info:
Submitted: 08-02-25 ***Final Revised***: 17-04-25 ***Accepted***: 25-04-25 ***Published***: 26-04-25

**Correspondence Author: M. Baihaqi*
Email: baihaqi@uinsa.ac.id

PENDAHULUAN

Evaluasi pembelajaran merupakan komponen esensial dalam proses Pendidikan (Armini, 2024; Rusnawati, 2020). Evaluasi hadir sebagai alat untuk mengukur pencapaian tujuan yang telah ditetapkan, sekaligus menjadi panduan untuk mencapai tujuan tersebut. Dengan demikian, evaluasi memiliki peran yang sangat penting dalam keberhasilan pendidikan (Bariyyah et al., 2018). Namun, kelemahan pada berbagai instrumen tes dapat mengurangi tingkat akurasi pengukuran dalam praktik pendidikan. Padahal, evaluasi yang akurat dan tepat berkontribusi signifikan terhadap pengambilan keputusan yang lebih baik dalam pendidikan. Salah satu solusi untuk mengatasi masalah ini adalah penerapan tes standar yang memenuhi syarat validitas dan reliabilitas (Bashooir & Supahar, 2018).

Di Indonesia, ujian nasional serta berbagai tes yang dirancang oleh guru menjadi bagian integral dari evaluasi (Pahri, 2021). Tes didefinisikan sebagai kumpulan pertanyaan dengan jawaban benar atau salah, yang bertujuan untuk mengukur kemampuan dan pemahaman seseorang (Defitasari et al., 2011). Dalam konteks pembelajaran Bahasa Arab, salah satu cara untuk menilai kemampuan siswa adalah melalui tes. Tes ini dapat berupa tugas atau latihan yang dirancang untuk mengukur berbagai aspek kemampuan siswa (Rosmana et al., 2024). Mengingat peran dan fungsinya yang sangat penting, berbagai tes diharapkan memenuhi standar validitas dan reliabilitas. Evaluasi keberhasilan pembelajaran mencakup tiga komponen utama: aspek kognitif (pengetahuan), aspek psikomotorik (keterampilan fisik), dan aspek afektif (sikap, nilai, dan emosi siswa terhadap materi).

Tes sumatif merupakan penilaian yang dilakukan pada akhir semester dalam kalender pendidikan, dan hasilnya dicantumkan dalam rapor untuk menentukan kelulusan atau kenaikan kelas (Anshari et al., 2024). Dalam pelaksanaannya, hasil tes sumatif dibandingkan dengan kriteria ketuntasan minimal (KKM) untuk mengevaluasi sejauh mana tujuan pembelajaran telah tercapai (Anshari et al., 2024). Tes ini menjadi alat penting dalam menilai kemampuan siswa secara keseluruhan. Salah satu bentuk tes yang umum digunakan adalah pilihan ganda, di mana siswa diminta memilih jawaban paling tepat dari beberapa opsi yang tersedia. Tes semacam ini menuntut siswa untuk berpikir kritis dan cermat. Agar efektif, setiap butir soal dalam tes harus memiliki validitas dan reliabilitas yang memadai (Wulandari & Pramusinto, 2020). Analisis butir soal diperlukan untuk mengidentifikasi kelemahan atau peluang perbaikan, sehingga pengukuran kemampuan siswa menjadi lebih valid dan kredibel. Perbaikan ini diperlukan agar mampu menunjang proses pembelajaran yang efektif (Jundi et al., 2025).

Dalam penelitian sebelumnya, Validitas dan reliabilitas suatu instrumen dipengaruhi oleh subjek yang diukur, pengguna instrumen, dan instrumen itu sendiri (Yusup, 2018). Proses ini dipengaruhi oleh respons siswa, kondisi lokasi

penelitian, serta kesesuaian alat yang digunakan. Validitas dan reliabilitas hasil penelitian berpengaruh langsung pada kualitas dan kuantitas kesimpulan yang dihasilkan. Validitas dan reliabilitas merupakan isu mendasar dalam penelitian ilmiah, yang menentukan nilai penelitian ilmiah (Arslan, 2022).

Penelitian ini sangat relevan karena observasi awal di lapangan menunjukkan bahwa beberapa guru, khususnya guru Bahasa Arab di SMK Ma'arif NU Sunan Giri, belum menerapkan analisis butir soal pada tes sumatif tengah semester ganjil tahun ajaran 2024/2025. Hal ini disebabkan oleh berbagai faktor, seperti keterbatasan waktu akibat pemadatan materi dan kurangnya pemanfaatan teknologi. Sementara pembelajaran Bahasa arab sangat perlu menggunakan teknologi kekinian (M. Baihaqi et al., 2023). Meskipun tes sumatif telah dilaksanakan, analisis spesifik terhadap soal belum dilakukan. Padahal, analisis ini sangat penting untuk mengevaluasi kualitas soal. Seorang pendidik harus memiliki kemampuan untuk menganalisis soal secara mendalam guna memastikan instrumen yang digunakan benar-benar efektif dalam mengukur pencapaian belajar siswa. Sebagaimana guru-guru sekolah dasar di Filipina dan Indonesia menunjukkan kemampuan yang baik dalam menganalisis pertanyaan sains dengan pemikiran tingkat tinggi, masing-masing dengan 81% dan 79,8% menganalisis pertanyaan dengan benar (Atmojo et al., 2019).

