TA'LIMUNA
JURNAL PENDIDIKAN ISLAM
P-ISSN 2085-2975 E-ISSN 2622-9889

EVALUASI VALIDITAS, RELIABILITAS, DAN DAYA PEMBEDA INSTRUMEN UJIAN PAI KELAS XII MELALUI ANALISIS BUTIR SOAL

Evaluation of the Validity, Reliability, and Discriminating Power of Islamic Education Exam Items for Grade XII through Item Analysis

Fahriana Nurrisa
UIN Antasari Banjarmasin, Indonesia
fahriananurrissa@gmail.com

Muhamad Sabirin
UIN Antasari Banjarmasin, Indonesia
m.sabirin@uin-antasari.ac.id

Nuril Huda
UIN Antasari Banjarmasin, Indonesia
nurilhuda@uin-antasari.ac.id

Keywords
Analisis Butir Soal
Daya Pembeda Soal
Evaluasi Pembelajaran
Reliabilitas
Validitas

Abstract

This study aims to evaluate the quality of a school exam instrument for the Islamic Religious Education (PAI) subject in Grade XII by analyzing the validity, reliability, and discriminating power of each test item. The study employed a descriptive quantitative approach. The research subjects consisted of 15 Grade XII students from a state vocational school (SMK) in South Kalimantan who took the PAI school examination. The instrument analyzed comprised 15 multiple-choice questions. Data were obtained through documentation, including test sheets, answer keys, and student response sheets. The item analysis was conducted based on Classical Test Theory (CTT). The results showed that only 2 out of 15 items met the validity criteria, with correlation values exceeding the critical r-table value (0.5324). The instrument's reliability, calculated using the KR-20 formula, was 0.0227, indicating a very low level of internal consistency. Meanwhile, the discriminating power analysis revealed that most items were categorized as poor or rejected, although a few were classified as good to excellent. These

findings suggest that the test instrument is not yet suitable for use in learning evaluation without significant revision and improvement of most test items.
Penelitian ini bertujuan mengevaluasi kualitas instrumen ujian sekolah mata pelajaran Pendidikan Agama Islam (PAI) kelas XII melalui analisis validitas, reliabilitas, dan daya pembeda butir soal. Penelitian ini menggunakan pendekatan kuantitatif deskriptif. Subjek penelitian terdiri dari 15 peserta didik kelas XII di salah satu SMK Negeri di Kalimantan Selatan yang mengikuti ujian sekolah PAI. Instrumen yang dianalisis terdiri dari 15 butir soal pilihan ganda. Data diperoleh melalui dokumentasi berupa lembar soal, kunci jawaban, dan lembar jawaban siswa. Teknik analisis yang digunakan adalah analisis butir soal berdasarkan teori tes klasik. Hasil penelitian menunjukkan bahwa hanya 2 dari 15 soal yang memenuhi kriteria validitas dengan nilai korelasi di atas r tabel (0,5324). Reliabilitas instrumen berdasarkan perhitungan menggunakan KR-20 adalah sebesar 0,0227 dan termasuk kategori sangat rendah. Sementara itu, daya pembeda menunjukkan bahwa sebagian besar soal berada dalam kategori ditolak, meskipun terdapat beberapa soal dengan daya pembeda baik hingga sangat baik. Temuan ini menunjukkan bahwa instrumen secara keseluruhan belum layak digunakan dalam evaluasi pembelajaran tanpa melalui perbaikan dan revisi pada sebagian besar butir soal.

Received	Revised	Accepted	Available Online
1 Agustus 2025	22 September 2024	31 Agustus 2025	30 September 2025

A. PENDAHULUAN

Pendidikan merupakan proses yang terencana untuk membentuk manusia yang beriman, bertakwa, berakhlak mulia, serta memiliki pengetahuan, keterampilan, dan sikap yang bermanfaat bagi dirinya dan masyarakat. Dalam proses pendidikan, pembelajaran menjadi inti dari kegiatan yang menjembatani antara kompetensi dasar yang dirancang dalam kurikulum dengan kemampuan nyata yang dimiliki peserta didik. Agar proses pembelajaran berjalan efektif dan terukur, maka perlu dilakukan evaluasi secara sistematis, objektif, dan berkesinambungan. Evaluasi pembelajaran adalah upaya untuk mengetahui sejauh mana peserta didik telah mencapai tujuan pembelajaran yang telah ditetapkan (Muryadi, 2017).

Salah satu bentuk evaluasi yang paling umum diterapkan di sekolah adalah ujian tertulis. Ujian tersebut dirancang untuk mengukur kemampuan kognitif peserta didik dalam memahami materi yang telah diajarkan. Namun, efektivitas evaluasi sangat tergantung pada kualitas instrumen yang digunakan, khususnya jika instrumen tersebut berupa soal pilihan ganda. Soal yang baik harus dapat mengukur

kompetensi yang relevan, memberikan hasil yang konsisten, dan mampu membedakan antara peserta didik yang memahami dan tidak memahami materi.

Dalam konteks mata pelajaran Pendidikan Agama Islam (PAI), evaluasi memiliki peran yang sangat penting tidak hanya dalam mengukur aspek kognitif, tetapi juga untuk menilai sejauh mana nilai-nilai ajaran Islam dipahami dan diinternalisasi oleh peserta didik. Oleh karena itu, kualitas instrumen ujian PAI perlu mendapat perhatian khusus. Instrumen yang tidak valid, tidak reliabel, atau memiliki daya pembeda yang rendah dapat mengakibatkan hasil evaluasi yang bias, sehingga tidak mencerminkan kondisi sebenarnya dari kemampuan peserta didik (Devi et al., 2022). Sayangnya, dalam praktik di lapangan, masih banyak guru yang menyusun soal ujian tanpa melalui proses analisis empirik. Penyusunan soal dilakukan secara konvensional, berdasarkan pengalaman dan kebiasaan, tanpa uji coba dan tanpa mengetahui apakah soal tersebut benar-benar layak digunakan. Padahal, dalam dunia evaluasi pendidikan, terdapat metode analisis butir soal yang dapat digunakan untuk mengevaluasi kualitas masing-masing item dalam suatu instrumen.

