Jurnal Pendidikan Biologi
Universitas Negeri Malang, Jawa Timur, Indonesia
p-ISSN 2085-6873 | e-ISSN 2540-9271
Edisi Agustus 2025, Volume 16, Nomor 2, pp 110-119

OPTIMALISASI ASESMEN BIOLOGI: ANALISIS KOMPREHENSIF KUALITAS BUTIR SOAL UNTUK PENINGKATAN TES KOMPETENSI GURU

Frida Kunti Setiowati*, Siti Zubaidah
Sekolah Pascasarjana Universitas Negeri Malang, Jl. Semarang No. 5 Malang, Jawa Timur, Indonesia
* corresponding author | email : frida.fmipa@um.ac.id

Received: 20 Juli 2025 *Accepted: 28 Agustus 2025* *Published: 31 Agustus 2025*

ABSTRAK

Penelitian ini bertujuan untuk menganalisis kualitas instrumen tes kompetensi guru biologi melalui uji validitas butir, daya beda, tingkat kesukaran, efektivitas distraktor, dan reliabilitas. Subjek penelitian adalah 54 mahasiswa Pendidikan Profesi Guru (PPG) Biologi di Universitas Negeri Malang yang mengerjakan tes berjumlah 35 butir soal pilihan ganda dengan lima opsi jawaban. Analisis data dilakukan menggunakan pendekatan teori tes klasik. Hasil penelitian mengonfirmasi bahwa dari 35 butir soal, 22 memenuhi kriteria validitas (CITC > 0,30), sedangkan 13 butir dinyatakan tidak valid. Analisis daya beda memperlihatkan mayoritas soal berkategori jelek (27 butir), dan 8 cukup. Tingkat kesukaran didominasi soal mudah (25 butir), dengan 5 soal sedang dan 2 soal sulit. Analisis distraktor menunjukkan 14 soal memiliki pengecoh berfungsi dan 21 lainnya tidak berfungsi. Uji reliabilitas menghasilkan koefisien Cronbach's Alpha sebesar 0,900, yang menunjukkan konsistensi internal sangat tinggi. Berdasarkan hasil tersebut, 27 butir digunakan kembali setelah revisi, sedangkan 8 butir digugurkan. Penelitian ini menegaskan pentingnya analisis butir soal untuk menjamin validitas dan reliabilitas instrumen, serta memberikan dasar pengembangan instrumen tes kompetensi guru biologi yang lebih representatif, sehingga dapat mendukung peningkatan kualitas asesmen dan profesionalisme guru.

doi http://dx.doi.org/10.17977/um052v16i2p110-119

Kata Kunci : *analisis butir soal, tes kompetensi, guru biologi, validitas, reliabilitas*

This study aimed to analyze the quality of a biology teacher competency test instrument by examining item validity, discriminating power, difficulty level, distractor effectiveness, and reliability. The research subjects were 54 Biology Teacher Professional Education (PPG) students at Universitas Negeri Malang, who completed a 35-item multiple-choice test designed with five answer options. Data analysis was conducted using the classical test theory approach. The results indicated that out of 35 items, 22 met the validity criteria (CITC > 0.30), while 13 items were declared invalid. The discriminating power analysis showed that the majority of items fell into the poor category (27 items), with 8 items being rated as sufficient. Regarding difficulty level, the test was predominantly easy (25 items), with 5 moderate and 2 difficult items. Distractor analysis revealed that 14 items had functioning distractors, whereas the other 21 did not. The reliability test yielded a Cronbach's Alpha coefficient of 0.900, which indicated very high internal consistency. Based on these findings, 27 items were reused after revision, while 8 items were discarded. This study underscores the importance of item analysis in ensuring instrument validity and reliability. Furthermore, it provides a crucial foundation for the development of more representative biology teacher competency test instruments, thereby supporting the continuous improvement of assessment quality and teacher professionalism.