Penelitian ini bertujuan mendorong guru Bahasa Arab untuk merefleksikan kualitas soal sumatif yang digunakan, dengan harapan dapat memberikan panduan praktis dalam memperbaiki butir soal yang kurang valid atau kredibel. Dengan fokus pada analisis validitas dan reliabilitas soal sumatif tengah semester ganjil, penelitian ini tidak hanya memberikan wawasan tentang kualitas instrumen evaluasi, tetapi juga berkontribusi dalam meningkatkan efektivitas dan akurasi penilaian terhadap siswa. Secara khusus, penelitian ini diharapkan membantu guru di SMK Ma’arif NU Sunan Giri Driyorejo serta guru lainnya dalam mendukung pencapaian tujuan pembelajaran Bahasa Arab secara lebih optimal.

Penelitian ini memperkenalkan pendekatan berbasis data untuk memvalidasi dan menilai reliabilitas butir soal sumatif dalam pendidikan Bahasa Arab, khususnya untuk siswa kelas XII di SMK Maarif NU Sunan Giri Driyorejo. Berbeda dengan studi sebelumnya yang sering berfokus pada prinsip validitas dan reliabilitas umum (Yusup, 2018; Arslan, 2022), penelitian ini memberikan bukti empiris menggunakan perangkat lunak SPSS untuk menganalisis 30 butir soal, mengidentifikasi 28 butir valid dan 2 tidak valid, dengan koefisien reliabilitas tinggi sebesar 0,742. Selain itu, penelitian ini mengatasi kesenjangan dalam penerapan praktis dengan menekankan pentingnya analisis rutin butir soal oleh guru, sebuah praktik yang sering diabaikan karena keterbatasan waktu dan penggunaan teknologi (M. Baihaqi et al., 2023). Penelitian juga menyoroti sifat dinamis validitas tes, menyarankan pembaruan berkala untuk menyesuaikan dengan perubahan kurikulum, serta memberikan rekomendasi spesifik untuk memperbaiki butir soal yang tidak valid, yang kurang ditekankan dalam penelitian sebelumnya (Anshari et

al., 2024; Wulandari & Pramusinto, 2020).

METODE PENELITIAN

Penelitian ini menggunakan pendekatan kuantitatif dengan metode *ex post facto,* di mana data dikumpulkan setelah peristiwa tertentu terjadi. Penelitian dilakukan di SMK Maarif NU Sunan Giri, yang berlokasi di Jalan Raya Mulung No. 129, Driyorejo, Gresik, setelah penilaian tengah semester ganjil tahun akademik 2024/2025. Data yang diperlukan diperoleh melalui dua pendekatan utama, yaitu dokumentasi dan wawancara dengan guru Bahasa Arab. Penelitian ini bertujuan untuk menganalisis hubungan antara variabel-variabel yang telah ada tanpa melakukan intervensi terhadap variabel tersebut.

Penelitian ini melibatkan 60 siswa dari kelas XIIAK, XIITPM, dan XIITKR sebagai subjek penelitian, serta guru Bahasa Arab sebagai informan. Sampel penelitian berupa kumpulan soal ujian sumatif tengah semester ganjil mata pelajaran Bahasa Arab, yang terdiri atas tiga puluh soal pilihan ganda. Tujuan dari penelitian ini adalah untuk memperoleh informasi terkait materi ujian, hasil ujian, dan identitas siswa dari ketiga kelas tersebut.

Pengumpulan data dilakukan dengan cara mendokumentasikan lembar jawaban dan soal ujian siswa. Selanjutnya, setiap item soal diuji validitas dan kredibilitasnya menggunakan jawaban benar dan salah dari ujian tersebut. Data yang diperoleh kemudian dianalisis dengan metode deskriptif kuantitatif, di mana hasil ujian siswa diubah menjadi angka-angka. Analisis ini bertujuan untuk memastikan bahwa setiap item soal yang digunakan dalam mengukur pemahaman siswa terhadap materi pelajaran Bahasa Arab memiliki tingkat validitas dan reliabilitas yang memadai.

Adapun tahapan pengelolaan dan penganalisisannya sebagai berikut:

Pertama, Menentukan nilai validitas butir item dengan rumus:

$$r_{xy} = \frac{N\sum XY - (\sum X)(\sum Y)}{\sqrt{\left\{N\sum X^2 - (\sum X)^2\right\}\left\{N\sum Y^2 - (\sum Y)^2\right\}}}$$

Keterangan :

rxy: koefisien korelasi yang dicari N: banyaknya peserta tes

X: nilai variabel X (skor item) Y: nilai variabel Y (skor item)

Adapun dasar pengambilan keputusan pada uji validitas adalah: Jika rhitung > rtabel maka butir item valid.

