

ITEM ANALYSIS OF BUFFER SOLUTION MULTIPLE-CHOICE TEST

¹Kiky Setiawan, ²Rufi'i, ³Sabariah

^{1, 2, 3}Program Studi Pascasarjana Teknologi Pendidikan, Universitas PGRI
Adibuwana, Surabaya, Indonesia

[¹kie.306@gmail.com](mailto:kie.306@gmail.com), [²rufi'i@unipasby.ac.id](mailto:rufi'i@unipasby.ac.id), [³sabariah@unipasby.ac.id](mailto:sabariah@unipasby.ac.id)

ABSTRACT

This study evaluated the quality of an end-of-chapter multiple-choice test on buffer solution material by examining reliability, item difficulty, discrimination, and distractor effectiveness. A descriptive quantitative design based on classical test theory was used. The data consisted of responses from 35 students to 25 multiple-choice items and were analyzed using ANATES. The results showed a reliability coefficient of 0.72, indicating acceptable overall consistency. However, all items were categorized as very easy, with correct response rates ranging from 88.57% to 100%. Discrimination analysis showed that 17 items were poor, 5 were moderate, 2 were good, and 1 was very good. Distractor analysis also indicated that most incorrect options were non-functional because they were rarely or never selected. Only items 2, 7, and 23 were recommended for retention without revision, while most items required revision or removal. These findings indicate that total-score reliability alone is insufficient for judging assessment quality. Routine item analysis is therefore essential to strengthen the diagnostic usefulness of chemistry classroom assessment.

Keywords: *Item analysis, classroom assessment, multiple-choice test, test reliability, buffer solution*

ABSTRAK

Penelitian ini bertujuan mengevaluasi kualitas tes pilihan ganda akhir bab pada materi larutan penyangga berdasarkan reliabilitas, tingkat kesukaran, daya pembeda, dan efektivitas pengecoh. Penelitian menggunakan desain kuantitatif deskriptif berbasis teori tes klasik. Data berupa respons 35 siswa terhadap 25 butir soal yang dianalisis dengan ANATES. Hasil menunjukkan koefisien reliabilitas tes sebesar 0,72, sehingga konsistensi umum instrumen tergolong dapat diterima. Namun, seluruh butir berada pada kategori sangat mudah dengan rentang jawaban benar 88,57%-100%. Sebanyak 17 butir memiliki daya pembeda buruk, 5 sedang, 2 baik, dan 1 sangat baik. Analisis pengecoh juga

menunjukkan sebagian besar opsi salah tidak berfungsi karena jarang atau tidak pernah dipilih. Hanya butir 2, 7, dan 23 direkomendasikan untuk dipertahankan tanpa revisi, sementara sebagian besar butir perlu direvisi atau dibuang. Temuan ini menegaskan bahwa reliabilitas total belum cukup untuk menyatakan kualitas instrumen, sehingga analisis butir perlu dilakukan secara rutin dalam penilaian pembelajaran kimia.

Kata Kunci: Analisis butir, penilaian kelas, tes pilihan ganda, reliabilitas tes, larutan penyangga

A. Pendahuluan

Penilaian merupakan salah satu komponen utama dalam proses pembelajaran yang berfungsi untuk memperoleh informasi mengenai pencapaian kompetensi peserta didik sekaligus menjadi dasar dalam pengambilan keputusan pembelajaran. Hasil penilaian yang akurat sangat bergantung pada kualitas instrumen yang digunakan. Instrumen yang memiliki karakteristik psikometrik yang baik akan menghasilkan informasi yang valid dan reliabel mengenai kemampuan peserta didik, sedangkan instrumen yang kurang berkualitas dapat menghasilkan interpretasi yang keliru terhadap capaian belajar siswa (Kubo, 2023; DeCarlo, 2023). Oleh karena itu, pengembangan dan evaluasi instrumen penilaian menjadi

bagian yang tidak terpisahkan dari upaya meningkatkan kualitas pembelajaran.