Keywords : *item analysis, competency test, biology teacher, validity, reliability*

Evaluasi hasil belajar memiliki peran strategis dalam menjamin mutu pendidikan karena menjadi dasar untuk menilai sejauh mana tujuan pembelajaran tercapai dan kualitas pendidik dapat ditingkatkan (Schoepp, 2019; Alhazmi, 2025). Melalui evaluasi yang sistematis, lembaga pendidikan

dapat memperoleh informasi yang akurat mengenai efektivitas proses pembelajaran serta kompetensi tenaga pendidik dalam menjalankan perannya. Dalam konteks pendidikan biologi, asesmen tidak hanya berfungsi untuk mengukur capaian peserta didik, tetapi juga memiliki makna penting dalam memetakan kompetensi guru biologi sebagai faktor penentu keberhasilan pembelajaran di kelas. Guru biologi dituntut untuk tidak hanya menguasai konsep-konsep biologi secara mendalam, tetapi juga mampu menerapkan pendekatan pedagogis yang adaptif terhadap karakteristik peserta didik dan kompleksitas materi yang diajarkan (Nurdianti, 2017; Mulvidatin & Kurniawati, 2024). Oleh karena itu, asesmen terhadap kompetensi guru menjadi bagian integral dari upaya peningkatan profesionalisme pendidik dan mutu pembelajaran biologi.

Salah satu instrumen yang digunakan untuk memastikan penguasaan kompetensi guru adalah tes kompetensi guru biologi, yang dirancang untuk menilai kemampuan guru pada aspek profesional dan pedagogik. Tes ini tidak hanya berfungsi sebagai alat ukur kemampuan individual, tetapi juga sebagai dasar dalam merancang program pengembangan profesionalisme guru, seperti Pendidikan Profesi Guru (PPG) atau pelatihan berbasis Musyawarah Guru Mata Pelajaran (MGMP) (Kabieva dkk., 2014; Rohayati dkk., 2018). Melalui analisis butir soal tes kompetensi, penyelenggara pendidikan dapat mengidentifikasi kelemahan dalam instrumen maupun dalam pemahaman konseptual guru. Dengan demikian, hasil evaluasi dapat dimanfaatkan untuk memperbaiki proses pelatihan, menyusun kurikulum PPG yang lebih tepat sasaran, serta memberikan umpan balik terhadap kebutuhan peningkatan kapasitas guru biologi di lapangan.

Meskipun memiliki peran penting, kualitas instrumen tes kompetensi guru di lapangan masih sering menjadi persoalan. Beberapa penelitian mengindikasikan banyak butir soal belum melalui proses analisis kualitas yang komprehensif, seperti uji validitas, reliabilitas, dan efektivitas pengecoh (Astorga & González-Carrasco, 2025; Sudarso dkk., 2023). Permasalahan umum yang ditemukan antara lain daya beda yang rendah, distribusi tingkat kesukaran yang tidak seimbang, serta opsi jawaban salah (pengecoh) yang tidak berfungsi dengan baik sehingga gagal membedakan kemampuan guru berkemampuan tinggi dan rendah. Kondisi tersebut dapat menurunkan keakuratan hasil evaluasi dan berimplikasi pada pengambilan keputusan yang kurang tepat, misalnya dalam menentukan kelayakan guru mengikuti sertifikasi atau pelatihan lanjutan.

Lebih jauh, kajian empiris mengenai analisis kualitas butir soal tes kompetensi pedagogik guru biologi masih sangat terbatas. Sebagian besar penelitian terdahulu lebih menekankan pada evaluasi capaian belajar atau pengembangan instrumen asesmen peserta didik (Caspersen dkk., 2017; Goss, 2022). Sementara itu, evaluasi terhadap instrumen yang digunakan untuk mengukur kompetensi guru, khususnya pada bidang biologi, belum dilakukan secara sistematis dan komprehensif (Cortes dkk., 2022). Padahal, kualitas butir soal memiliki pengaruh langsung terhadap keabsahan hasil tes dan kredibilitas sistem asesmen guru secara keseluruhan. Analisis yang mendalam terhadap butir soal dapat mengungkap sejauh mana instrumen benar-benar merepresentasikan kompetensi yang diukur, serta membantu memastikan bahwa tes berjalan secara objektif, adil, dan proporsional terhadap tingkat kemampuan peserta.