Jika rhitung < rtabel maka butir item tidak valid.

Kedua, menentukan nilai reliabilitas. Cara menghitung reliabilitas suatu tes adalah dengan menggunakan rumus:

$$r_{11} = \left(\frac{k}{k-1} \right) \left(1 - \frac{\Sigma s_i}{s_t} \right)$$

Keterangan:

r11 = Reliabilitas instrumen

k = Jumlah butir soal

ΣSi = Jumlah varians masing-masing item

St = Varian soal

Adapun dasar pengambilan keputusan pada uji reliabilitas adalah: Jika *Cronbach alpha* > 0,6 maka instrumen tes dapat dikatakan reliabel

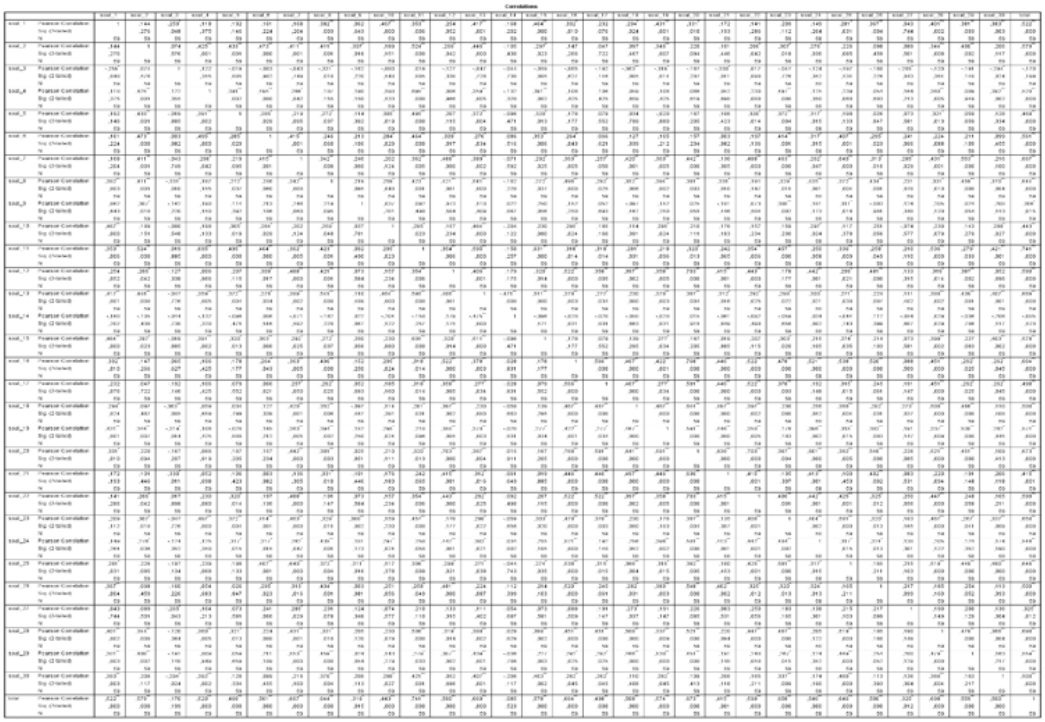
Jika *Cronbach alpha* < 0,6 maka instrumen tes dapat dikatakan tidak reliabel.

HASIL DAN PEMBAHASAN

1. Uji Validitas

Peneliti melakukan pengujian pada tes sumatif Tengah Semester Ganjil mata pelajaran Bahasa Arab, yang terdiri atas 30 butir soal pilihan ganda. Pengujian ini dilakukan di dua kelas, yaitu XIIAK dan XIITKR, dengan total 60 siswa sebagai responden. Dalam penelitian ini, validitas empiris setiap butir soal dianalisis menggunakan korelasi. Suatu butir soal dinyatakan valid apabila nilai koefisien korelasi yang diperoleh (r-hitung) sama dengan atau lebih besar dari nilai korelasi yang tercantum dalam tabel (r-tabel) pada tingkat signifikansi 5%. Hasil pengujian tersebut disajikan sebagai berikut:

Tabel 1. Hasil Validitas Butir Soal berdasarkan SPSS



Tabel 2. Intervensi Hasil Perhitungan Validitas

No. Soal	Korelasi (r)	Signifikansi (p-value)	Keputusan	Intervensi yang Dibutuhkan
1	0,522	0,000	Valid	Tidak perlu intervensi
2	0,579	0,000	Valid	Tidak perlu intervensi
3	-0,170	0,199	Tidak Valid	Perbaikan butir soal diperlukan
29	0,559	0,000	Valid	Tidak perlu intervensi
30	0,500	0,000	Valid	Tidak perlu intervensi

Catatan

a) Korelasi (r) menunjukkan kekuatan hubungan butir soal dengan total skor.
b) Signifikansi (p-value) mengacu pada uji hipotesis di mana $p < 0,05$ menunjukkan bahwa korelasi signifikan.
c) Keputusan validitas ditentukan berdasarkan nilai r dan p-value. Jika r positif dan $p < 0,05$, maka soal dinyatakan valid.
d) Intervensi yang dibutuhkan ditujukan untuk butir soal yang tidak valid. Perlu adanya revisi butir soal atau penghapusan butir yang tidak valid.