Analisis butir soal merupakan metode kuantitatif yang digunakan untuk mengetahui karakteristik tiap butir soal, termasuk validitas (keterkaitan soal dengan skor total), reliabilitas (tingkat konsistensi soal), dan daya pembeda (kemampuan soal membedakan peserta didik berkemampuan tinggi dan rendah). Dengan menerapkan analisis ini, guru atau peneliti dapat mengidentifikasi soal-soal yang perlu diperbaiki, direvisi, atau dihapus, sehingga instrumen yang digunakan benar-benar mencerminkan alat evaluasi yang efektif dan berkualitas. Berdasarkan latar belakang tersebut, penelitian ini dilakukan untuk mengevaluasi kualitas instrumen ujian sekolah mata pelajaran PAI kelas XII melalui analisis butir soal. Adapun fokus dari penelitian ini adalah pada tiga aspek utama, yaitu validitas, reliabilitas, dan daya pembeda butir soal pilihan ganda yang digunakan dalam ujian sekolah.

B. METODE

Penelitian ini menggunakan pendekatan kuantitatif deskriptif adalah metode penelitian yang bertujuan untuk menggambarkan atau mendeskripsikan suatu fenomena, kondisi, atau karakteristik tertentu secara sistematis dan didukung oleh data numerik atau angka. Pendekatan ini mengumpulkan data yang bersifat kuantitatif (angka-angka, statistik), lalu dianalisis untuk memahami gambaran umum tentang objek yang diteliti (Unaradjan, 2019).

Penelitian kuantitatif deskriptif ini bertujuan mengevaluasi kualitas instrumen ujian sekolah mata pelajaran PAI kelas XII. Fokus evaluasi diarahkan pada tiga aspek utama, yaitu validitas, reliabilitas, dan daya pembeda butir soal pilihan ganda. Penelitian ini tidak bertujuan menguji hubungan antar variabel, melainkan untuk mendeskripsikan karakteristik butir soal berdasarkan data empiris dari hasil ujian peserta didik. Penelitian dilaksanakan di salah satu SMK Negeri di Kalimantan Selatan. Subjek penelitian terdiri

dari 15 peserta didik kelas XII yang telah mengikuti ujian sekolah PAI secara lengkap. Instrumen yang dianalisis terdiri dari 15 butir soal pilihan ganda, yang masing-masing memiliki lima opsi jawaban (A, B, C, D, dan E), dengan satu jawaban benar sebagai kunci.

Teknik pengambilan sampel menggunakan metode *purposive sampling*, dengan kriteria siswa yang menyelesaikan seluruh soal tanpa ada jawaban kosong. Sedangkan pengumpulan data dilakukan melalui teknik dokumentasi, yaitu dengan mengumpulkan naskah soal, kunci jawaban, dan lembar jawaban siswa. Teknik analisis data dilakukan berdasarkan pendekatan teori tes klasik (*Classical Test Theory*), dengan tahapan sebagai berikut:

1) Validitas dihitung menggunakan teknik korelasi *Pearson Product-Moment* antara skor tiap butir soal dan skor total siswa. Butir soal dinyatakan valid apabila nilai koefisien korelasi ≥ 0,5324 (nilai r tabel untuk n = 15 pada taraf signifikansi 0,05).
2) Reliabilitas dianalisis menggunakan rumus Kuder-Richardson Formula 20 (KR-20), yang sesuai untuk soal pilihan ganda dengan *skoring dikotomus* (benar–salah). Instrumen dikatakan reliabel apabila nilai KR-20 ≥ 0,60.
3) Daya Pembeda dihitung berdasarkan selisih proporsi siswa kelompok atas dan kelompok bawah yang menjawab benar. Kelompok atas dan bawah masing-masing terdiri dari 7–8 siswa dengan skor tertinggi dan terendah. Daya pembeda ≥ 0,30 dikategorikan baik, sedangkan < 0,20 menunjukkan soal perlu direvisi atau diganti.

Seluruh proses perhitungan dilakukan dengan bantuan perangkat lunak Microsoft Excel, dan hasil analisis disajikan secara deskriptif untuk memberikan gambaran mengenai kualitas masing-masing butir soal sebagai alat evaluasi pembelajaran.

C. HASIL DAN PEMBAHASAN.

1. Hasil Penelitian

Sehubungan dengan analisis butir soal dengan menghitung validitas, reliabilitas, dan daya pembeda. Tes hasil belajar adalah 15 butir soal tes obyektif pilihan ganda dengan lima pilihan, jawaban 15 orang siswa dilaporkan hasilnya sebagai berikut:

Tabel 1
Laporan Hasil Belajar Siswa Kelas XII

No	Siswa	Butir Soal															Jumlah
		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	
1	Aulia Ramadhan	1	1	1	1	0	1	1	1	1	0	1	0	1	1	1	12
2	Muhammad Hairul wapi	1	0	1	1	0	1	0	0	1	0	1	0	1	1	0	8
3	Azkia Azizah	1	0	1	1	1	1	1	1	1	0	1	0	1	1	1	12
4	Muhammad Julianur	1	0	1	0	0	1	1	0	1	1	0	1	1	1	0	9
5	NurJanah	1	0	0	1	1	0	1	1	1	0	1	0	1	1	0	9
6	Amilia sari	1	1	1	1	1	1	1	1	1	1	1	0	1	1	0	13
7	Muhammad Ripani	0	1	1	1	1	1	1	1	0	0	1	0	0	1	0	9
8	Akhmad adam	1	0	1	1	0	1	1	1	1	0	1	0	1	1	0	10
9	Dimas Al Qhoiri	1	0	1	1	0	0	0	1	1	1	1	1	1	1	0	10
10	Miranda santika nurul juliani	1	0	1	1	1	1	1	1	1	0	0	0	1	1	1	11
11	Siti Maifatus As-Sa'adah	1	0	1	1	0	0	1	1	1	1	1	0	1	1	0	10
12	Muhammad Dhani Saputra	1	1	1	1	0	1	0	1	1	0	1	0	1	1	1	11
13	Muhammad Asdy Ramadhan	1	0	1	1	0	1	1	0	1	0	1	0	1	1	0	9
14	Siti Maifatus As-Sa'adah	1	0	1	1	0	0	1	1	1	1	1	0	1	1	0	10
/ABAN BENAR		13	4	13	13	5	10	11	11	13	5	12	2	13	14	4	