Menjawab kesenjangan tersebut, penelitian ini bertujuan untuk menganalisis kualitas butir soal tes kompetensi pedagogik guru biologi yang dikembangkan oleh tim pengajar PPG Biologi Universitas Negeri Malang. Analisis dilakukan dengan meninjau beberapa aspek utama, meliputi validitas butir, reliabilitas tes, tingkat kesukaran, daya beda, dan efektivitas pengecoh. Melalui pendekatan ini, diharapkan dapat diperoleh gambaran empiris mengenai kekuatan dan kelemahan instrumen yang digunakan dalam menilai kompetensi guru biologi. Hasil penelitian ini tidak hanya memberikan kontribusi teoretis dalam bidang evaluasi pendidikan, tetapi juga kontribusi praktis bagi lembaga penyelenggara PPG dan MGMP Biologi dalam merancang instrumen asesmen yang lebih berkualitas.

METODE

Penelitian ini menggunakan desain kuantitatif deskriptif dengan tujuan menganalisis kualitas instrumen tes kompetensi guru biologi. Subjek penelitian terdiri atas 54 orang mahasiswa Pendidikan Profesi Guru (PPG) bidang Biologi di Universitas Negeri Malang yang mengikuti tes kompetensi sebagai bagian dari program akademik mereka. Instrumen penelitian berupa tes kompetensi pedagogik guru biologi yang dikembangkan oleh tim PPG Biologi UM. Tes dikembangkan dalam bentuk pilihan ganda dengan 35 butir soal dan lima opsi jawaban pada setiap butir, termasuk satu jawaban benar dan empat

pengecoh. Struktur instrumen ini dipilih karena dapat mengukur secara lebih objektif penguasaan konsep biologi serta keterampilan pedagogik calon guru, sekaligus memungkinkan analisis mendalam terkait validitas, daya beda, tingkat kesukaran, efektivitas pengecoh, dan reliabilitas tes. Seluruh analisis dilakukan dengan bantuan perangkat lunak statistik SPSS versi 27 dan Microsoft Excel 2019.

Analisis data dilakukan untuk mengevaluasi kualitas butir soal dalam instrumen tes kompetensi guru biologi. Validitas isi (*content validity*) terlebih dahulu ditelaah oleh dua ahli pendidikan biologi untuk memastikan kesesuaian butir dengan indikator kompetensi dan konteks pembelajaran. Selanjutnya, validitas butir dianalisis secara empiris menggunakan *Corrected Item-Total Correlation* (CITC), yaitu korelasi antara skor butir dengan skor total tes (dikurangi skor butir tersebut). Item dinyatakan valid apabila nilai CITC ≥ 0,30, sedangkan item dengan CITC < 0,30 dianggap tidak valid karena kontribusinya terhadap konstruk yang diukur sangat rendah. Jika nilai CITC tinggi (≥ 0,30), berarti butir tersebut konsisten dengan konstruk yang diukur oleh keseluruhan tes. Sebaliknya, nilai rendah menunjukkan butir itu mungkin tidak sesuai atau mengukur hal lain (Suseno, 2014; Rachmawati & Pradana, 2025).

Daya beda soal dihitung untuk menilai sejauh mana butir mampu membedakan peserta dengan kemampuan tinggi dan rendah. Indeks daya beda diperoleh dari selisih proporsi peserta yang menjawab benar pada kedua kelompok tersebut dengan rumus $D = P_U - P_L$, di mana P_U adalah proporsi peserta kelompok atas yang menjawab benar dan P_L adalah proporsi peserta kelompok bawah yang menjawab benar. Daya beda dikategorikan berdasarkan kriteria Ebel & Frisbie (1991), yaitu: <0,20 (buruk), 0,20–0,39 (cukup), 0,40–0,69 (baik), dan ≥0,70 (sangat baik). Butir dengan daya beda ≥0,20 dinyatakan layak digunakan (Arikunto, 2016). Nilai daya beda yang tinggi menjadi indikasi soal efektif dalam mengklasifikasikan responden berdasarkan tingkat kompetensinya, sedangkan nilai yang mendekati nol mengindikasikan soal tidak dapat membedakan kedua kelompok.