Tes pilihan ganda merupakan bentuk instrumen yang paling banyak digunakan dalam evaluasi pembelajaran karena memiliki tingkat objektivitas yang tinggi, mudah diadministrasikan kepada banyak peserta didik, serta memungkinkan proses penskoran dilakukan secara cepat dan konsisten. Namun demikian, penyusunan soal berdasarkan validasi isi atau telaah ahli saja belum menjamin bahwa setiap butir soal mampu mengukur kemampuan siswa secara optimal. Sebuah tes dapat memiliki reliabilitas yang cukup baik, tetapi masih mengandung butir soal yang terlalu mudah, tidak mampu membedakan siswa berkemampuan tinggi dan rendah, atau memiliki pengecoh yang

tidak berfungsi. Kondisi tersebut menyebabkan tes kehilangan fungsi diagnostiknya sehingga kurang mampu memberikan gambaran nyata mengenai penguasaan konsep peserta didik (Slepkov et al., 2021; DeCarlo, 2023).

Dalam teori tes klasik (*Classical Test Theory*), kualitas suatu instrumen tidak hanya dinilai berdasarkan reliabilitas skor total, tetapi juga melalui karakteristik empiris setiap butir soal, yaitu tingkat kesukaran (*item difficulty*), daya pembeda (*item discrimination*), dan efektivitas pengecoh (*distractor effectiveness*). Analisis terhadap indikator-indikator tersebut memungkinkan guru mengidentifikasi butir soal yang layak dipertahankan, direvisi, atau dieliminasi sehingga instrumen penilaian dapat terus disempurnakan. Analisis butir juga menjadi salah satu prosedur yang direkomendasikan dalam pengembangan instrumen karena mampu meningkatkan ketepatan pengukuran dan kualitas keputusan pembelajaran yang dihasilkan

(DeCarlo, 2023; Nurjanah et al., 2024).

Berbagai penelitian menunjukkan bahwa masih banyak instrumen evaluasi yang digunakan di sekolah belum memenuhi karakteristik psikometrik yang baik. Nurjanah et al. (2024) melaporkan bahwa analisis kualitas tes menggunakan pendekatan *Classical Test Theory* mampu mengidentifikasi kelemahan instrumen yang sebelumnya tidak terlihat hanya melalui skor total. Demikian pula, Hermanto et al. (2025) menunjukkan bahwa analisis butir soal memberikan informasi penting mengenai kualitas kemampuan kognitif yang diukur sehingga dapat menjadi dasar revisi instrumen evaluasi. Temuan-temuan tersebut menunjukkan bahwa reliabilitas tes saja belum cukup untuk menyimpulkan bahwa suatu instrumen telah memiliki kualitas yang baik.

Pada praktiknya, evaluasi pembelajaran di sekolah masih lebih berorientasi pada perolehan nilai akhir siswa dibandingkan evaluasi terhadap kualitas instrumen yang

digunakan. Setelah pelaksanaan tes, guru umumnya langsung memanfaatkan skor sebagai dasar penilaian tanpa melakukan analisis terhadap karakteristik setiap butir soal. Akibatnya, butir-butir yang terlalu mudah, memiliki daya pembeda rendah, ataupun pengecoh yang tidak berfungsi tetap digunakan pada evaluasi berikutnya. Padahal, analisis butir merupakan bagian penting dari evaluasi pembelajaran karena mampu memberikan umpan balik bagi guru untuk memperbaiki kualitas instrumen secara berkelanjutan dan meningkatkan akurasi hasil pengukuran (Implementasi Classic Test dan Item Response Theory, 2023; Yulisharyasti et al., 2023).

Materi larutan penyangga merupakan salah satu topik penting dalam pembelajaran kimia yang menuntut pemahaman konseptual serta kemampuan menerapkan konsep dalam berbagai situasi ilmiah. Kompleksitas materi ini mengharuskan instrumen evaluasi mampu membedakan tingkat penguasaan konsep peserta didik

secara tepat. Oleh karena itu, kualitas setiap butir soal perlu dievaluasi secara empiris agar instrumen yang digunakan benar-benar mampu mengukur kemampuan siswa sesuai dengan tujuan pembelajaran.