a. Hasil Uji Validitas

Untuk mengetahui sejauh mana setiap butir soal dapat mengukur kompetensi yang diharapkan, dilakukan uji validitas secara empiris melalui korelasi antara skor tiap butir soal dan skor total peserta didik. Hasil analisis validitas disajikan pada tabel berikut ini:

Tabel 2
Hasil Uji Validitas

VALIDITAS															
CORRELATION	0.2455	0.4774	0.2455	0.2455	0.3182	0.2140	0.2084	0.5890	0.2455	0.1009	0.0638	-0.2125	0.2455	#DIV/0!	0.5926
R TABEL	0.5324	0.5324	0.5324	0.5324	0.5324	0.5324	0.5324	0.5324	0.5324	0.5324	0.5324	0.5324	0.5324	0.5324	0.5324
KEPUTUSAN	Tidak Valid	Tidak Valid	Tidak Valid	Tidak Valid	Tidak Valid	Tidak Valid	Tidak Valid	Valid	Tidak Valid	Tidak Valid	Tidak Valid	Tidak Valid	Tidak Valid	#DIV/0!	Valid

Validitas merupakan aspek penting dalam penyusunan instrumen evaluasi, karena berhubungan langsung dengan ketepatan soal dalam mengukur kompetensi yang dituju. Dalam penelitian ini, validitas dianalisis secara empiris dengan menggunakan teknik korelasi *Pearson Product Moment* antara skor setiap butir soal dan skor total peserta didik. Penentuan valid atau tidaknya suatu butir soal dilakukan dengan cara membandingkan nilai r hitung dengan r tabel. Dengan jumlah responden sebanyak 15 siswa, maka r tabel pada taraf signifikansi 0,05 (df = 13) adalah sebesar 0,5324. Artinya, suatu soal dikatakan valid apabila nilai r hitungnya sama dengan atau lebih besar dari 0,5324.

Berdasarkan hasil perhitungan, dari total 15 butir soal yang dianalisis, hanya 2 butir soal, yaitu soal nomor 8 dan 15, yang memiliki nilai korelasi >0,5324, sehingga dikategorikan valid. Sementara itu, 13 butir soal lainnya memiliki nilai r hitung di bawah batas kritis tersebut, sehingga dinyatakan tidak valid.Temuan ini menunjukkan bahwa sebagian besar soal dalam instrumen ujian belum memenuhi kriteria validitas yang diharapkan. Soal-soal yang tidak valid kemungkinan besar disebabkan oleh

ketidaksesuaian antara redaksi soal dan indikator pembelajaran, bahasa soal yang ambigu, atau pengecoh (distraktor) yang tidak berfungsi dengan baik.

Menurut Arikunto, soal yang tidak valid berpotensi menimbulkan hasil penilaian yang bias dan tidak mencerminkan kemampuan peserta didik secara objektif. Oleh karena itu, soal-soal yang tidak memenuhi kriteria validitas perlu direvisi atau diganti, dan sebaiknya diuji kembali sebelum digunakan dalam evaluasi sumatif. Hasil ini juga menekankan pentingnya penguasaan guru terhadap teknik analisis butir soal sebagai bagian dari kompetensi profesional dalam merancang dan mengevaluasi instrumen penilaian. Tanpa uji validitas, kualitas instrumen akan sulit dijamin, dan hasil penilaian menjadi kurang dapat dipercaya.

b. Hasil Uji Reliabilitas

Setelah validitas dianalisis, langkah selanjutnya adalah mengukur konsistensi atau kestabilan instrumen secara keseluruhan melalui uji reliabilitas dengan rumus Kuder-Richardson 20 (KR-20). Hasil perhitungan reliabilitas disajikan dalam tabel berikut ini:

Tabel 3
Hasil Uji Reliabilitas

RELIABILITAS															
N	14														
n-1	13														
P	0.929	0.286	0.929	0.929	0.357	0.714	0.786	0.786	0.929	0.357	0.857	0.143	0.929	1	0.286
Q	0.071	0.714	0.071	0.071	0.643	0.286	0.214	0.214	0.071	0.643	0.143	0.857	0.071	0	0.714
PQ	0.066	0.204	0.066	0.066	0.230	0.204	0.168	0.168	0.066	0.230	0.122	0.122	0.066	0	0.204
SPQ	1.985														
VARIANT SCORE	2.0274725														
KR 20	0.0227225														
KEPUTUSAN	TIDAK RELIABEL														

Reliabilitas merupakan indikator penting dalam evaluasi instrumen tes, karena menyangkut konsistensi atau kestabilan hasil yang diperoleh dari suatu alat ukur. Dalam konteks evaluasi pembelajaran, reliabilitas menjadi dasar untuk memastikan bahwa nilai atau hasil yang diperoleh peserta didik benar-benar mencerminkan kemampuan mereka, dan bukan hasil dari ketidaktepatan alat ukur (Hidayat, 2021b). Dalam penelitian ini, reliabilitas instrumen dihitung menggunakan rumus Kuder-Richardson Formula 20 (KR-20), yang dirancang khusus untuk mengukur reliabilitas instrumen dengan soal pilihan ganda berdikotomi (skor 1 untuk benar dan 0 untuk salah). KR-20 dianggap efektif dalam menganalisis tes dengan tipe soal yang seragam dan jumlah responden terbatas. Hasil perhitungan menunjukkan bahwa nilai KR-20 adalah 0,0227.