Tingkat kesukaran dihitung berdasarkan proporsi peserta yang menjawab benar pada setiap butir. Soal dengan indeks kesukaran mendekati 1,00 termasuk kategori sangat mudah, sedangkan yang mendekati 0,00 tergolong sangat sulit. Instrumen yang baik idealnya memiliki variasi soal dengan tingkat kesukaran sedang (Son, 2019). Kriteria tingkat kesukaran ditentukan berdasarkan proporsi jawaban benar (p), dengan kategori: p < 0,30 (sulit), 0,30 ≤ p ≤ 0,70 (sedang), dan p > 0,70 (mudah) (Arikunto, 2016; Ebel & Frisbie, 1991). Selain itu, efektivitas distraktor dianalisis untuk menilai sejauh mana opsi jawaban salah berfungsi sebagai pengecoh. Distraktor dinyatakan efektif apabila dipilih oleh minimal 5% responden dan menunjukkan pola pilihan lebih tinggi pada peserta berkemampuan rendah (Haladyna dkk., 2002). Distraktor yang tidak pernah dipilih atau dipilih <5% responden dianggap tidak berfungsi.

Reliabilitas instrumen kemudian diuji menggunakan koefisien Cronbach's Alpha untuk menilai konsistensi internal antarbutir. Nilai alpha ≥ 0,70 menunjukkan instrumen reliabel dan dapat dipercaya. Melalui analisis validitas, reliabilitas, daya beda, tingkat kesukaran, serta efektivitas distraktor, kualitas instrumen dapat ditentukan sekaligus diperoleh rekomendasi perbaikan agar instrumen lebih valid, reliabel, dan representatif dalam mengukur kompetensi guru biologi. Proses revisi dilakukan secara konseptual berdasarkan hasil analisis tanpa dilakukan uji ulang terhadap butir yang telah direvisi.

HASIL DAN PEMBAHASAN

HASIL

Validitas Butir Soal

Berdasarkan hasil analisis validitas butir soal dengan menggunakan *Corrected Item-Total Correlation* (CITC) dan batas kriteria sebesar 0,3, diperoleh bahwa dari 35 butir soal terdapat 22 butir yang memenuhi kriteria validitas (CITC > 0,3), yaitu item nomor 1, 4, 5, 7, 8, 10, 11, 13, 14, 15, 16, 17, 18, 19, 20, 21, 25, 27, 28, 31, 32, 33, 34 (Tabel 1). Sementara itu, sebanyak 13 butir soal memiliki koefisien CITC di bawah 0,3 sehingga dinyatakan tidak valid, yaitu item nomor 2, 3, 6, 9, 12, 22, 23, 24, 26, 29, 30, dan 35. Hal ini mengonfirmasi sebagian besar soal sudah mampu mengukur konstruk yang

dimaksud secara konsisten, namun masih ada sejumlah butir yang perlu direvisi atau diganti agar instrumen tes dapat mencapai kualitas yang lebih baik.

Tabel 1. Hasil Analisis Validitas Menggunakan CITC

Cut-off CICT	Nomor Item	Interpretasi
>0.3	1, 4, 5, 7, 8, 10, 11, 13, 14, 15, 16, 17, 18, 19, 20, 21, 25, 27, 28, 31, 32, 33, 34	Valid
<0.3	2, 3, 6, 9, 12, 22, 23, 24, 26, 29, 30, 35	Tidak Valid

Daya Beda

Hasil analisis daya beda memperlihatkan bahwa tidak ada butir soal yang memiliki daya pembeda baik (D > 0,40). Sebanyak 8 butir soal berada pada kategori cukup (0,20 ≤ D < 0,40), yaitu item nomor 3, 6, 8, 15, 21, 29, 30, dan 32. Sementara itu, mayoritas butir soal, yaitu 27 item, masuk kategori jelek (D < 0,20), antara lain item nomor 1, 2, 4, 5, 7, 9, 10, 11, 12, 13, 14, 16, 17, 18, 19, 20, 22, 23, 24, 25, 26, 27, 28, 31, 33, 34, dan 35 (Tabel 2). Temuan ini mengindikasikan bahwa sebagian besar butir soal belum mampu membedakan dengan baik antara peserta didik yang berkemampuan tinggi dengan peserta didik yang berkemampuan rendah, sehingga perlu dilakukan revisi terutama pada butir-butir yang memiliki daya beda rendah agar instrumen tes lebih efektif dalam mengukur kemampuan secara akurat.

