

Bagaimana Kualitas Instrumen Penilaian Pembelajaran yang Digunakan? Kajian Empiris Kualitas Butir Soal

Yuliana Olga Siba Sabon^{1*}, Lamba Toding Palimbu², Hajra Yansa³, Dite Umbara Alfansuri⁴

^{1,3,4}Universitas Musamus, Jl. Kedungmundu Raya No. 18, Kec. Tembalang, Kota Semarang, Jawa Tengah.

²Universitas Satya Wiyata Mandala, Jl. Sutamsu, Kalibobo, Distrik Nabire, Kab. Nabire, Papua Tengah
yulianasabon@unmus.ac.id

Abstract

Instrument quality analysis provides important information for test developers and teachers in ensuring the quality of educational measurement instruments. This study aimed to analyze the quality of an Economics learning assessment instrument based on empirical data. A quantitative descriptive approach was employed to examine item characteristics, including item difficulty, item discrimination, and instrument reliability. The data were obtained from 10 multiple-choice items administered to 38 students and analyzed using Classical Test Theory (CTT). The results showed that the distribution of item difficulty was predominantly moderate. In terms of item discrimination, 80% of the items demonstrated good discrimination, while 20% were categorized as having adequate discrimination. The instrument's reliability coefficient was categorized as low. Based on the item analysis, 8 of the 10 items were considered suitable for use. These findings indicate that empirical item analysis can provide valuable information for evaluating and improving the quality of learning assessment instruments.

Keywords: assessment, Classical Test Theory, instrument quality

Abstrak

Analisis kualitas instrumen memberikan informasi penting bagi pengembang instrumen dan guru dalam memastikan kualitas alat ukur pembelajaran. Penelitian ini bertujuan untuk menganalisis kualitas instrumen penilaian pembelajaran Ekonomi berdasarkan data empiris. Penelitian ini menggunakan pendekatan kuantitatif deskriptif untuk menganalisis karakteristik butir soal yang meliputi tingkat kesukaran, daya pembeda, dan reliabilitas instrumen. Data diperoleh dari 10 butir soal pilihan ganda yang diujikan kepada 38 siswa dan dianalisis menggunakan pendekatan Classical Test Theory (CTT). Hasil analisis menunjukkan bahwa distribusi tingkat kesukaran butir didominasi oleh kategori sedang. Sebanyak 80% butir memiliki daya pembeda yang baik, sedangkan 20% berada pada kategori cukup. Koefisien reliabilitas instrumen berada pada kategori rendah. Berdasarkan hasil analisis karakteristik butir, 8 dari 10 butir soal dinyatakan layak digunakan. Temuan ini menunjukkan bahwa analisis empiris butir soal dapat memberikan informasi penting dalam mengevaluasi dan meningkatkan kualitas instrumen penilaian pembelajaran.

Kata kunci: Classical Test Theory, kualitas instrumen, penilaian

Copyright (c) 2026 Yuliana Olga Siba Sabon, Lamba Toding Palimbu, Hajra Yansa, Dite Umbara Alfansuri

✉ Corresponding author: Yuliana Olga Siba Sabon

Email Address: yulianasabon@unmus.ac.id (Jl. Kedungmundu Raya No. 18, Kota Semarang, Jawa Tengah.)

Received 08 Agustus 2026, Accepted 15 September 2026, Published 22 September 2026

PENDAHULUAN

Penilaian pembelajaran merupakan cara yang dilakukan untuk mengetahui capaian belajar siswa dan aspek perkembangan siswa. Penilaian memberikan gambaran tentang pengetahuan dan keterampilan siswa secara akurat (Gao et al., 2020). Oleh karena itu penggunaan instrumen penilaian pembelajaran yang tepat perlu diperhatikan. Instrumen penilaian yang tidak memenuhi kriteria kualitas secara psikometrik dapat menghasilkan data yang bias dan tidak menggambarkan kemampuan nyata siswa (Rahim et al., 2025; Syadiah & Hamdu, 2020). Agar hasil penilaian dapat memberikan gambaran yang akurat tentang capaian belajar siswa dan dapat menjadi acuan evaluasi, maka instrumen penilaian yang digunakan harus berkualitas.

Kualitas instrumen dapat diketahui melalui analisis butir soal. Analisis butir soal memastikan

bahwa instrument penilaian yang digunakan berkualitas sesuai dengan standar kurikulum dan tujuan pembelajaran (Fatimah & Alfath, 2019). Analisis butir soal berperan penting dalam evaluasi kualitas instrumen penilaian karena dapat mengidentifikasi butir-butir yang baik, perlu direvisi, atau harus dihapus dari paket tes.