Berdasarkan uraian tersebut, penelitian ini bertujuan mengevaluasi kualitas tes pilihan ganda akhir bab pada materi larutan penyangga menggunakan pendekatan *Classical Test Theory* melalui analisis reliabilitas, tingkat kesukaran, daya pembeda, dan efektivitas pengecoh dengan bantuan perangkat lunak ANATES. Hasil penelitian diharapkan memberikan informasi empiris mengenai kualitas instrumen yang digunakan sekaligus menjadi dasar perbaikan butir soal sehingga penilaian pembelajaran kimia dapat menghasilkan informasi yang lebih valid, reliabel, dan bermanfaat dalam mendukung pengambilan keputusan pembelajaran.

B. Metode Penelitian

Penelitian ini menggunakan desain kuantitatif deskriptif yang berlandaskan pada teori tes klasik.

Desain ini sesuai karena penelitian bertujuan untuk mendeskripsikan kualitas empiris dari instrumen penilaian kelas, bukan untuk menguji hubungan sebab-akibat. Objek analisis adalah tes pilihan ganda yang terdiri dari 25 soal pada materi larutan penyangga. Setiap soal diberi skor dua poin, sehingga skor maksimum yang dapat diperoleh adalah 100.

Data berasal dari respons 35 siswa yang menyelesaikan tes akhir bab tersebut. Hasil ANATES dibagi menjadi kelompok siswa atas dan bawah yang masing-masing terdiri dari sembilan siswa, untuk analisis diskriminasi. ANATES digunakan sebagai alat analisis utama. Prosedur terdiri dari memasukkan respons siswa, menghitung statistik deskriptif tes, memperkirakan reliabilitas, mengidentifikasi kesulitan item, memeriksa daya diskriminasi, meninjau kinerja pengecoh, dan mengintegrasikan indikator-indikator ini untuk keputusan per item. Reliabilitas diuji melalui koefisien yang dilaporkan dan korelasi ganjil-

genap.

C. Hasil Penelitian dan Pembahasan

Hasil penelitian menunjukkan bahwa koefisien reliabilitas tes sebesar 0,72 dan korelasi ganjil-genap sebesar 0,56. Ini menunjukkan konsistensi internal yang dapat diterima untuk penggunaan dasar di kelas. Namun, tes ini memiliki kesulitan yang sangat mudah, di mana seluruh soal termasuk dalam kategori sangat mudah dengan rentang jawaban benar antara 88,57% hingga 100%. Sebanyak 17 item memiliki daya diskriminasi buruk, 5 item sedang, 2 item baik, dan 1 item sangat baik. Analisis pengecoh menunjukkan bahwa sebagian besar opsi salah tidak berfungsi dengan baik karena jarang atau tidak pernah dipilih.

Temuan ini mengindikasikan bahwa reliabilitas skor total saja tidak cukup untuk menilai kualitas instrumen. Oleh karena itu, analisis rutin item perlu dilakukan untuk memastikan kualitas instrumen penilaian.

1. Reliabilitas Tes

Hasil analisis menggunakan ANATES menunjukkan bahwa tes memiliki koefisien reliabilitas sebesar **0,72** dengan korelasi belah dua (*split-half correlation*) sebesar **0,56**. Nilai reliabilitas tersebut menunjukkan bahwa instrumen memiliki konsistensi internal yang berada pada kategori **dapat diterima** (*acceptable*) untuk digunakan sebagai instrumen penilaian kelas. Reliabilitas yang memadai mengindikasikan bahwa skor yang diperoleh peserta didik relatif stabil apabila tes digunakan pada kelompok dengan karakteristik yang sebanding.

Meskipun demikian, reliabilitas yang tinggi tidak serta-merta menunjukkan bahwa seluruh butir soal telah berkualitas baik. Dalam teori tes klasik (*Classical Test Theory*), reliabilitas menggambarkan konsistensi skor secara keseluruhan, sedangkan kualitas setiap butir soal tetap harus dievaluasi melalui indeks kesukaran, daya pembeda, dan efektivitas pengecoh (DeCarlo, 2023). Oleh karena itu, suatu tes dapat memiliki reliabilitas yang memadai,

tetapi tetap mengandung sejumlah butir yang kurang mampu mengukur kemampuan peserta didik secara akurat.