Berdasarkan klasifikasi Sudijono (2015), nilai reliabilitas:

1) ≥ 0,90 dikategorikan sangat tinggi,

2) 0,70–0,89 tinggi,
3) 0,50–0,69 sedang,
4) 0,30–0,49 rendah,
5) dan < 0,30 sangat rendah.

Dengan demikian, nilai 0,0227 tergolong sangat rendah, yang berarti bahwa instrumen ujian tersebut tidak reliabel. Artinya, jika instrumen yang sama digunakan kembali kepada kelompok siswa yang serupa, hasilnya berpotensi sangat berbeda dan tidak konsisten. Nilai reliabilitas yang sangat rendah ini menunjukkan adanya masalah serius dalam kualitas instrumen secara keseluruhan. Salah satu penyebab utamanya adalah validitas butir soal yang rendah, karena sebanyak 13 dari 15 soal dinyatakan tidak valid.

Dalam teori tes klasik, validitas dan reliabilitas saling berkaitan. Soal yang tidak valid berarti tidak mengukur apa yang seharusnya diukur, dan ketika banyak soal tidak valid, maka konsistensi total tes secara keseluruhan pun terganggu, yang berdampak langsung pada reliabilitas. Selain itu, daya pembeda yang juga sebagian besar berada pada kategori "ditolak" atau bahkan negatif, turut memperburuk nilai reliabilitas. Daya pembeda yang rendah menyebabkan perbedaan antar peserta didik tidak tertangkap secara efektif oleh soal, sehingga nilai total siswa menjadi tidak representatif.

Faktor lain yang mungkin mempengaruhi rendahnya reliabilitas dalam penelitian ini adalah:

1) Jumlah butir soal yang relatif sedikit (hanya 15),
2) Jumlah responden terbatas (15 peserta),
3) Redaksi soal yang kurang jelas atau menimbulkan penafsiran ganda,
4) Susunan soal yang tidak sistematis atau acak

Menurut (Arikunto, 2021), instrumen yang reliabel akan memberikan hasil yang sama meskipun diberikan dalam kondisi atau waktu yang berbeda, selama variabel dan situasi tetap serupa. Maka, nilai reliabilitas yang rendah mengindikasikan bahwa hasil evaluasi dengan instrumen ini tidak dapat diandalkan, baik sebagai bahan laporan nilai maupun untuk pengambilan keputusan pembelajaran. Dengan hasil ini, dapat disimpulkan bahwa instrumen ujian sekolah PAI yang dianalisis belum layak digunakan dalam penilaian sumatif, karena ketidakstabilan hasil yang dihasilkan dari penggunaannya. Maka dari itu, guru perlu memperbaiki terlebih dahulu kualitas individual setiap soal melalui uji validitas dan daya pembeda sebelum menguji ulang reliabilitas instrumen secara menyeluruh.

c. Hasil Uji Daya Pembeda

Analisis berikutnya adalah daya pembeda, yang bertujuan untuk mengetahui kemampuan butir soal dalam membedakan peserta didik yang berkemampuan tinggi dan rendah. Hasil perhitungan daya pembeda untuk setiap butir soal disajikan pada tabel berikut ini:

Tabel 4
Hasil Uji Daya Pembeda

No	Siswa	Butir Soal															Jumlah
		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	
1	Amilia sari	1	1	1	1	1	1	1	1	1	1	1	0	1	1	0	13
2	Aulia Ramadhan	1	1	1	1	0	1	1	1	1	0	1	0	1	1	1	12
3	Azkia Azizah	1	0	1	1	1	1	1	1	1	0	1	0	1	1	1	12
4	Miranda santika nurul juliani	1	0	1	1	1	1	1	1	1	0	0	0	1	1	1	11
5	Muhammad Dhani Saputra	1	1	1	1	0	1	0	1	1	0	1	0	1	1	1	11
K.ATAS	BA/JA	1	0.6	1	1	0.6	1	0.8	1	1	0.2	0.8	0	1	1	0.8	
6	Akhmad adam	1	0	1	1	0	1	1	1	1	0	1	0	1	1	0	10
7	Dimas Al Qhoiri	1	0	1	1	0	0	0	1	1	1	1	1	1	1	0	10
8	Siti Maifatus As-Sa'adah	1	0	1	1	0	0	1	1	1	1	1	0	1	1	0	10
9	Siti Maifatus As-Sa'adah	1	0	1	1	0	0	1	1	1	1	1	0	1	1	0	10
10	Muhammad Julianur	1	0	1	0	0	1	1	0	1	1	0	1	1	1	0	9
11	NurJanah	1	0	0	1	1	0	1	1	1	0	1	0	1	1	0	9
12	Muhammad Ripani	0	1	1	1	1	1	1	1	0	0	1	0	0	1	0	9
13	Muhammad Asdy Ramadhan	1	0	1	1	0	1	1	0	1	0	1	0	1	1	0	9
14	Muhammad Hairul wapi	1	0	1	1	0	1	0	0	1	0	1	0	1	1	0	8
K.BAWAH	BB/JB	0.9	0.1	0.9	0.9	0.2	0.6	0.8	0.7	0.9	0.4	0.9	0.2	0.9	1	0	
DAYA PEMBEDA		0.1	0.5	0.1	0.1	0.4	0.4	0.0	0.3	0.1	-0.2	-0.1	-0.2	0.1	0.0	0.8	
KEPUTUSAN		Ditolak	Sangat Bagus	Ditolak	Ditolak	Bagus	Sangat Bagus	Ditolak	Bagus	Ditolak	Ditolak	Ditolak	Ditolak	Ditolak	Ditolak	Sangat Bagus	

Daya pembeda adalah salah satu indikator penting dalam menentukan kualitas sebuah butir soal. Daya pembeda berfungsi untuk menunjukkan sejauh mana sebuah soal mampu membedakan antara peserta didik yang menguasai materi (kelompok atas) dan yang tidak menguasai materi (kelompok bawah). Semakin besar daya pembeda suatu soal, semakin baik kualitas soal tersebut karena mampu secara tajam mengklasifikasikan peserta didik berdasarkan tingkat kemampuannya.