Kualitas instrumen mencakup indeks kesukaran, daya pembeda, efektivitas pengecoh (distractor), validitas, dan reliabilitas. Indeks kesukaran mengungkap sejauh mana peserta didik mampu menjawab butir soal dengan benar. Kriteria indeks kesukaran $> 0,7$ menunjukkan soal tergolong mudah; sedangkan $0,3 - 0,7$ tingkat kesukaran soal sedang; dan indeks kesukaran $< 0,3$ menunjukkan soal sulit. Kualitas instrument juga ditunjukkan oleh indeks daya beda. Indeks daya beda menggambarkan kemampuan butir tersebut dalam membedakan antara siswa berkemampuan tinggi dan rendah dengan rentang sebagai berikut: $0,3 - 1,00$ daya beda soal baik; $0,2 - 0,29$ daya beda soal diterima, tetapi soal perlu perbaikan; $< 0,2$ daya beda soal buruk dan butir soal dapat dieliminasi. Parameter kualitas instrument lainnya adalah efektivitas pengecoh; mengukur keberfungsian alternatif jawaban yang tidak benar dalam menarik respon peserta yang kurang paham materi (Istiyono, 2020). Estimasi reliabilitas dan pembuktian validitas instrument sangat penting untuk dilakukan. Reliabilitas merupakan konsistensi instrumen dalam mengukur konstruk yang sama. Kriteria reliabilitas minimum menurut Bujang, et al. (2024) adalah 0,60. Validitas menunjukkan tingkat kesesuaian instrument dengan konstruk yang hendak diukur. Suatu instrumen disebut memiliki validitas yang tinggi apabila dapat menjalankan fungsi ukur secara tepat atau memberikan hasil ukur yang sesuai dengan maksud dilakukannya pengukuran (Ramadhan et al., 2024; Souza et al., 2017). Dengan menganalisis parameter-parameter tersebut, membantu guru atau pengembang soal menentukan soal yang layak digunakan (Saat, 2020). Analisis butir soal memberikan umpan balik yang berharga dan berkelanjutan bagi pengembang soal untuk menyempurnakan instrumen penilaian.

Salah satu pendekatan analisis butir soal adalah Classical Test Theory (CTT). CTT merupakan pendekatan fundamental dalam evaluasi kualitas instrumen penilaian pembelajaran yang sampai saat ini masih relevan dan banyak digunakan dalam analisis butir soal. CTT berfokus pada hubungan antara skor observasi, skor sebenarnya, dan kesalahan pengukuran, serta menyediakan parameter-parameter dasar seperti indeks kesukaran butir, indeks daya pembeda, reliabilitas, dan efektivitas pengecoh sebagai indikator kualitas butir soal. Pemahaman yang baik tentang CTT meletakkan fondasi bagi prinsip-prinsip pengukuran (Wu et al., 2017). Keunggulan pendekatan CTT adalah praktis dan relatif mudah diterapkan, sehingga dapat digunakan dalam konteks evaluasi pembelajaran kelas, ujian sekolah maupun ujian nasional (Fitraynsyah et al., 2024; Ridha et al., 2025; Setiawan et al., 2025).

Dalam konteks pelajaran ekonomi, guru di sekolah menengah sering menghadapi kesulitan dalam menyusun soal yang sesuai dengan standar kurikulum. Proses penilaian yang dilakukan di sekolah sebatas mengetahui sejauh mana hasil pembelajaran peserta didik. Guru jarang memanfaatkan hasil penilaian sebagai acuan evaluasi terhadap instrument penilaian yang digunakan (Rasimah et al., 2026; Sabon et al., 2025). Hal ini bukan hanya berdampak pada akurasi penilaian, tetapi juga dapat

memengaruhi motivasi belajar siswa serta kepercayaan terhadap proses evaluasi itu sendiri (Rahim et al., 2025). Oleh karena itu, perlu adanya upaya sistematis untuk menilai dan menyempurnakan soal melalui pendekatan ilmiah.

Penelitian relevan tentang analisis butir soal dengan pendekatan CTT telah banyak dilakukan dengan menggunakan berbagai software yang berbeda. Banyak penelitian yang menampilkan hasil analisis kualitas instrument yang sangat baik dan memenuhi kriteria sesuai dengan angka patokan parameter kualitas soal yang telah disepakati oleh ahli (Rachmawati & Pradana, 2025; Rachmawati & Rahman, 2025; Riani et al., 2020; Setiawan et al., 2025). Namun demikian, data di lapangan ditemukan bahwa tidak semua instrument penilaian yang digunakan memenuhi standar instrument berkualitas secara psikometri meskipun instrumen tersebut disusun oleh guru yang kompeten. Kajian empirik tentang kualitas butir soal yang belum memenuhi standar juga perlu dilakukan untuk memberikan informasi yang berguna bagi pengembang instrument ataupun guru.