1Dari paparan diatas menunjukkan bahwa 28 butir soal dari tes sumatif Tengah Semester Ganjil mata pelajaran Bahasa Arab dinyatakan valid. Hasil ini mencerminkan keberhasilan sebagian besar butir soal dalam mengukur pemahaman siswa terhadap materi pelajaran. Validitas butir soal yang tinggi mengindikasikan bahwa instrumen evaluasi mampu menggambarkan tingkat pemahaman siswa sesuai dengan tujuan pembelajaran yang telah ditetapkan. Selain itu, validitas yang dicapai juga mencerminkan upaya guru dalam menyusun instrumen evaluasi yang sesuai dengan kurikulum dan materi pembelajaran yang diterapkan di SMK Maarif NU Sunan Giri Driyorejo.

Namun demikian, terdapat 2 butir soal yang dinyatakan tidak valid. Hal ini menunjukkan adanya kelemahan yang perlu diperbaiki untuk meningkatkan kualitas instrumen evaluasi. Analisis mendalam terhadap butir soal yang tidak valid diperlukan untuk memahami penyebab ketidakvalidan tersebut. Proses ini akan memberikan wawasan yang berguna bagi guru dan penyusun kurikulum dalam menyempurnakan desain soal, sehingga kualitas instrumen evaluasi di masa mendatang dapat lebih baik.

Validitas butir soal merupakan aspek yang dinamis dan dapat berubah seiring dengan perkembangan kurikulum serta kebutuhan pembelajaran. Oleh karena itu, penting bagi guru untuk melakukan pemantauan secara berkala terhadap kualitas instrumen evaluasi. Penyesuaian dan pembaruan instrumen secara rutin diperlukan agar tetap relevan dengan perubahan kurikulum dan memastikan instrumen tersebut memenuhi standar validitas yang diperlukan.

Secara konseptual, validitas instrumen terdiri atas tiga komponen utama, yaitu validitas isi, validitas konstruk, dan validitas kriteria. Validitas isi memastikan

bahwa butir soal mencakup materi yang sesuai dengan tujuan pembelajaran. Validitas konstruk mengacu pada sejauh mana instrumen mencerminkan aspek psikologis atau indikator pembelajaran yang akan diukur. Sedangkan validitas kriteria berfokus pada hubungan antara hasil tes dan kinerja siswa dalam situasi nyata atau prediktif.

Validitas empiris, yang merupakan salah satu jenis validitas tambahan, mengacu pada keakuratan instrumen berdasarkan pengalaman. Misalnya, validitas prediksi memungkinkan instrumen mengukur atau memprediksi hasil di masa depan, seperti keberhasilan siswa dalam studi. Contohnya, tes masuk universitas dirancang untuk memprediksi kemampuan siswa dalam menjalani studi di perguruan tinggi. Dengan membandingkan hasil tes dengan data historis, guru atau penyusun instrumen dapat mengevaluasi sejauh mana tes mencerminkan pencapaian siswa secara akurat dan valid (Sudrajat, 2022).

Dengan pemahaman ini, validitas instrumen tidak hanya berfungsi sebagai tolok ukur keberhasilan evaluasi dalam mengukur pencapaian siswa terhadap tujuan pembelajaran, tetapi juga menjadi landasan penting untuk melakukan perbaikan berkelanjutan pada sistem pembelajaran dan evaluasi di sekolah, sehingga instrumen yang digunakan tidak hanya relevan dengan kebutuhan kurikulum yang terus berkembang, tetapi juga mampu memberikan gambaran yang akurat tentang kemampuan siswa dan mendukung tercapainya proses pembelajaran yang lebih efektif dan berkualitas.