Dalam penelitian ini, daya pembeda dihitung berdasarkan proporsi siswa kelompok atas dan bawah yang menjawab soal dengan benar. Kelompok atas terdiri dari siswa dengan nilai tertinggi, sedangkan kelompok bawah terdiri dari siswa dengan nilai terendah, yang masing-masing diambil dari 27% atau separuh responden dalam jumlah kecil. Metode ini sesuai dengan pendekatan teori tes klasik dalam menganalisis butir soal secara empirik. Hasil perhitungan menunjukkan bahwa dari 15 butir soal:

1) 3 soal memiliki daya pembeda sangat baik (nilai ≥ 0,40), yaitu soal nomor 2, 6, dan 15.
2) 2 soal memiliki daya pembeda baik (0,30 ≤ nilai < 0,40), yaitu soal nomor 5 dan 8.
3) 10 soal lainnya, termasuk soal nomor 1, 3, 4, 7, 9, 10, 11, 12, 13, dan 14, tergolong dalam kategori "ditolak" karena memiliki nilai < 0,20, bahkan beberapa di antaranya bernilai negatif.

Soal-soal dengan daya pembeda negatif, seperti soal nomor 11, 12, dan 13, merupakan tanda adanya permasalahan serius dalam kualitas soal. Nilai negatif menunjukkan bahwa siswa berkemampuan rendah justru lebih banyak menjawab benar dibanding siswa berkemampuan tinggi. Ini dapat terjadi karena, kunci jawaban salah (tidak sesuai dengan logika atau isi soal), distraktor tidak berfungsi, sehingga siswa asal memilih jawaban dan kebetulan benar, redaksi soal tidak jelas, ambigu, atau menjebak siswa berkemampuan tinggi untuk memilih jawaban yang salah.

Soal dengan daya pembeda < 0,20 sebaiknya dihilangkan atau direvisi secara total, karena soal seperti ini tidak mampu berfungsi sebagai alat seleksi atau penilaian (Qadir et al., 2024). Dalam proses evaluasi pendidikan, penting bagi guru untuk dapat membedakan siswa berdasarkan tingkat pemahamannya. Soal yang tidak mampu melakukan hal ini akan menurunkan kualitas keseluruhan dari instrumen evaluasi, dan berpotensi menghasilkan kesimpulan yang keliru tentang kemampuan siswa. Lebih lanjut, Arikunto menegaskan bahwa soal dengan daya pembeda tinggi merupakan salah satu ciri utama soal yang baik, karena memberikan informasi yang tajam tentang distribusi kemampuan siswa. Sebaliknya, soal dengan daya pembeda rendah atau negatif perlu dihindari karena dapat mengaburkan hasil evaluasi.

Temuan ini menunjukkan bahwa hanya sepertiga dari jumlah soal yang mampu berfungsi optimal, sedangkan mayoritas soal perlu diperbaiki, baik dari segi redaksional, konten, maupun struktur pilihan jawaban. Kondisi ini menunjukkan bahwa guru atau penyusun soal perlu memahami teknik analisis butir soal dan menggunakannya secara berkala sebagai alat bantu untuk meningkatkan kualitas asesmen. Dengan demikian, penting bagi sekolah untuk membekali guru dengan pelatihan tentang analisis daya pembeda dan teknik penyusunan soal yang baik. Jika instrumen evaluasi disusun dan dianalisis dengan cermat, maka hasil penilaian yang diperoleh dapat lebih adil, objektif, dan bermanfaat untuk pengambilan keputusan pendidikan yang lebih tepat.

2. Pembahasan

a. Evaluasi Pembelajaran

Grondlund dan Linn, mendefinisikan evaluasi pembelajaran adalah suatu proses mengumpulkan, menganalisis dan menginterpretasi informasi secara sistematik untuk menetapkan sejauh mana ketercapaian tujuan pembelajaran (Devi et al., 2022). Tujuan evaluasi pembelajaran adalah untuk menghimpun informasi yang dijadikan dasar untuk mengetahui taraf kemajuan, perkembangan, dan pencapaian belajar siswa, serta keefektifan pengajaran guru.

Evaluasi pembelajaran mencakup kegiatan pengukuran dan penilaian. Ditinjau dari tujuannya, evaluasi pembelajaran dibedakan atas evaluasi diagnostik, selektif, penempatan, formatif dan sumatif. Ditinjau ari sasarannya, evaluasi pembelajaran dapat dibedakan atas evaluasi konteks, input, proses, hasil dan *outcom*. Proses evaluasi melalui tiga tahap: perencanaan, pelaksanaan, pengolahan dan pelaporan.

Pengertian evaluasi pembelajaran adalah proses untuk menentukan nilai pembelajaran yang dilaksanakan, dengan melalui kegiatan pengukuran dan penilaian pembelajaran. Pengukuran yang dimaksud adalah proses membandingkan tingkat keberhasilan pembelajaran dengan ukuran keberhasilan pembelajaran yang telah ditentukan secara kuantitatif. Kemampuan pembelajar dalam menyampaikan materi kepada pembelajar dan bagi pembelajar sebagai penjajagan seberapa banyak materi yang mampu mereka serap selama proses pembelajaran (Ambiyar & D, 2019). Dari hasil tes, pembelajar/penyusun

silabus dapat mengubah/memperbaiki silabus, metode, dan media. Tes merupakan pengumpul informasi.