Temuan penelitian ini sejalan dengan penelitian Priyani dan Sugiharto (2024) yang menunjukkan bahwa reliabilitas instrumen dapat berada pada kategori sedang hingga baik, tetapi masih ditemukan sejumlah butir soal yang memiliki kualitas rendah berdasarkan analisis daya pembeda dan pengecoh. Hal tersebut menunjukkan bahwa evaluasi kualitas instrumen tidak cukup hanya berlandaskan pada koefisien reliabilitas, melainkan harus dilengkapi dengan analisis karakteristik setiap butir soal.

2. Tingkat Kesukaran Butir Soal

Hasil analisis menunjukkan bahwa **seluruh 25 butir soal berada pada kategori sangat mudah**, dengan persentase jawaban benar berkisar antara **88,57% hingga 100%**. Temuan ini mengindikasikan bahwa hampir seluruh peserta didik mampu menjawab setiap soal dengan benar sehingga variasi skor antarpeserta didik menjadi sangat

kecil. Dalam teori tes klasik, soal yang terlalu mudah tidak memberikan informasi yang optimal mengenai kemampuan peserta didik karena hampir semua siswa memperoleh skor yang sama. Kondisi tersebut menyebabkan instrumen kehilangan kemampuan untuk membedakan peserta didik yang memiliki kemampuan tinggi dengan peserta didik yang memiliki kemampuan rendah (DeCarlo, 2023). Sebaliknya, instrumen yang baik umumnya memiliki distribusi tingkat kesukaran yang bervariasi sehingga mampu mengukur kemampuan peserta didik pada berbagai tingkat penguasaan materi.

Hasil penelitian ini berbeda dengan penelitian pada instrumen pembelajaran matematika yang melaporkan distribusi tingkat kesukaran lebih proporsional, yaitu terdiri atas kategori sulit, sedang, dan mudah sehingga instrumen mampu memberikan informasi kemampuan siswa secara lebih komprehensif.

Temuan tersebut mengindikasikan bahwa soal larutan penyangga yang dianalisis

kemungkinan lebih banyak mengukur kemampuan mengingat (*remembering*) dibandingkan kemampuan memahami, menerapkan, atau menganalisis konsep. Oleh karena itu, penyusunan soal selanjutnya perlu memperhatikan variasi tingkat kognitif sesuai Taksonomi Bloom Revisi sehingga diperoleh komposisi soal yang lebih seimbang.

3. Daya Pembeda Butir Soal

Analisis daya pembeda menunjukkan bahwa **17 butir (68%)** berada pada kategori **buruk**, **5 butir (20%)** berkategori **sedang**, **2 butir (8%)** berkategori **baik**, dan hanya **1 butir (4%)** berkategori **sangat baik**. Hasil tersebut menunjukkan bahwa sebagian besar butir soal belum mampu membedakan siswa yang memiliki kemampuan tinggi dengan siswa yang berkemampuan rendah.

Daya pembeda merupakan salah satu indikator paling penting dalam analisis butir karena menunjukkan kemampuan suatu soal dalam mengidentifikasi perbedaan kompetensi antar peserta didik. Butir soal dengan daya pembeda rendah

biasanya disebabkan oleh tingkat kesukaran yang terlalu tinggi atau terlalu rendah, adanya kunci jawaban yang kurang tepat, maupun alternatif jawaban yang kurang berfungsi (DeCarlo, 2023).

Rendahnya daya pembeda pada penelitian ini diduga merupakan konsekuensi langsung dari seluruh butir yang berada pada kategori sangat mudah. Ketika hampir semua peserta didik mampu menjawab benar, skor siswa kelompok atas dan kelompok bawah menjadi relatif sama sehingga indeks diskriminasi menjadi rendah. Hubungan antara tingkat kesukaran dengan daya pembeda ini juga dilaporkan dalam berbagai penelitian analisis butir yang menunjukkan bahwa soal dengan tingkat kesukaran ekstrem cenderung memiliki kemampuan diskriminasi yang rendah.

Implikasinya, guru perlu menyusun butir soal yang lebih menantang melalui penggunaan konteks kehidupan nyata, soal berbasis pemecahan masalah, maupun pertanyaan yang mengukur kemampuan berpikir tingkat tinggi

(*Higher Order Thinking Skills/HOTS*) sehingga kemampuan diskriminasi instrumen dapat meningkat.