2. Uji Reliabilitas

Peneliti melaksanakan pengujian pada tes sumatif Tengah Semester Ganjil yang terdiri atas 50 soal pilihan ganda mata pelajaran Bahasa Arab di tiga kelas, yaitu XIIAK, XIITPM, dan XIITKR, dengan total jumlah siswa sebanyak 60 orang. Untuk mengukur reliabilitas instrumen penelitian, peneliti menggunakan perhitungan dengan rumus *Alpha Cronbach*. Suatu instrumen dianggap reliabel apabila nilai reliabilitas yang diperoleh sama dengan atau lebih besar dari 0,65. Berikut adalah hasil pengujian reliabilitas:

Tabel 3. Hasil Uji Reliabilitas Instrumen Soal

Reliability Statistics

Cronbach's Alpha	N of Items
,742	31

Nilai reliabilitas sebesar 0,742 yang diperoleh jauh melampaui ambang batas minimal reliabilitas sebesar 0,65. Hal ini menegaskan bahwa instrumen tes pilihan ganda memiliki tingkat konsistensi yang sangat tinggi, sehingga hasil evaluasi yang diperoleh dapat diandalkan dan akurat. Hasil uji reliabilitas ini memberikan dukungan kuat terhadap validitas instrumen penelitian dan

memperkuat kesimpulan yang dapat diambil dari analisis butir soal sebelumnya. Dengan reliabilitas yang tinggi, tes ini menunjukkan kemampuan untuk mengukur pemahaman siswa terhadap mata pelajaran Bahasa Arab secara konsisten. Selain itu, tingkat reliabilitas yang tinggi dapat meminimalkan kemungkinan kesalahan pengukuran, sehingga meningkatkan keakuratan hasil penelitian.

Keberhasilan dalam mendapatkan nilai *Cronbach's Alpha* sebesar 0,742 menegaskan bahwa instrumen penelitian ini dapat memberikan kontribusi signifikan dalam mengevaluasi pemahaman siswa terhadap materi Bahasa Arab di SMK Maarif NU Sunan Giri Driyorejo. Dalam konteks evaluasi pembelajaran, hasil ini dapat menjadi dasar untuk pembaruan dan penyesuaian instrumen di masa mendatang. Guru dapat menggunakan informasi ini untuk merancang strategi pengajaran yang lebih efektif, sementara sekolah dapat mempertimbangkan penerapan instrumen ini dalam ujian dan evaluasi lainnya. Dengan demikian, hasil reliabilitas yang baik tidak hanya mendukung penelitian ini tetapi juga memberikan dampak positif dalam meningkatkan kualitas sistem evaluasi pendidikan.

Pemahaman terhadap hasil analisis butir soal ini memiliki manfaat yang luas, baik bagi guru maupun pihak-pihak terkait dalam proses pengambilan keputusan di tingkat sekolah. Hasil penelitian ini dapat menjadi landasan untuk mengembangkan strategi evaluasi dan pengajaran yang lebih baik, serta memberikan kontribusi positif terhadap sistem evaluasi pendidikan secara keseluruhan. Oleh karena itu, temuan ini penting untuk didiskusikan dan dimanfaatkan secara kolaboratif di antara staf pengajar, penyusun kurikulum, dan pimpinan sekolah guna mencapai tujuan bersama dalam meningkatkan kualitas pendidikan.

Uji reliabilitas digunakan untuk mengukur konsistensi alat ukur, seperti kuesioner yang terdiri atas indikator-indikator berdasarkan variabel atau konstruk tertentu. Uji reliabilitas memastikan bahwa alat pengukur dapat menghasilkan hasil yang konsisten sepanjang percobaan atau pengukuran, sehingga dapat diandalkan. Instrumen dianggap reliabel jika memberikan hasil yang sama dalam berbagai situasi pengukuran. Hal ini menunjukkan tingkat stabilitas dan konsistensi yang diperlukan untuk menghasilkan evaluasi yang akurat.

Sebagaimana dijelaskan oleh Nursalam, reliabilitas mengacu pada kesamaan hasil pengukuran atau pengamatan yang dilakukan dalam periode waktu berbeda. Faktor alat dan metode yang digunakan dalam pengukuran sangat penting untuk memastikan kesamaan hasil yang konsisten. Reliabilitas kuesioner didefinisikan sebagai konsistensi jawaban responden terhadap pernyataan-pernyataan dalam kuesioner. Jika jawaban responden tetap konsisten dalam berbagai pengukuran, kuesioner dianggap reliabel. Dengan demikian, uji reliabilitas memastikan bahwa suatu tes memberikan hasil yang konsisten setiap kali digunakan (Riza Desima, 2013).

Reliabilitas juga terkait dengan kecocokan instrumen tes untuk mengukur variabel tertentu. Instrumen yang menghasilkan nilai yang sama pada pengukuran

yang sama dan dalam kondisi yang serupa disebut reliabel. Menurut Surapranata (2006), pengulangan pengukuran dengan instrumen yang sama diharapkan menghasilkan hasil yang serupa. Untuk menguji reliabilitas, berbagai metode dapat digunakan, seperti uji test-retest, uji ekuivalen, dan uji internal consistency. Teknik uji internal consistency, termasuk KR-20, KR-21, dan Alpha Cronbach split-half, digunakan untuk berbagai jenis instrumen sesuai kebutuhan (Slamet & Wahyuningsih, 2022).