Berdasarkan latar belakang tersebut penelitian ini dilakukan untuk mengetahui kualitas butir soal yang digunakan dalam penilaian pembelajaran ekonomi dengan pendekatan CTT serta memberikan rekomendasi kelayakan butir soal. Kajian mendalam tentang kualitas butir soal yang memenuhi maupun tidak memenuhi kualitas instrument secara psikometrik dipaparkan sebagai referensi bagi guru maupun penyusun soal.

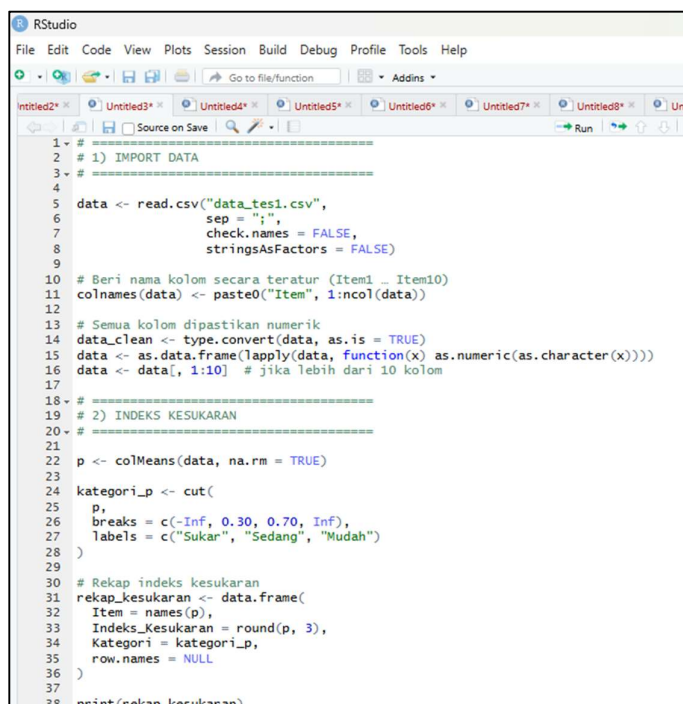
METODE

Penelitian ini merupakan penelitian kuantitatif dengan pendekatan deskriptif. Penelitian dilakukan dengan tujuan untuk menganalisis kualitas butir soal. Analisis butir soal yang dilakukan secara kuantitatif yaitu penelaahan soal didasarkan pada data empirik yang diperoleh dari soal yang telah diujikan (Fatimah & Alfath, 2019). Karakteristik butir soal yang dianalisis dalam penelitian ini adalah tingkat kesukaran, daya pembeda, dan reliabilitas dengan prinsip-prinsip *Classical Test Theory* (CTT). Objek penelitian ini adalah 10 butir soal pilihan ganda mata pelajaran ekonomi pada sebuah kelas di Sekolah Menengah Atas di Nabire Papua Tengah yang berjumlah 38 siswa. Instrumen utama dalam penelitian ini adalah paket soal berbentuk pilihan ganda menggunakan aplikasi Exambro. Data hasil ujian yang dikumpulkan berbentuk skor 0: salah dan skor 1: benar. Oleh karena itu karakteristik soal yang dianalisis dalam penelitian ini adalah tingkat kesukaran butir, indeks daya beda, dan reliabilitas. Data dianalisis menggunakan pendekatan *Classical Test Theory* (CTT) dengan bantuan software analisis RSudio dengan pustaka CTT dan Psych.

HASIL DAN DISKUSI

Hasil

Analisis kualitas butir soal dengan karakteristik parameter tingkat kesukaran butir, indeks daya beda, dan reliabilitas alpha-cronbach dilakukan dengan skrip analisis bantuan Pustaka “CTT” dan “Psych” pada software analisis RStudio yang dapat dilihat pada Gambar 1 dan 2.