Tujuan evaluasi ada dua, yaitu secara umum dan secara khusus. Secara umum tujuan evaluasi pembelajaran adalah: *Pertama*, Untuk mengetahui keefektifan dan efisiensi sistem pembelajaran, baik yang menyangkut tentang tujuan, materi, metode, media, sumber belajar, lingkungan maupun sistem penilaian itu sendiri. *Kedua*, Untuk menghimpun bahan keterangan (data) yang dijadikan sebagai bukti mengenai tarap kemajuan anak didik dalam mengalami proses pendidikan selama jangka waktu tertentu.

Adapun tujuan khusus evaluasi pembelajaran adalah: *Pertama*, *Keeping track*. Yaitu untuk menelusuri dan melacak proses belajar peserta didik sesuai dengan rencana pelaksanaan pembelaj aran yang telah ditetapkan. Untuk itu, guru harus mengumpulkan data dan informasi dalam kurun waktu tertentu melalui berbagai jenis dan teknik penilaian untuk memperoleh gambaran tentang pencapaian kemajuan belajar peserta didik. *Kedua*, *Checking-up*. Yaitu untuk mengecek ketercapaian kemampuan peserta didik dalam proses pembelajaran dan kekurangan-kekurangan peserta didik selama mengikuti proses pembelajaran. Dengan kata lain, guru perlu melakukan penilaian untuk mengetahui bagian mana dari materi yang sudah dikuasai peserta didik dan bagian mana dari materi yang belum dikuasai.

Ketiga, *Finding-out*. Yaitu untuk mencari, menemukan dan mendeteksi kekurangan kesalahan atau kelemahan peserta didik dalam proses pembelajaran, sehingga guru dapat dengan cepat mencari alternatif solusinya. *Keempat*, *Summing-up*. Yaitu untuk menyimpulkan tingkat penguasaan peserta didik terhadap kompetensi yang telah ditetapkan. Hasil penyimpulan ini dapat digunakan guru untuk menyusun laporan kemajuan belajar ke berbagai pihak yang berkepentingan.

b. Analisis Butir Soal

Menganalisis butir soal merupakan kegiatan yang harus dilakukan guru untuk meningkatkan mutu soal yang telah ditulis. Kegiatan ini merupakan proses pengumpulan, peringkasan, dan penggunaan informasi dari jawaban siswa untuk membuat keputusan tentang setiap penilaian. Tujuan penelaahan kualitas butir soal, menurut Aiken, memiliki tiga tujuan antara lain: Untuk mengkaji dan menelaah setiap butir soal agar diperoleh soal yang bermutu sebelum soal digunakan, untuk membantu meningkatkan tes melalui revisi atau membuang soal yang tidak efektif, untuk mengetahui informasi diagnostik pada siswa apakah mereka sudah/belum memahami materi yang telah diajarkan. Untuk hal itu, (Qadir et al., 2024) menegaskan bahwa tujuan utama dari analisis butir soal dalam sebuah tes yang dibuat guru adalah untuk mengidentifikasi kekurangan-kekurangan dalam tes atau dalam pembelajaran.

Berdasarkan tujuan ini, maka kegiatan analisis butir soal memiliki banyak manfaat, di antaranya adalah dapat membantu para pengguna tes dalam evaluasi atas tes yang digunakan, sangat relevan bagi penyusunan tes informal dan lokal seperti tes yang disiapkan guru untuk siswa di kelas, mendukung penulisan butir soal yang efektif, secara materi dapat memperbaiki tes di kelas, meningkatkan validitas soal

dan reliabilitas. Untuk hal itu, Linn dan Gronlund, menambahkan tentang pelaksanaan kegiatan analisis butir soal yang biasanya didesain untuk menjawab pertanyaan-pertanyaan antara lain, apakah fungsi soal sudah tepat?, apakah soal ini memiliki tingkat kesukaran yang tepat?, apakah soal bebas dari hal-hal yang tidak relevan?, apakah pilihan jawabannya efektif?

c. Validitas dan Reliabilitas

Validitas berasal dari kata *validity* yang mempunyai arti sejauh mana ketepatan dan kecermatan suatu instrumen pengukur (tes) dalam melakukan fungsi ukurnya. Suatu tes dikatakan memiliki validitas yang tinggi apabila alat tersebut menjalankan fungsi ukur secara tepat atau memberikan hasil ukur yang sesuai dengan maksud dilakukannya pengukuran tersebut. Artinya hasil ukur dari pengukuran tersebut merupakan besaran yang mencerminkan secara tepat fakta atau keadaan sesungguhnya dari apa yang diukur (Hakim et al., 2021). Validitas tes pada dasarnya menunjuk kepada derajat fungsi pengukurnya suatu tes, atau derajat kecermatan ukurnya sesuatu tes. Validitas suatu tes mempermasalahkan apakah tes tersebut benarbenar mengukur apa yang hendak diukur. Maksudnya adalah seberapa jauh suatu tes mampu mengungkapkan dengan tepat ciri atau keadaan yang sesungguhnya dari obyek ukur, akan tergantung dari tingkat validitas tes yang bersangkutan.

Validitas berkenaan dengan ketepatan alat penilaian terhadap konsep yang dinilai sehingga betul-betul menilai apa yang seharusnya dinilai. Suatu tes yang valid untuk tujuan tertentu atau pengambilan keputusan tertentu, mungkin tidak valid untuk tujuan atau pengambilan keputusan lain. Jadi validitas suatu tes, harus selalu dikaitkan dengan tujuan atau pengambilan keputusan tertentu. Tes masuk di SMA misalnya harus selalu dikaitkan dengan seberapa jauh tes masuk tersebut dapat mencerminkan prestasi atau hasil belajar para calon peserta didik baru setelah belajar nanti. Konsep validitas tes dapat dibedakan atas tiga macam yaitu validitas isi (*content validity*), validitas konstruk (*construct validity*), dan validitas empiris atau validitas kriteria.