3. Faktor yang Mempengaruhi Reliabilitas

Reliabilitas tes cenderung meningkat seiring dengan bertambahnya jumlah butir soal yang disertakan dalam instrumen. Hal ini disebabkan oleh kenyataan bahwa semakin banyak butir soal yang digunakan, semakin besar kemungkinan instrumen tersebut mampu mengukur dengan lebih akurat. Seperti yang diilustrasikan melalui rumus Spearman-Brown, tingkat reliabilitas akan meningkat sebagai fungsi dari panjang tes, menunjukkan hubungan positif antara jumlah butir soal dan konsistensi hasil pengukuran.

$$r_n = \frac{nr}{1 + (n-1)r}$$

r_n : indeks reliabilitas setelah ditambahkan soal

n : perkalian penambahan awal

r :indeks reliabilitas awal

Dalam menyusun instrumen tes, penting untuk mempertimbangkan jumlah butir soal yang tepat (Zakiyah dkk, 2024). Kontinuitas instrumen akan meningkat jika didasarkan pada pemahaman mendalam tentang variabel yang diukur. Meskipun penambahan butir soal dapat meningkatkan reliabilitas, bukan berarti semakin banyak soal akan selalu lebih baik. Oleh karena itu, perlu diperhatikan parameter serta batas ukuran variabel yang diuji agar instrumen tetap efektif dan efisien dalam mengukur kompetensi yang diinginkan.

Selain jumlah butir soal, variabilitas kelompok juga menjadi faktor yang memengaruhi reliabilitas instrumen. Variabilitas ini mengacu pada perbedaan nilai subjek dalam variabel yang diukur oleh tes. Misalnya, jika tes digunakan untuk mengukur kecakapan berbahasa Arab, maka variabilitas kelompok akan mencerminkan perbedaan tingkat kecakapan berbahasa Arab di antara peserta didik yang diuji. Semakin besar variasi dalam kelompok, semakin besar peluang instrumen tersebut menunjukkan hasil yang lebih akurat.

Koefisien reliabilitas dipengaruhi langsung oleh penyebaran nilai dalam kelompok yang diukur. Koefisien ini menunjukkan sejauh mana tes dapat secara konsisten menempatkan siswa dalam posisi yang sesuai di dalam kelompok mereka. Hal ini sejalan dengan pendapat Robert M. Thorndike, yang menyatakan bahwa "variabilitas kelompok yang diberikan tes" merupakan salah satu faktor utama yang memengaruhi reliabilitas suatu instrumen. Semakin besar penyebaran skor dalam

kelompok yang diuji, semakin tinggi koefisien reliabilitas yang dapat diperoleh (Setiyawan, 2014). Ebel juga mencatat bahwa kelompok dengan variasi kemampuan yang lebih besar cenderung memiliki koefisien reliabilitas yang lebih tinggi dibandingkan dengan kelompok yang lebih homogen. Dengan kata lain, semakin heterogen suatu kelompok, semakin besar kemungkinan tes memberikan hasil yang lebih reliabel.

Selain variabilitas kelompok, objektivitas dalam penskoran juga berperan penting dalam menentukan reliabilitas tes. Reliabilitas penilai mengacu pada tingkat kesepakatan antara berbagai penilai dalam memberikan skor pada suatu instrumen. Semakin tinggi kesesuaian antarpenilai, semakin kuat keandalan penilaian. Sebaliknya, jika terdapat perbedaan penilaian yang signifikan, maka reliabilitas pemberian skor akan menurun. Oleh karena itu, dalam penyusunan tes, perlu dipertimbangkan metode penskoran yang dapat mengurangi subjektivitas dan meningkatkan konsistensi hasil penilaian.

Banyak ujian akademik standar dan ujian bakat memiliki tingkat objektivitas yang tinggi, terutama dalam soal-soal pilihan ganda yang skornya tidak dipengaruhi oleh pendapat atau subjektivitas penilai. Objektivitas dalam situasi ini menunjukkan bahwa evaluasi dilakukan secara konsisten tanpa dipengaruhi faktor eksternal atau keputusan subjektif dari penilai. Karena skor tes objektif dapat diukur dengan cara yang seragam oleh berbagai evaluator, hasilnya memiliki tingkat objektivitas yang tinggi. Reliabilitas tes cenderung berkorelasi positif dengan tingkat objektivitasnya. Oleh karena itu, penggunaan tes objektif dapat meningkatkan reliabilitas instrumen dalam pengukuran.

Dalam pengujian koefisien reliabilitas tes standar, pemilihan metode atau teknik yang digunakan untuk menentukan besar koefisien reliabilitas sangatlah penting. Nilai koefisien reliabilitas yang diestimasi secara langsung dipengaruhi oleh metode yang digunakan. Salah satu metode yang umum digunakan adalah metode tes ulang (*test-retest method*). Jika interval waktu antara tes pertama dan tes ulang lebih singkat, respons siswa terhadap tes akan lebih stabil dan konsisten, sehingga koefisien reliabilitas yang diperoleh cenderung lebih tinggi dibandingkan dengan metode belah dua. Namun, jika interval waktu antara tes pertama dan tes ulang semakin panjang, koefisien reliabilitas cenderung lebih rendah. Hal ini disebabkan oleh kemungkinan perubahan dalam kemampuan siswa atau faktor lain yang memengaruhi respons mereka terhadap tes yang berbeda. Perubahan ini dapat berdampak negatif pada reliabilitas tes dan menyebabkan penurunan koefisien reliabilitas yang diestimasi.