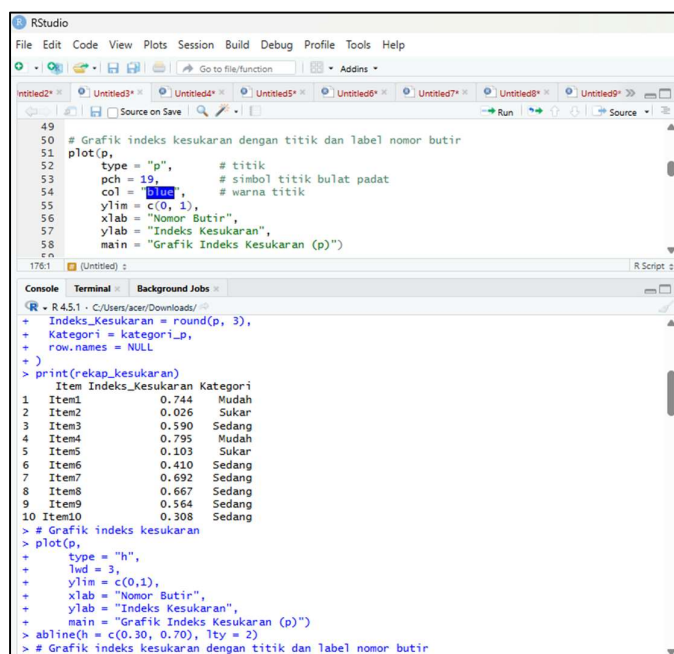


```

1 # =====
2 # 1) IMPORT DATA
3 # =====
4
5 data <- read.csv("data_tes1.csv",
6                 sep = ";",
7                 check.names = FALSE,
8                 stringsAsFactors = FALSE)
9
10 # Beri nama kolom secara teratur (Item1 .. Item10)
11 colnames(data) <- paste0("Item", 1:ncol(data))
12
13 # Semua kolom dipastikan numerik
14 data_clean <- type.convert(data, as.is = TRUE)
15 data <- as.data.frame(lapply(data, function(x) as.numeric(as.character(x))))
16 data <- data[, 1:10] # jika lebih dari 10 kolom
17
18 # =====
19 # 2) INDEKS KESUKARAN
20 # =====
21
22 p <- colMeans(data, na.rm = TRUE)
23
24 kategori_p <- cut(
25   p,
26   breaks = c(-Inf, 0.30, 0.70, Inf),
27   labels = c("Sukar", "Sedang", "Mudah")
28 )
29
30 # Rekap indeks kesukaran
31 rekap_kesukaran <- data.frame(
32   Item = names(p),
33   Indeks_Kesukaran = round(p, 3),
34   Kategori = kategori_p,
35   row.names = NULL
36 )
37
38 print(rekap_kesukaran)

```

Gambar 1. Skrip analisis pada Rstudio



```

49
50 # Grafik indeks kesukaran dengan titik dan label nomor butir
51 plot(p,
52      type = "p", # titik
53      pch = 19, # simbol titik bulat padat
54      col = "blue", # warna titik
55      ylim = c(0, 1),
56      xlab = "Nomor Butir",
57      ylab = "Indeks Kesukaran",
58      main = "Grafik Indeks Kesukaran (p)")
59
60
176:1 (Untitled) :

```

```

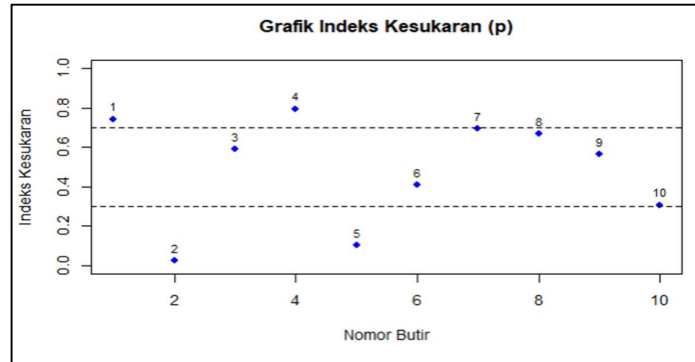
+ Indeks_Kesukaran = round(p, 3),
+ Kategori = kategori_p,
+ row.names = NULL
+ )
> print(rekap_kesukaran)
  Item Indeks_Kesukaran Kategori
1 Item1          0.744    Mudah
2 Item2          0.026     Sukar
3 Item3          0.590    Sedang
4 Item4          0.795    Mudah
5 Item5          0.103     Sukar
6 Item6          0.410    Sedang
7 Item7          0.692    Sedang
8 Item8          0.667    Sedang
9 Item9          0.564    Sedang
10 Item10         0.308    Sedang
> # Grafik indeks kesukaran
> plot(p,
+      type = "h",
+      lwd = 3,
+      ylim = c(0,1),
+      xlab = "Nomor Butir",
+      ylab = "Indeks Kesukaran",
+      main = "Grafik Indeks Kesukaran (p)")
> abline(h = c(0.30, 0.70), lty = 2)
> # Grafik indeks kesukaran dengan titik dan label nomor butir

```

Gambar 2. Tampilan analisis pada Rstudio

Indeks Kesukaran Butir

Indeks kesukaran butir merujuk pada tingkat kesukaran sebuah butir soal. Indeks kesukaran butir soal berperan dalam menentukan sejauh mana sebuah soal dapat dijawab dengan benar oleh peserta didik. Hasil analisis tingkat kesukaran butir soal ekonomi penilaian akhir semester dengan bantuan R studio ditampilkan dalam Gambar 3.