4. Efektivitas Pengecoh

Hasil analisis menunjukkan bahwa sebagian besar alternatif jawaban salah (*distractors*) tidak berfungsi secara optimal karena jarang bahkan tidak pernah dipilih oleh peserta didik. Kondisi tersebut menunjukkan bahwa pengecoh belum cukup menarik bagi siswa yang belum menguasai materi sehingga peserta didik dengan mudah mengidentifikasi jawaban yang benar.

Pengecoh yang efektif seharusnya dipilih oleh sebagian peserta didik yang belum memahami konsep, sedangkan peserta didik yang memahami materi akan memilih jawaban yang benar. Apabila suatu pengecoh hampir tidak pernah dipilih, maka alternatif tersebut tidak lagi berfungsi sebagai pengecoh dan perlu direvisi. Keberadaan pengecoh yang tidak berfungsi akan menurunkan kualitas butir soal dan berdampak pada rendahnya daya pembeda instrumen. Hasil penelitian ini memperkuat temuan sebelumnya

bahwa efektivitas pengecoh memiliki hubungan erat dengan indeks kesukaran dan daya pembeda. Semakin banyak pengecoh yang berfungsi dengan baik, semakin besar peluang suatu butir mampu membedakan peserta didik berdasarkan tingkat penguasaan materi.

5. Implikasi Hasil Analisis Butir

Berdasarkan integrasi keempat indikator analisis diperoleh bahwa hanya **butir nomor 2, 7, dan 23** yang memenuhi kriteria untuk dipertahankan tanpa revisi. Sebagian besar butir lainnya memerlukan revisi, terutama pada aspek tingkat kesukaran, kualitas pengecoh, dan kemampuan membedakan peserta didik. Temuan ini menegaskan bahwa **reliabilitas tes sebesar 0,72 belum cukup dijadikan dasar untuk menyatakan bahwa instrumen telah berkualitas baik**. Evaluasi instrumen harus dilakukan secara komprehensif dengan mempertimbangkan seluruh indikator analisis butir. Pendekatan ini sejalan dengan rekomendasi dalam teori tes klasik bahwa keputusan penggunaan

suatu instrumen sebaiknya didasarkan pada kombinasi reliabilitas, tingkat kesukaran, daya pembeda, dan efektivitas pengecoh agar instrumen benar-benar mampu menghasilkan informasi yang akurat mengenai kemampuan peserta didik (DeCarlo, 2023).

Secara praktis, hasil penelitian ini memberikan implikasi bahwa guru kimia perlu menjadikan analisis butir sebagai bagian dari evaluasi rutin setelah pelaksanaan tes. Melalui analisis tersebut, bank soal dapat diperbaiki secara berkelanjutan sehingga instrumen yang digunakan pada evaluasi berikutnya memiliki kualitas psikometrik yang lebih baik dan mampu mendukung pengambilan keputusan pembelajaran secara lebih tepat. Hasil ini juga mendukung penelitian sebelumnya yang merekomendasikan penggunaan analisis butir secara sistematis untuk meningkatkan mutu penilaian di kelas.

D. Kesimpulan

Berdasarkan hasil analisis, dapat disimpulkan bahwa tes pilihan

ganda akhir bab pada materi larutan penyangga memiliki konsistensi yang dapat diterima tetapi kualitas item pada level yang lebih rendah. Koefisien reliabilitas sebesar 0,72 menunjukkan bahwa skor total stabil. Namun, seluruh item sangat mudah, sebagian besar item memiliki daya diskriminasi buruk, dan sebagian besar pengecoh tidak berfungsi dengan baik. Sebagai hasilnya, tes ini memiliki kapasitas terbatas untuk membedakan siswa menurut tingkat pemahamannya.

Hanya tiga item, yaitu item 2, 7, dan 23 yang cocok untuk dipertahankan tanpa revisi. Sebagian besar item lainnya perlu direvisi atau dihapus. Penelitian ini menegaskan pentingnya analisis item secara rutin untuk memastikan bahwa instrumen penilaian kelas dapat digunakan dengan efektif untuk mendukung keputusan pembelajaran.