Validitas isi suatu tes mempermasalahkan seberapa jauh suatu tes mengukur tingkat penguasaan terhadap isi atau konten atau materi tertentu yang seharusnya dikuasai sesuai dengan tujuan pengajaran. Dengan kata lain tes yang mempunyai validitas isi yang baik ialah tes yang benar-benar mengukur penguasaan materi yang seharusnya dikuasai sesuai dengan konten pengajaran yang tercantum dalam Garis-garis Besar Program Pengajaran (GBPP). Validitas isi menunjukkan sejauhmana pertanyaan, tugas atau butir dalam suatu tes atau instrumen mampu mewakili secara keseluruhan dan proporsional perilaku sampel yang dikenai tes tersebut. Artinva tes itu valid apabila butir-butir tes itu mencerminkan keseluruhan konten atau materi yang diujikan atau yang seharusnya dikuasai secara proporsional (Slamet & Wahyuningsih, 2022).

Validitas konstruk (*construct validity*) adalah validitas yang mempermasalahkan seberapa jauh butir-butir tes mampu mengukur apa yang benar-benar hendak diukur sesuai dengan konsep khusus atau definisi konseptual yang telah ditetapkan. Validitas konstruk biasa digunakan untuk instrumen yang dimaksudkan mengukur variabel konsep, baik yang sifatnya performansi tipikal seperti instrumen untuk mengukur sikap, minat konsep diri, lokus kontrol, gaya kepemimpinan, motivasi berprestasi, dan lain-lain, maupun yang sifatnya performansi maksimum seperti instrumen untuk mengukur bakat (tes bakat), inteligansi (kecerdasan intelektual), kecerdasan, emosional dan lain-lain.

Reliabilitas berasal dari kata *reliability* berarti sejauh mana hasil suatu pengukuran dapat dipercaya. Suatu hasil pengukuran dapat dipercaya apabila dalam beberapa kali pelaksanaan pengukuran terhadap kelompok subyek yang sama, diperoleh hasil pengukuran yang relatif sama, selama aspek yang diukur dalam diri subyek memang belum berubah. Reliabilitas ukuran menyangkut seberapa jauh skor deviasi individu, atau skor-z, relatif konsisten apabila dilakukan pengulangan pengadministrasian dengan tes yang sama atau tes yang ekivalen. reliabilitas merupakan salah-satu ciri atau karakter utama instrumen pengukuran yang baik.

Tujuan utama menghitung reliabilitas skor tes adalah untuk mengetahui tingkat ketepatan (*precision*) dan keajegan (*consistency*) skor tes. Indeks reliabilitas berkisar antara 0-1. Semakin tinggi koefisien reliabilitas suatu tes (mendekati 1), makin tinggi pula keajegan/ ketepatannya. Tes yang memiliki konsistensi reliabilitas tinggi adalah akurat, *reproducibel*, dan *generalized* terhadap kesempatan *testing* dan instrumen tes lainnya (Hidayat, 2021a).

Secara rinci faktor yang mempengaruhi reliabilitas skor tes di antaranya: Semakin banyak jumlah butir soal, semakin ajek suatu tes, semakin lama waktu tes, semakin ajeg, semakin sempit *range* kesukaran butir soal semakin besar keajegan, soal-soal yang saling berhubungan akan mengurangi keajegan, semakin objektif pemberian skor, semakin besar keajegan, ketidaktepatan pemberian skor, menjawab besar soal dengan cara menebak, semakin homogen materi semakin besar keajegan, pengalaman peserta ujian, salah penafsiran terhadap butir soal, menjawab soal dengan buru-buru/cepat, kesiapan mental peserta ujian, adanya gangguan dalam pelaksanaan tes, jarak antara tes pertama dengan tes kedua, mencontek dalam mengerjakan tes, posisi individu dalam belajar, kondisi fisik peserta ujian.

Ada tiga cara yang dapat dilakukan untuk menentukan reliabilitas skor tes: *Pertama*, Keajegan pengukuran ulang: kesesuaian antara hasil pengukuran pertama dan kedua dari sesuatu alat ukur terhadap kelompok yang sama. *Kedua*, Keajegan pengukuran setara: kesesuaian hasil pengukuran dan 2 atau lebih alat ukur berdasarkan kompetensi kisi-kisi yang lama. *Ketiga*, Keajegan belah dua: kesesuaian antara hasil pengukuran belahan pertama dan belahan kedua dari alat ukur yang sama.

d. Daya Pembeda Soal

Daya pembeda suatu butir soal adalah kemampuan butir soal dalam membedakan antara siswa yang dapat menjawab soal dengan tepat dan siswa yang menjawab soal kurang tepat. Analisis daya beda memiliki makna mengkaji suatu butir soal dari segi kesanggupan tes tersebut dalam membedakan siswa yang memiliki kategori rendah dan kategori tinggi. Daya pembeda butir soal adalah kemampuan suatu butir tes untuk dapat membedakan antara testee yang berkemampuan tinggi dan berkemampuan rendah (Fatimah & Alfath, 2019).

Menganalisis daya pembeda artinya mengkaji soal-soal tes dari segi keanggupan tes tersebut dalam membedakan siswa yang termasuk kedalam katagori lemah/rendah dan kategori kuat/tinggi presentasinya. Penentuan daya beda butir soal pada Anates dapat diketahui dalam tabel daya pembeda pada kolom DP persen. Daya pembeda butir soal memiliki manfaat yaitu untuk meningkatkan mutu setiap-butir soal melalui data empiriknya dan untuk mengetahui seberapa jauh masing-masing butir soal dapat membedakan kemapuan siswa,yaitu siswa yang telah memahami atau belum memahami materi yang diajarkan pendidik.