Gambar 3. Indeks kesukaran butir

Berdasarkan Gambar 3, garis batas kuadran bawah dengan indeks kesukaran berkisar dari 0 – 0,3 menunjukkan soal yang sulit. Sedangkan pada kuadran atas > 0,7 menunjukkan soal yang mudah.

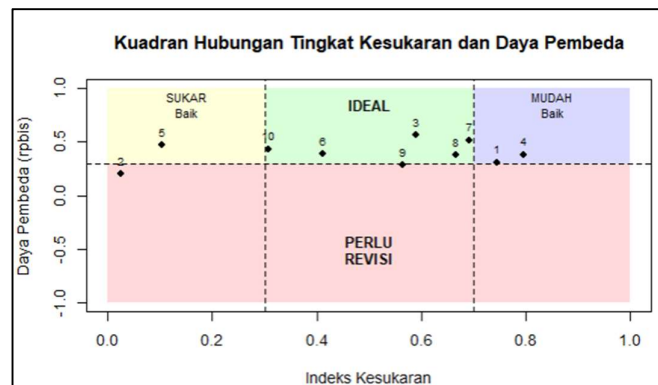
Indeks Daya Beda

Indeks daya beda menunjukkan kemampuan sebuah soal untuk membedakan peserta didik kemampuan tinggi dan yang memiliki kemampuan rendah. Hasil analisis daya beda soal disajikan pada Tabel 1.

Tabel 1. Indeks daya beda setiap butir soal

Item	Point Biserial	Kategori
Item1	0.312	Baik
Item2	0.202	Cukup Diterima
Item3	0.566	Baik
Item4	0.382	Baik
Item5	0.470	Baik
Item6	0.390	Baik
Item7	0.518	Baik
Item8	0.375	Baik
Item9	0.283	Cukup Diterima
Item10	0.435	Baik

Rekap tingkat kesukaran butir soal dan indeks daya beda soal memberikan informasi tentang butir ideal yang dapat digunakan untuk penilaian selanjutnya. Gambar 4 menunjukkan bagaimana posisi setiap butir berdasarkan kesukaran dan daya beda soal.



Gambar 4. Rekap Indeks kesukaran dan daya beda soal

Reliabilitas

Reliabilitas instrumen menunjukkan tingkat konsistensi soal ketika digunakan kembali dalam situasi yang sama. Reliabilitas dalam analisis ini menunjukkan tingkat konsistensi internal yakni semua soal mengukur karakteristik yang sama yakni pengetahuan peserta didik tentang pembelajaran ekonomi. Estimasi reliabilitas diketahui bahwa soal mata pelajaran ekonomi yang digunakan dalam penilaian akhir semester memiliki indeks reliabilitas yang rendah yakni 0.48.

Diskusi

Analisis butir soal dapat dilakukan dengan pendekatan teori tes klasik (CTT) maupun modern (IRT). Dalam konteks penelitian ini analisis butir soal dilakukan dengan menggunakan CTT karena dalam skala kecil atau tingkat kelas. Hal ini didukung oleh penelitian sebelumnya yang menyatakan bahwa meskipun IRT merupakan pendekatan yang lebih unggul dalam mengevaluasi kualitas butir soal dan dapat menganalisis kemampuan peserta didik secara akurat namun pendekatan yang lebih relevan terutama untuk pengukuran di kelas dengan skala kecil sebaiknya menggunakan CTT (Fitraynsyah et al., 2024; Ridha et al., 2025; Setiawan et al., 2025). Hal tersebut menunjukkan bahwa penggunaan analisis butir soal dengan pendekatan CTT masih relevan dan potensial dalam meningkatkan kualitas evaluasi pembelajaran di sekolah.

Salah satu kunci kualitas evaluasi pembelajaran di sekolah terletak pada kualitas instrument yang digunakan. Salah satu kualitas instrument penilaian yang dikaji dalam penelitian ini tingkat kesukaran soal. Soal yang baik adalah soal yang tidak terlalu mudah dan tidak terlalu sukar atau distribusi tingkat kesukaran soal tidak didominasi oleh soal yang sukar atau soal yang mudah (Fatimah & Alfath, 2019). Berdasarkan hasil analisis diketahui bahwa distribusi tingkat kesukaran soal pada penilaian akhir semester mata pelajaran ekonomi adalah 20 % soal mudah; 60 % dengan tingkat kesukaran sedang, dan 20 % soal sulit. Berdasarkan distribusi soal tersebut maka dapat dikatakan bahwa distribusi tingkat kesukaran soal sudah baik karena mayoritas soal berada pada tingkat kesukaran sedang. Hal ini didukung oleh beberapa penelitian terdahulu bahwa soal yang baik adalah soal yang tidak terlalu mudah atau terlalu sukar. Soal yang terlalu mudah tidak memotivasi peserta didik untuk berusaha; sedangkan soal yang terlalu sulit menyebabkan peserta didik menjadi frustrasi dan bahkan dapat menurunkan motivasi belajar serta mengaburkan penilaian objektif atas kompetensi yang dimiliki (Aini et al., 2025; Qadir et al., 2024).