DAFTAR PUSTAKA

- Alenezi, dkk. (2024). *Item analysis: the impact of distractor efficiency on the difficulty index and discrimination power of multiple-choice items*. *BMC Medical Education*
- Allahverdipour, H., Badri, M., Shaghaghi, A., Mahmoodi, H., Heizomi, H., Shirzadi, S., & Jafarabadi, M. A. (2023). Medications Non-Adherence Reasoning Scale (MedNARS): Development and psychometric properties appraisal. *Health Promotion Perspectives*, 13(3), 212-218.
<https://doi.org/10.34172/hpp.2023.26>
- Butz, A. R., & Branchaw, J. (2020). Entering Research Learning Assessment (ERLA): Validity evidence for an instrument to measure undergraduate and graduate research trainee development. *CBE-Life Sciences Education*, 19(2), ar18.
<https://doi.org/10.1187/cbe.19-07-0146>
- Chauhan, G. R., et al. (2023). *Relations of the Number of Functioning Distractors With the Item Difficulty Index and the Item Discrimination Power in the Multiple Choice Questions*. *Cureus*. DOI: <https://doi.org/10.7759/cureus.42492>
- DeCarlo, L. T. (2023). *Classical Item Analysis from a Signal Detection Perspective*. *Journal of Educational Measurement*, 60(3), 520–547.

- <https://doi.org/10.1111/jedm.12358>
- Donnelly, J. P., Finn, K. E., & Stewart, A. D. (2023). Development of a brief college embeddedness scale. *Journal of College Student Retention: Research, Theory & Practice*, 27(4), 995-1010.
<https://doi.org/10.1177/15210251231217094>
- Hermanto, F. Y., Anggraeni, A. D., & Sutirman. (2025). *Analisis Butir Soal dengan Classical Test Theory untuk Mengukur Kemampuan Kognitif Siswa*. Efisiensi: Kajian Ilmu Administrasi.
<https://doi.org/10.21831/efisiensi.v20i2.80225>
- Kubo, S. (2023). *Test Theory Underpinning Learner Assessment*. *Journal of the Japan Society for Medical Education*, 54(4), 367–375.
https://doi.org/10.11307/mededjapan.54.4_367
- Nurjanah, S., Iqbal, M., Zafrullah, Z., Mahmud, M. N., Seran, D. S. F., Suardi, I. K., & Arriza, L. (2024). *Psychometric quality of multiple-choice tests under classical test theory (CTT): AnBuso, Iteman, and R*. *Jurnal Penelitian dan Evaluasi Pendidikan*, 28(2), 161–172.
<https://doi.org/10.21831/pep.v28i2.71542>
- Priyani, T., & Sugiharto, B. (2024). *Analysis of biology midterm exam items using a comparison of the classical theory test and the Rasch model*. *Jurnal Pendidikan Biologi Indonesia*. DOI:
<https://doi.org/10.22219/jpbi.v10i3.34345>
- Slepkov, A. D., Bussel, M. L. V., Fitze, K. M., & Burr, W. S. (2021). A baseline for multiple choice testing in the university classroom. *SAGE Open*, 11(2).
<https://doi.org/10.1177/21582440211016838>
- Syamsuddin, Syaiful Syamsuddin (2023). Implementasi Classic Test dan Item Response Theory Pada Penilaian Tes Pembelajaran Matematika. *EDUSCOPE: Jurnal Pendidikan, Pembelajaran, dan Teknologi*.
<https://doi.org/10.32764/eduscope.v8i2.3488>
- Yulisharyasti, L., Nurdin, A., Aulia, N., Arfa, F. A. H., & Fadryani. (2023). *Analyzing the Quality of Measurement Instruments of Multiple Choice Questions through Classical Test Theory and Rasch Models*.
<https://doi.org/10.22487/27765660.2023.v3.i1.16417>
- Zakiah, Z., & Hidayah, A. (2023). The COVID-19 pandemic and the characteristic comparison of English achievement tests. *Perspectives of Science and*

Education, 62(2), 307-329.
<https://doi.org/10.32744/pse.2023.2.18>