Tabel 2. Hasil Analisis Daya Beda Butir Soal dan Kriterianya

Daya beda (D)	Nomor Soal	Interpretasi
D > 0.40	-	Baik
0.20 ≤ D < 0.40	3, 6, 8, 15, 21, 29, 30, 32	Cukup
D < 0.20	1, 2, 4, 5, 7, 9, 10, 11, 12, 13, 14, 16, 17, 18, 19, 20, 22, 23, 24, 25, 26, 27, 28, 31, 33, 34, 35	Jelek

Tingkat Kesukaran

Hasil analisis tingkat kesukaran menyajikan bukti bahwa sebagian besar butir soal termasuk dalam kategori mudah (p ≥ 0,70), yaitu sebanyak 25 item, meliputi nomor 1, 2, 3, 4, 5, 7, 8, 10, 11, 12, 13, 14, 16, 17, 18, 19, 20, 21, 22, 25, 26, 27, 28, 30, dan 31. Sementara itu, terdapat 5 butir soal yang berkategori sedang (0,30 ≤ p < 0,70), yakni item nomor 6, 15, 23, 29, dan 32. Adapun hanya 2 butir soal yang tergolong sulit (p < 0,30), yaitu item nomor 9 dan 24 (Tabel 3). Pola distribusi ini mengindikasikan bahwa dominasi soal kategori mudah berpotensi kurang optimal dalam menilai variasi kemampuan peserta, sehingga diperlukan penyeimbangan dengan menambah jumlah soal kategori sedang dan sulit agar instrumen tes dapat mengukur kompetensi guru biologi secara lebih proporsional.

Tabel 3. Hasil Tingkat Kesukaran Butir Soal

Nilai p	Nomor Soal	Kategori
p ≥ 0.70	1, 2, 3, 4, 5, 7, 8, 10, 11, 12, 13, 14, 16, 17, 18, 19, 20, 21, 22, 25, 26, 27, 28, 30, 31	Mudah
0.30 ≤ p < 0.70	6, 15, 23, 29, 32	Sedang
p < 0.30	9, 24	Sulit

Efektivitas Pengecoh

Berdasarkan analisis efektivitas distraktor, sebanyak 14 butir memiliki distraktor yang berfungsi dengan baik, sedangkan 21 butir lainnya memiliki distraktor yang tidak berfungsi. Distraktor dinyatakan berfungsi apabila mampu menarik minimal 5% peserta tes untuk memilihnya dan tersebar secara relatif merata pada kelompok peserta dengan kemampuan rendah maupun sedang (Nurgiyantoro, 2010). Sebaliknya, distraktor yang tidak berfungsi terlihat dari tidak adanya responden yang memilihnya atau pilihan hanya dipilih dalam jumlah sangat kecil, sehingga tidak efektif dalam mengalihkan peserta yang kurang menguasai materi dari jawaban benar. Temuan ini memperlihatkan

bahwa masih terdapat lebih dari separuh butir soal yang memerlukan revisi, khususnya dalam penyusunan opsi jawaban pengecoh agar dapat meningkatkan kualitas tes secara keseluruhan.

Butir Soal Gugur dan Digunakan Setelah Revisi

Pengambilan keputusan terkait penggunaan butir soal dalam penelitian ini didasarkan pada hasil analisis validitas, daya beda, tingkat kesukaran, efektivitas distraktor, serta reliabilitas instrumen secara keseluruhan. Butir soal dengan validitas ≥ 0,30 dan tingkat kesukaran yang masih dalam kategori ideal (mudah, sedang, atau sulit) dipertahankan untuk digunakan, baik tanpa revisi maupun dengan revisi pada bagian tertentu. Namun, hasil analisis daya beda dan efektivitas distraktor juga dipertimbangkan sebagai dasar untuk menentukan apakah butir perlu direvisi. Apabila suatu butir memiliki skor validitas mendekati batas minimal, misalnya antara 0,24–0,299, maka butir tersebut tetap dipertimbangkan untuk dipakai dengan revisi terlebih dahulu, baik pada pokok soal maupun pilihan jawaban, dengan tetap memperhatikan keseimbangan keterwakilan indikator soal.

Berdasarkan hasil analisis, sebanyak delapan butir soal (nomor 2, 9, 12, 22, 23, 24, 26, dan 35) dinyatakan gugur karena tidak memenuhi kriteria kelayakan. Sebanyak 27 butir lainnya (nomor 1, 3–8, 10–21, 25, dan 27–34) dipertahankan dengan beberapa revisi sesuai hasil uji validitas, daya beda, tingkat kesukaran, dan efektivitas distraktor. Butir-butir yang telah direvisi kemudian dirakit kembali menjadi instrumen tes kompetensi guru biologi yang lebih representatif. Adapun analisis reliabilitas digunakan untuk menilai konsistensi internal instrumen secara keseluruhan, guna memastikan bahwa seluruh butir mengukur konstruk yang sama secara konsisten.