Parameter kualitas instrumen penilaian lainnya adalah indeks daya beda. Hasil analisis daya beda soal penilaian akhir semester mata pelajaran ekonomi pada Tabel 1 diketahui bahwa hanya terdapat 80 % soal dapat membedakan peserta didik dengan kemampuan tinggi dan kemampuan rendah dengan baik dan 20% soal cukup membedakan peserta didik dengan kemampuan tinggi dan rendah. Artinya soal – soal tersebut perlu perbaikan dari segi konstruksi dan substansi.

Nilai korelasi point biserial yang rendah terletak pada soal nomor 2 yakni 0,202 dan pada soal nomor 9 yakni 0,283. Nilai korelasi point biserial yang rendah pada analisis daya beda soal berdampak pada indeks reliabilitas soal. Hal ini didukung oleh penelitian lainnya bahwa item-total correlation

(termasuk point-biserial) yang rendah menurunkan nilai reliabilitas (Marianti et al., 2023; Souza et al., 2017; Stefana et al., 2025).

Hasil analisis menunjukkan bahwa indeks reliabilitas soal rendah yakni 0,48. Nilai koefisien reliabilitas yang ideal adalah minimal 0,60. Apabila nilai reliabilitas $< 0,60$ maka soal tidak dapat diadalkan (Bujang et al., 2024; Slamet & Wahyuningsih, 2022). Namun demikian reliabilitas yang rendah bukan berarti soal tersebut tidak dapat berfungsi karena reliabilitas soal sangat bergantung pada karakteristik sampel yang mengikuti tes (Ikhsanudin et al., 2024; Lee & Zhang, 2017; Taber, 2017).

Soal dengan indeks reliabilitas yang rendah mengindikasikan butir soal yang digunakan dalam pengukuran memiliki jawaban yang tidak konsisten (Souza et al., 2017). Sedangkan menurut Ohiri *et al.* (2024) reliabilitas soal yang rendah disebabkan oleh pertanyaan yang ambigu sehingga ditafsirkan oleh peserta tes dengan cara yang berbeda. Beberapa penelitian menyarankan bahwa dengan menghilangkan butir soal dengan nilai korelasi yang rendah dapat meningkatkan indeks reliabilitas soal, selain itu perlu dicek kembali konstruk instrumen, dan karakteristik peserta tes (Hossan et al., 2025; Souza et al., 2017; Zakariya, 2022).

Implikasi dari hasil evaluasi butir soal yang telah dilakukan adalah dari 10 butir soal instrument penilaian pembelajaran ekonomi jumlah butir yang layak digunakan dalam penilaian akhir tahun adalah 8 butir yakni butir 1, 3, 4, 5, 6, 7, 8, dan 10. Butir soal tersebut tergolong ideal dan tanpa perbaikan dapat digunakan. Butir 2 merupakan butir yang sukar tetapi memiliki daya beda yang cukup. Menurut Fatimah & Alfath (2019) tingkat kesukaran butir soal memiliki hubungan dengan daya beda soal. Butir yang memiliki tingkat kesukaran yang maksimal, maka daya pembedanya akan rendah, demikian pula bila butir soal yang terlalu mudah tidak memiliki daya pembeda. Dalam penelitian ini ditemukan juga bahwa butir dengan tingkat kesukaran sedang (butir 9) juga memiliki daya beda yang cukup. Hal ini menunjukkan bahwa butir 2 dan 9 perlu diperbaiki agar dapat membedakan peserta didik yang berkemampuan tinggi dan kemampuan rendah. Perbaikan dapat dilakukan berdasarkan tingkat kognitif butir soal atau tata bahasa pada pertanyaan yang digunakan. Hal ini didukung oleh penelitian Yazdi, *et al.* (2021) yang menyatakan bahwa beberapa faktor yang mempengaruhi peserta didik menjawab soal yakni tingkat kognitif dari pertanyaan, situasi, kondisi tes, dan kondisi mental peserta tes.