Selain faktor metode yang digunakan, koefisien reliabilitas juga dipengaruhi oleh karakteristik kelompok yang diuji. Tingkat kemampuan individu yang diuji dapat memengaruhi akurasi pengukuran instrumen tes. Namun, tidak ada aturan pasti yang menjelaskan korelasi ini, karena hubungan tersebut sangat bergantung pada desain tes. Jika suatu tes terlalu mudah bagi kelompok tertentu sehingga hampir semua peserta dapat menjawab dengan benar, maka tes tersebut mungkin

kurang efektif dalam membedakan tingkat kemampuan individu dalam kelompok tersebut. Sebaliknya, jika tes terlalu sulit, peserta mungkin cenderung menebak jawaban, yang dapat mengurangi akurasi pengukuran. Oleh karena itu, guru harus mengevaluasi pembelajarannya (Huda, Baihaqi, & Fathoni, 2024), tingkat kesulitan tes harus disesuaikan agar dapat membedakan kemampuan individu secara efektif.

Reliabilitas juga dipengaruhi oleh homogenitas tes. Misalnya, ujian bahasa Arab dengan 100 butir soal untuk siswa kelas XII akan lebih akurat dalam mengukur kemampuan mereka dibandingkan dengan ujian yang mencakup berbagai tingkat pendidikan, seperti SMK secara keseluruhan. Prinsip yang sama berlaku untuk bidang studi lainnya, seperti matematika, yang memerlukan pendekatan lebih terstruktur dengan penekanan pada logika, aturan, dan keterampilan. Hal ini berbeda dengan mata pelajaran seperti sejarah, yang mungkin lebih terbuka terhadap interpretasi. Dengan demikian, struktur dan homogenitas tes menjadi faktor tambahan yang memengaruhi reliabilitas dalam pengukuran.

Secara keseluruhan, reliabilitas instrumen tes sangat dipengaruhi oleh berbagai faktor, seperti jumlah butir soal, variabilitas kelompok, metode pengujian, objektivitas penskoran, serta tingkat kesulitan dan homogenitas tes. Semua faktor ini harus dipertimbangkan secara matang agar instrumen yang digunakan dapat memberikan hasil yang akurat, konsisten, dan dapat diandalkan dalam mengukur kemampuan individu secara valid dan objektif. Sebagaimana amanat kurikulum merdeka, diharapkan siswa memiliki kesiapan yang baik dalam mengimplementasikannya (Huda, Baihaqi, Jundi, et al., 2024).

KESIMPULAN

Berdasarkan analisis validitas dan reliabilitas terhadap soal ujian sumatif Tengah Semester Ganjil mata pelajaran Bahasa Arab tahun ajaran 2024/2025 di SMK Ma’arif NU Sunan Giri Driyorejo, ditemukan bahwa soal pilihan ganda untuk kelas XII memiliki kualitas yang baik dan dapat digunakan sebagai instrumen evaluasi yang andal dalam jangka panjang. Temuan ini menegaskan pentingnya uji kualitas soal sebelum digunakan agar hasil evaluasi mencerminkan kemampuan siswa secara akurat. Oleh karena itu, guru disarankan untuk rutin melakukan uji coba butir soal serta mengikuti pelatihan atau workshop yang difasilitasi oleh sekolah guna meningkatkan keterampilan analisis soal. Untuk penelitian selanjutnya, disarankan dilakukan analisis soal pada jenjang kelas X dan XI guna melihat konsistensi kualitas instrumen, serta memperluas kajian pada aspek tingkat kesukaran, daya pembeda, dan efektivitas pengecoh. Penelitian kualitatif mengenai persepsi guru dan siswa terhadap soal yang telah divalidasi serta studi komparatif antar sekolah juga direkomendasikan untuk memperkaya pemahaman tentang implementasi evaluasi yang efektif dalam pembelajaran Bahasa Arab.

BIBLIOGRAFI

Anshari, M. I., Nasution, R., Irsyad, M., Alifa, A. Z., & Zuhriyah, I. A. (2024). Analisis Validitas dan Reliabilitas Butir Soal Sumatif Akhir Semester Ganjil Mata Pelajaran PAI. *Edukatif : Jurnal Ilmu Pendidikan*, *6*(1), 964–975. https://doi.org/10.31004/edukatif.v6i1.5931

Armini, N. K. (2024). Evaluasi metode penilaian perkembangan siswa dan pendidikan karakter dalam kurikulum merdeka pada sekolah dasar. *Metta: Jurnal Ilmu Multidisiplin*, *4*(1), 98–112.