Manfaat daya pembeda butir soal adalah: *Pertama*, Untuk meningkatkan mutu setiap butir soal melalui data empiriknya. Berdasarkan indeks daya pembeda, setiap butir soal dapat diketahui apakah butir soal itu baik, direvisi, atau ditolak. *Kedua*, Untuk mengetahui seberapa jauh setiap butir soal dapat mendeteksi/membedakan kemampuan siswa, yaitu siswa yang telah memahami atau belum memahami materi yang diajarkan guru.

Apabila suatu butir soal tidak dapat membedakan kedua kemampuan siswa itu, maka butir soal itu dapat dicurigai "kemungkinannya" seperti, kunci jawaban butir soal itu tidak tepat, butir soal itu memiliki 2 atau lebih kunci jawaban yang benar, kompetensi yang diukur tidak jelas, pengecoh tidak berfungsi, materi yang ditanyakan terlalu sulit sehingga banyak siswa yang menebak, sebagian besar siswa yang memahami materi yang ditanyakan berpikir ada yang salah informasi dalam butir soalnya. Indeks daya pembeda setiap butir soal biasanya juga dinyatakan dalam bentuk proporsi.

Semakin tinggi indeks daya pembeda soal berarti semakin mampu soal yang bersangkutan membedakan warga belajar/siswa yang telah memahami materi dengan warga belajar/peserta didik yang belum memahami materi. Indeks daya pembeda berkisar antara -1,00 sampai dengan +1,00. Semakin tinggi daya pembeda suatu soal, maka semakin kuat/baik soal itu. Semakin tinggi daya pembeda suatu soal, maka semakin kuat/baik soal itu. Jika daya pembeda negatif (<0) bearti lebih banyak kelompok bawah (peserta didik yang tidak memahami materi) menjawab soal benar dibanding dengan kelompok atas (peserta didik yang memahami materi yang diajarkan guru).

DP = daya pembeda soal, BA = jumlah jawaban benar pada kelompok atas, BB = jumlah jawaban benar pada kelompok bawah, N=jumlah siswa yang mengerjakan tes.

D. KESIMPULAN

Berdasarkan hasil analisis terhadap instrumen ujian sekolah pelajaran PAI kelas XII menggunakan pendekatan teori tes klasik, dapat disimpulkan bahwa kualitas instrumen secara keseluruhan masih belum memenuhi kriteria sebagai alat evaluasi yang baik dan layak digunakan. Hasil uji validitas menunjukkan bahwa hanya dua dari lima belas butir soal yang memenuhi syarat validitas empiris, sedangkan sisanya dinyatakan tidak valid. Hal ini menunjukkan bahwa sebagian besar soal tidak secara akurat mengukur kompetensi yang dituju dalam pembelajaran.

Selain itu, hasil uji reliabilitas menggunakan rumus KR-20 menghasilkan nilai sebesar 0,0227, yang termasuk dalam kategori sangat rendah. Nilai ini mengindikasikan bahwa instrumen tidak memiliki konsistensi internal yang memadai, sehingga hasil evaluasi yang dihasilkan tidak dapat diandalkan. Kondisi ini diperburuk oleh hasil uji daya pembeda, di mana hanya sebagian kecil soal yang mampu membedakan peserta didik berkemampuan tinggi dan rendah, sementara sebagian besar lainnya termasuk dalam kategori ditolak atau bahkan memiliki daya pembeda negatif.

Revisi menyeluruh terhadap sebagian besar butir soal sangat diperlukan, baik dari segi redaksi, pemilihan pengecoh, kunci jawaban, maupun kesesuaian dengan indikator capaian pembelajaran. Artikel ini menekankan pentingnya kompetensi guru dalam menyusun dan menganalisis soal secara empiris sebagai bagian dari peningkatan mutu evaluasi pendidikan.

DAFTAR PUSTAKA

Ambiyar, & D, M. (2019). *Metodologi Penelitian Evaluasi Program*. Alfabeta.

Arikunto, S. (2021). *Dasar-Dasar Evaluasi Pendidikan*. Bumi Aksara.

Devi, M. Y., Hidayanthi, R., & Fitria, Y. (2022). Model-Model Evaluasi Pendidikan dan Model Sepuluh Langkah dalam Penilaian. *Jurnal Basicedu*, *6*(1), Article 1.

Fatimah, L. U., & Alfath, K. (2019). Analisis Kesukaran Soal, Daya Pembeda dan Fungsi Distraktor. *AL-MANAR : Jurnal Komunikasi Dan Pendidikan Islam*, *8*(2), 37–64.

Hakim, R. A., Mustika, I., & Yuliani, W. (2021). Validitas Dan Reliabilitas Angket Motivasi Berprestasi. *FOKUS : Kajian Bimbingan Dan Konseling Dalam Pendidikan*, *4*(4), Article 4.

Hidayat, A. A. (2021a). *Menyusun Instrumen Penelitian & Uji Validitas-Reliabilitas*. Health Books.

Muryadi, A. D. (2017). Model Evaluasi Program Dalam Penelitian Evaluasi. *PENJAS (Penelitian, Pendidikan Dan Pengajaran)*, *3*(1).

Qadir, A., Huda, N., & Hermina, D. (2024). Analisis Butir Tes: Tingkat Kesukaran, Daya Pembeda Dan Efektivitas Pengecoh. *Al-Furqan : Jurnal Agama, Sosial, Dan Budaya*, *3*(3), 1450–1467.

Slamet, R., & Wahyuningsih, S. (2022). Validitas Dan Reliabilitas Terhadap Instrumen Kepuasan Kerja. *Aliansi : Jurnal Manajemen Dan Bisnis*, *17*(2). https://doi.org/10.46975/aliansi.v17i2.428

Unaradjan, D. D. (2019). *Metode Penelitian Kuantitatif*. Unika Atma Jaya Jakarta.