KESIMPULAN

Analisis butir soal penting dilakukan untuk memperoleh informasi kualitas dan keberfungsian instrumen penilaian dalam mengukur capaian belajar peserta didik. Soal dengan kualitas yang baik dapat berfungsi memberikan informasi hasil penilaian secara akurat. Tingkat kesukaran butir soal penilaian pembelajaran ekonomi tergolong baik karena didominasi oleh soal dengan tingkat kesukaran sedang. Mayoritas butir soal yakni 80% butir soal memiliki daya beda yang baik. Dari 10 butir soal, jumlah butir yang layak digunakan dalam penilaian selanjutnya adalah 8 butir yakni butir 1, 3, 4, 5, 6, 7, 8, dan 10. Namun demikian reliabilitas soal pada kategori yang rendah. Oleh karena itu perlu dicek kembali konstruk instrumen, dan karakteristik peserta tes.

REFERENSI

- Aini, N., Ramadhan, M. R., & Wijdan, G. (2025). Analisis tingkat kesukaran butir soal try out busana PPG bagi guru tertentu tahap 1 Universitas Negeri Malang tahun 2025. *Pendas : Jurnal Ilmiah Pendidikan Dasar*, 10(4), 367–381. <https://doi.org/10.23969/jp.v10i4.33638>
- Bujang, M. A., Omar, E. D., Foo, D. H. P., & Hon, Y. K. (2024). Sample size determination for conducting a pilot study to assess reliability of a questionnaire. *Restorative Dentistry & Endodontics*, 49(1). <https://doi.org/10.5395/RDE.2024.49.E3>
- 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. <https://doi.org/10.36668/jal.v8i2.115>
- Fitraynsyah, M. A., Hilmiyati, F., & Sultan Maulana Hasanuddin, U. (2024). Peran tingkat kesukaran dan daya pembeda dalam analisis butir tes: Kajian literatur untuk pendidikan menengah. *Jurnal Riset Evaluasi Pendidikan*, 1(4), 252–262. <https://doi.org/10.51574/JREP.V1I4.2250>
- Gao, X., Li, P., Shen, J., & Sun, H. (2020). Reviewing assessment of student learning in interdisciplinary STEM education. *International Journal of STEM Education* 2020 7:1, 7(1), 24–. <https://doi.org/10.1186/s40594-020-00225-4>
- Hossan, D., Wolfs, B., Jesmin, N. S., & Petkovic, M. (2025). Questionnaire validity and reliability : a review with practical guidelines. *Journal of Entrepreneurship, Business and Economics*, 13(1), 135–186.
- Ikhsanudin, Subali, B., Retnawati, H., & Istiyono, E. (2024). Estimation of Cronbach reliability based on sample size, gender, and the grades. *International Journal of Evaluation and Research in Education* , 13(2), 759–766. <https://doi.org/10.11591/ijere.v13i2.24895>
- Istiyono, E. (2020). Pengembangan instrumen penilaian dan analisis hasil belajar fisika dengan teori tes klasik dan modern (2nd ed.). UNY Press.
- Lee, Y. H., & Zhang, J. (2017). Effects of differential item functioning on examinees' test performance and reliability of test. *International Journal of Testing*, 17(1), 23–54. <https://doi.org/10.1080/15305058.2016.1224888>
- Marianti, S., Rufaida, A., Hasanah, N., & Nuryanti, S. (2023). Comparing item-total correlation and item-theta correlation in test item selection: A simulation and empirical study. *Jurnal Penelitian Dan Evaluasi Pendidikan*, 27(2), 133–145. <https://doi.org/10.21831/PEP.V27I2.61477>
- Ohiri, S. C., Ihebom, D., & Nnennaya, C. (2024). Psychometric properties of a test: an overview. *International Journal of Research Publication and Reviews*, 5(2), 2217–2224. <https://doi.org/10.55248/gengpi.5.0224.0539>
- Qadir, A., Huda, N., & Hermina, D. (2024). Analisis butir tes: tingkat kesukaran, daya pembeda, eektivitas pengecoh. *Cv.Eureka Media Aksara*, 3(3), 105–134.
- Rachmawati, D., & Pradana, A. B. (2025). Analisis butir soal mata pelajaran ekonomi: Validitas, reliabilitas, tingkat kesukaran, dan daya pembeda. *Jurnal Pendidikan Ekonomi (JUPE)*, 13(3),