Uji reliabilitas instrumen tes kompetensi guru biologi dengan 27 butir soal yang digunakan setelah seleksi menunjukkan nilai koefisien Cronbach's Alpha sebesar 0,900. Nilai ini berada pada kategori sangat tinggi (≥ 0,90), yang mengindikasikan bahwa instrumen memiliki konsistensi internal yang sangat baik. Artinya, butir-butir soal yang tersisa mampu memberikan hasil pengukuran yang stabil dan dapat diandalkan dalam menilai kompetensi guru biologi. Dengan demikian, instrumen yang dikembangkan dapat dipertanggungjawabkan sebagai alat evaluasi yang reliabel dan layak digunakan untuk tujuan penelitian maupun praktik evaluasi pendidikan.

PEMBAHASAN

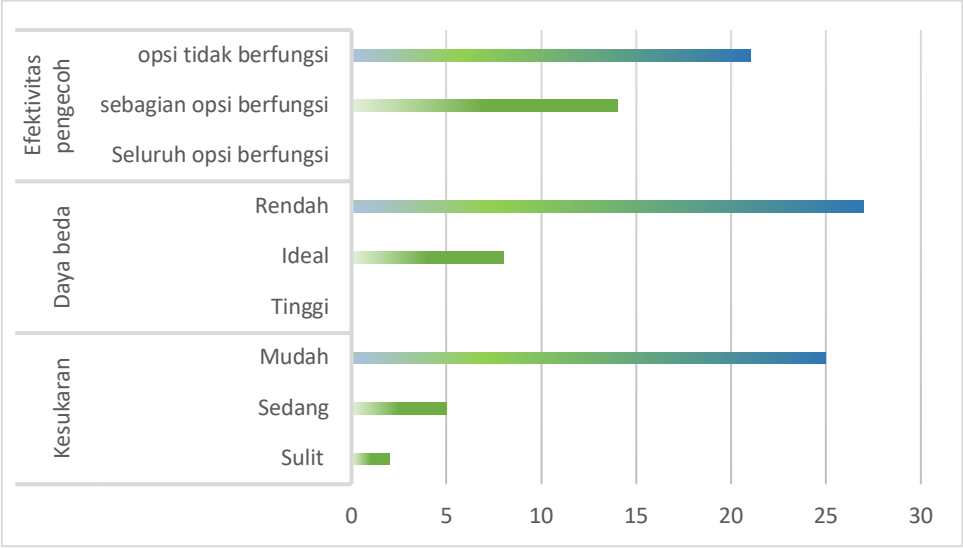
Hasil penelitian menunjukkan, dari 35 butir soal yang dianalisis, sebanyak 22 butir memenuhi kriteria validitas dengan *Corrected Item-Total Correlation* (CITC) ≥ 0,30, sementara 13 butir lainnya dinyatakan tidak valid karena nilai CITC < 0,30 (Askin dkk., 2020). Secara psikometrik, nilai CITC yang tinggi menandakan bahwa butir soal memiliki korelasi yang kuat dengan konstruk keseluruhan yang diukur, sehingga konsisten dengan kompetensi pedagogik yang menjadi sasaran tes (Zijlmans dkk., 2019). Sebaliknya, butir dengan CITC rendah menunjukkan ketidaksesuaian antara isi soal dengan konstruk atau indikator kompetensi yang hendak diukur (Matondang, 2009). Dalam konteks praktis, hal ini dapat disebabkan oleh redaksi soal yang terlalu terminologis atau menekankan hafalan konsep, bukan pada penerapan atau penalaran (Chou dkk., 2013). Kondisi tersebut menyebabkan butir gagal merepresentasikan kemampuan berpikir tingkat tinggi (HOTS) yang menjadi tuntutan kompetensi guru abad ke-21 (Nurdianti, 2017). Oleh karena itu, butir dengan validitas rendah perlu direvisi agar lebih selaras dengan indikator kompetensi dan konteks pembelajaran biologi yang sesungguhnya.