Arslan, E. (2022). Validity and Reliability in Qualitative Research. *Pamukkale University Journal of Social Sciences Institute*. https://doi.org/10.30794/pausbed.1116878

Atmojo, I. R. W., Tiana, R., & Karsono, K. (2019). Profile of elementary school teachers' ability in analyzing higher order thinking science question. *Journal of Physics: Conference Series*, *1318*(1). https://doi.org/10.1088/1742-6596/1318/1/012140

Bariyyah, K., Hastini, R. P., & Wulan Sari, E. K. (2018). Konseling Realita untuk Meningkatkan Tanggung Jawab Belajar Siswa. *Konselor*, *7*(1), 1–8. https://doi.org/10.24036/02018718767-0-00

Bashooir, K., & Supahar, S. (2018). Validitas dan reliabilitas instrumen asesmen kinerja literasi sains pelajaran Fisika berbasis STEM. *Jurnal penelitian dan evaluasi pendidikan*, *22*(2), 219–230.

Defitasari, Widayanti, S., Nur Indah, P., & Andrian Syah, M. (2011). Analisis Preferensi Konsumen Terhadap Minuman Jamu Tradisional Di Kecamatan Gondang Kabupaten Nganjuk. *Jurnal Ilmiah Mahasiswa Agroinfo Galuh*, *9*(2), 513–526.

Huda, H., Baihaqi, M., & Fathoni, F. (2024). Persepsi Dosen dalam Menerapkan Kurikulum Inovatif. *Jurnal Alfazuna : Jurnal Pembelajaran Bahasa Arab dan Kebahasaaraban*, *8*(2), 227–238. https://doi.org/10.15642/alfazuna.v8i2.2944

Huda, H., Baihaqi, M., Jundi, A., Mubarak, A., & Zaenuri, M. (2024). Perceptions of Arabic Language Education Students on the Implementation of an Innovative Curriculum. *Jurnal At Ta'rib UIN Palangkaraya*, *12*(2), 439–454.

Jundi, A., Baihaqi, M., & Zaenuri, M. (2025). Identification And Correction Of Pseudowords In Ilman Wa Ruhan Textbooks To Reduce Meaning Errors. *Ijaz Arabi: Journal Of Arabic Learning*, *8*(1), 169.

M. Baihaqi, Muflihah, & Ramadhani, T. F. (2023). The Effectiveness of Using Plotagon Story Media to Improve Listening Skills for Non-Arabic Speakers. *Journal of Arabic Language Studies and Teaching*, *3*(2), 141–150. https://doi.org/10.15642/jalsat.2023.3.2.141-150

Pahri, P. (2021). The Implementation of Total Physical Response (TPR) Method in Improving Arabic Speaking Skills. *Tanwir Arabiyyah: Arabic As Foreign Language Journal*, *1*(2), 63–72. https://doi.org/10.31869/aflj.v1i2.2872

Riza Desima. (2013). Tingkat Stres Kerja Perawat dengan Perilaku Caring Perawat. *Jurnal Keperawatan*, *4*(1), 43–55.

Rosmana, P. S., Iskandar, S., Rahma, A. R., Maria, S., Supriatna, S., & Wahyuningtyas, T. (2024). Efektivitas Penggunaan Media Pembelajaran Digital Pada Hasil Belajar Siswa Kelas 5 SDN 6 Nagrikaler. *Jurnal Sinektik*,

6(1), 10–17. https://doi.org/10.33061/js.v6i1.8205

Rusnawati, M. A. (2020). Komponen-Komponen Dalam Operasional Pendidikan. *Jurnal Azkia: Jurnal Aktualisasi Pendidikan Islam*, *15*(2).

Setiyawan, A. (2014). Faktor- faktor yang Mempengaruhi Reliabilitas Tes. *Jurnal An Nûr*, *VI*(2), 341–354.

Slamet, R., & Wahyuningsih, S. (2022). Validitas dan reliabilitas terhadap instrumen kepuasan kerja. *Aliansi: Jurnal Manajemen dan Bisnis*, *17*(2).

Wulandari, A. R., & Pramusinto, H. (2020). Analisis Kualitas Butir Soal Ujian Akhir Semester Ganjil Mata Pelajaran Otomatisasi Tata Kelola Sarana dan Prasarana Kelas XI Otomatisasi dan Tata Kelola Perkantoran. *Economic Education Analysis Journal*, *9*(2), 366–378. https://doi.org/10.15294/eeaj.v9i2.39000

Yusup, F. (2018). Uji Validitas dan Reliabilitas Instrumen Penelitian Kuantitatif. *Jurnal Tarbiyah: Jurnal Ilmiah Kependidikan*, *13*(1), 53–59. https://doi.org/10.21831/jorpres.v13i1.12884

Zakiyah dkk. (2024). *Evaluation of the Feasibility and Reliability of Arabic Multiple Choice Tests in Higher Education*. *11*(2), 164–192.