- 273–284. <https://doi.org/10.26740/jupe.v13n3.p273-284>
- Rahim, A., Sabon, Y. O. S., Nurhayati, N., Priyudahari, B. A., Islam, R. F., & Amri, H. (2025). Analisis butir soal pilihan ganda dengan pendekatan classical test theory pada ujian akhir semester mata pelajaran matematika. *Indo-MathEdu Intellectuals Journal*, 6(4), 4763–4776.
- Rahmawati, N. S., & Rahman, M. F. (2025). Analisis butir soal objektif penilaian sumatif akhir semester gasal mata pelajaran ekonomi. *JiIP - Jurnal Ilmiah Ilmu Pendidikan*, 8(8), 8863–8873. <https://doi.org/10.54371/jiip.v8i8.8837>
- Ramadhan, M. F., Siroj, R. A., & Afgani, M. W. (2024). Validitas dan reliabilitas. *Journal on Education*, 6(2), 10967–10975. <https://doi.org/10.31004/joe.v6i2.4885>
- Rasimah, E., Hayati, F., Nurhawan, & Rini, E. M. (2026). Analisis Distribusi Nilai pada Ujian Akhir Semester di SDN Bojonggede 06. *Jurnal Indragiri Penelitian Multidisiplin*, 6(1), 52–57. <https://ejournal.indrainstitute.id/index.php/jipm/article/view/1434>
- Riani, D., Almujaib, S., Dina, A., Fitriani, F., & Budiarto, R. (2020). Analisis butir soal dan kemampuan siswa dalam menjawab soal ujian nasional pada mata pelajaran ekonomi. *OIKOS: Jurnal Kajian Pendidikan Ekonomi Dan Ilmu Ekonomi*, 4(1), 70–79. <https://doi.org/10.23969/oikos.v4i1.2425>
- Ridha, A. R., Luthfi, M., & Mudijono. (2025). Evaluasi butir soal secara klasik dalam evaluasi pendidikan. *JIPDAS (Jurnal Ilmiah Pendidikan Dasar)*, 4(1), 100–103. <https://doi.org/10.65946/G6E91258>
- Saat, N. A. (2020). Analisis item kertas peperiksaan sumatif menggunakan teori ujian klasik. *Sains Humanika*, 12(2–2). <https://doi.org/10.11113/sh.v12n2-2.1788>
- Sabon, Y. O. S., Palimbu, L. T., & Edwar, R. C. (2025). Evaluasi Pembelajaran Ekonomi dalam Mendukung Pengalaman Belajar Mendalam. 6(8), 12370–12379.
- Setiawan, A. N., Ulfa Kamalia, P., & Java, E. (2025). Evaluating test item quality: A comprehensive analysis of economics multiple-choice questions in Indonesian high schools (Case study at SMA Negeri 1 Gedangan). *Asatiza: Jurnal Pendidikan*, 6(3), 287–297. <https://doi.org/10.46963/asatiza.v6i3.3168>
- 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>
- Souza, A. C. de, Alexandre, N. M. C., & Guirardello, E. de B. (2017). Psychometric properties in instruments evaluation of reliability and validity. *Epidemiologia e Serviços de Saúde*, 26(3), 649–659. <https://doi.org/10.5123/S1679-49742017000300022>
- Stefana, A., Damiani, S., Granziol, U., Provenzani, U., Solmi, M., Youngstrom, E. A., & Fusar-Poli, P. (2025). Psychological, psychiatric, and behavioral sciences measurement scales: best practice guidelines for their development and validation. *Frontiers in Psychology*, 15, 1494261. <https://doi.org/10.3389/FPSYG.2024.1494261>
- Syadiah, A. N., & Hamdu, G. (2020). Analisis rasch untuk soal tes berpikir kritis pada pembelajaran STEM di sekolah dasar. *Premiere Educandum : Jurnal Pendidikan Dasar Dan Pembelajaran*,

- 10(2), 138-148–138 – 148. <https://doi.org/10.25273/pe.v10i2.6524>
- Taber, K. S. (2017). The use of Cronbach's alpha when developing and reporting research instruments in science education. *Research in Science Education* 2016 48:6, 48(6), 1273–1296. <https://doi.org/10.1007/S11165-016-9602-2>
- Wu, M., Tam, H. P., & Jen, T.-H. (2017). Classical test theory. *Educational Measurement for Applied Researchers*, 73–90. https://doi.org/10.1007/978-981-10-3302-5_5
- Yazdi, M. S. G., Majid Haghighat Shoar, S., Sobhani, G., Vafi sani, F., Khoshkholgh, R., Mousavi bazaz, N., Mansourzadeh, A., Yazdi, G. M., Shoar, H. S., Gh, S., & Sani, V. F. (2021). Factors affecting students' guesswork in multiple choice questions and corrective strategies. *Medical Education Bulletin*, 2(4), 297–305. <https://doi.org/10.22034/MEB.2021.312176.1032>
- Zakariya, Y. F. (2022). Cronbach's alpha in mathematics education research: Its appropriateness, overuse, and alternatives in estimating scale reliability. *Frontiers in Psychology*, 13, 1074430. <https://doi.org/10.3389/FPSYG.2022.1074430/TEXT>