Analisis daya beda mengkonfirmasi tidak ada satu pun butir yang memiliki daya beda baik (D > 0,40), delapan butir tergolong cukup (0,20 ≤ D < 0,40), dan 27 butir berada pada kategori jelek (D < 0,20). Temuan ini menegaskan sebagian besar soal belum mampu membedakan peserta berkemampuan tinggi dan rendah. Secara konseptual, daya beda yang rendah dapat mengindikasikan bahwa butir terlalu mudah, terlalu sulit, atau ambigu, sehingga semua peserta menjawab dengan pola serupa (Iñarrairaegui et al., 2022; Dodeen, 2004). Dalam konteks asesmen guru biologi, rendahnya daya beda dapat pula disebabkan oleh kesamaan pengalaman akademik para peserta PPG yang mengikuti pelatihan serupa, serta pola soal yang menekankan pengetahuan deklaratif (Anderson & Krathwohl, 2001). Padahal, kompetensi guru biologi mencakup keterampilan mengaitkan konsep, proses ilmiah, dan aplikasi kontekstual (Malahayati dkk., 2015). Oleh karena itu, butir dengan daya

beda rendah perlu diperbaiki melalui perumusan ulang indikator agar menuntut analisis dan penerapan konsep, bukan sekadar recall informasi.

Analisis tingkat kesukaran menunjukkan distribusi yang tidak proporsional. Dari 35 butir soal, 25 tergolong mudah ($p \geq 0{,}70$), 5 sedang ($0{,}30 \leq p < 0{,}70$), dan hanya 2 sulit ($p < 0{,}30$). Dominasi soal mudah dapat mengurangi variasi skor, sehingga menghambat kemampuan tes dalam membedakan tingkat kompetensi antarindividu (Pratama dkk., 2021). Secara teoretis, distribusi tingkat kesukaran yang terlalu miring berdampak pada penurunan varians skor total, yang pada gilirannya dapat memengaruhi estimasi reliabilitas tes (Santen dkk., 2021). Tes yang baik umumnya memiliki proporsi terbesar pada kategori sedang karena memberikan daya beda paling optimal (Ferrando, 2012). Dalam konteks PPG Biologi, dominasi soal mudah bisa terjadi karena perancang instrumen berfokus pada penguasaan dasar pedagogik dan biologi umum, bukan pada penerapan integratif yang lebih menantang. Perbaikan sebaiknya diarahkan pada penyeimbangan tingkat kesukaran dengan menambah butir kategori sedang dan sulit yang mengukur kemampuan analisis, refleksi pedagogik, serta penerapan konsep biologi dalam situasi pembelajaran nyata.

Berdasarkan analisis efektivitas distraktor, dari 35 butir soal hanya 14 yang memiliki distraktor berfungsi, sedangkan 21 lainnya tidak. Distraktor dinyatakan berfungsi apabila dipilih oleh minimal 5% responden, khususnya dari kelompok berkemampuan rendah (Rahma dkk., 2017). Distraktor yang tidak berfungsi berarti, opsi jawaban tersebut tidak cukup *plausible* atau terlalu jelas salah sehingga tidak menantang bagi peserta (Haladyna dkk., 2002). Dalam konteks praktis, hal ini dapat terjadi karena distraktor tidak mewakili miskonsepsi umum yang sering muncul. Distraktor yang dirancang berdasarkan miskonsepsi nyata peserta akan lebih efektif dalam mengukur pemahaman konseptual dan meningkatkan daya beda soal (Hrnjicic dkk., 2022). Oleh karena itu, revisi butir perlu difokuskan pada penyusunan distraktor yang relevan dengan kesalahan konseptual yang umum ditemukan pada calon guru biologi.



Gambar 1. Hasil Analisis Butir Soal

Hasil uji reliabilitas memperlihatkan koefisien Cronbach's Alpha sebesar 0,90 untuk 27 butir yang dipertahankan, termasuk kategori sangat tinggi (Tighe dkk., 2010). Nilai ini menunjukkan konsistensi internal yang baik, namun juga bisa mengindikasikan homogenitas butir. Soal yang mengukur aspek sempit menghasilkan variasi jawaban rendah, sehingga tidak selalu mencerminkan representasi konstruk yang luas (Pratama dkk., 2021). Dengan demikian, reliabilitas harus dipahami bersama dengan validitas dan daya beda sebagai satu kesatuan indikator kualitas instrumen. Berdasarkan analisis keseluruhan, 27 butir soal diputuskan untuk dipertahankan dengan revisi, sementara 8 butir digugurkan karena tidak memenuhi kriteria validitas, daya beda, atau efektivitas distraktor. Keputusan ini sejalan dengan prinsip pengembangan tes yang menekankan keseimbangan antara validitas isi, daya pembeda, dan keandalan konstruk (Kaper dkk., 2015).

Hasil penelitian ini menegaskan pentingnya analisis butir soal dalam pengembangan asesmen guru biologi. Instrumen yang valid, memiliki daya beda baik, dan distraktor berfungsi akan memberikan informasi diagnostik yang akurat, baik untuk evaluasi formatif maupun sumatif dalam program PPG (Judijanto dkk., 2025). Secara kontekstual, temuan ini mengindikasikan perlunya penyusunan soal yang mengintegrasikan pendekatan pedagogik reflektif agar penilaian tidak hanya mengukur hafalan, tetapi juga pemahaman konseptual dan kemampuan aplikatif guru. Implikasi kebijakan dari penelitian ini adalah perlunya standardisasi instrumen asesmen guru berbasis kualitas psikometrik, serta pelatihan penyusun soal agar memahami prinsip validitas, reliabilitas, dan daya beda dalam konteks pendidikan biologi (Rachmaningtyas dkk., 2025). Dengan demikian, instrumen tes kompetensi guru dapat berfungsi tidak hanya sebagai alat seleksi, tetapi juga sebagai sarana pembinaan profesionalisme dan peningkatan mutu pendidikan biologi secara berkelanjutan.

KESIMPULAN DAN SARAN

Berdasarkan hasil analisis, kualitas instrumen tes kompetensi guru biologi yang dikembangkan masih memerlukan penyempurnaan pada beberapa aspek psikometrik. Dari 35 butir soal yang dianalisis, sebanyak 22 butir memenuhi kriteria validitas dengan nilai *Corrected Item-Total Correlation* (CITC) ≥ 0,30, sedangkan 13 butir lainnya dinyatakan tidak valid. Analisis daya beda menunjukkan bahwa mayoritas butir (27 soal) belum mampu membedakan peserta berkemampuan tinggi dan rendah, sedangkan 8 soal berada pada kategori cukup dan tidak ada yang berkategori baik. Distribusi tingkat kesukaran juga belum proporsional, dengan dominasi soal mudah (25 butir), sementara kategori sedang hanya 5 butir dan sulit 2 butir. Selain itu, 21 butir memiliki distraktor yang tidak berfungsi dengan baik, menunjukkan perlunya perbaikan pada opsi jawaban agar lebih representatif terhadap miskonsepsi umum di bidang biologi. Walaupun demikian, reliabilitas instrumen tergolong sangat tinggi ($\alpha = 0{,}900$), yang menunjukkan konsistensi internal yang baik antarbutir.

Secara praktis, penelitian ini memiliki implikasi penting bagi pengembangan asesmen guru biologi, terutama dalam program Pendidikan Profesi Guru (PPG) dan Musyawarah Guru Mata Pelajaran (MGMP). Instrumen yang telah dianalisis secara psikometrik dapat menjadi dasar bagi penyusunan tes kompetensi yang lebih akurat dalam memetakan kemampuan pedagogik dan konseptual guru biologi. Revisi instrumen perlu diarahkan untuk menyeimbangkan tingkat kesukaran, meningkatkan daya beda, serta menyusun distraktor yang mencerminkan miskonsepsi biologi yang umum dijumpai di kelas. Dengan demikian, hasil tes tidak hanya menggambarkan penguasaan hafalan, tetapi juga kemampuan berpikir analitis dan reflektif sesuai dengan tuntutan kompetensi abad ke-21.

Penelitian ini memiliki keterbatasan ukuran sampel yang relatif kecil (54 peserta PPG) dan karakteristik sampel yang hanya mencakup satu lembaga pendidikan. Kondisi ini dapat memengaruhi generalisasi hasil terhadap populasi guru biologi secara nasional. Selain itu, analisis dilakukan secara kuantitatif tanpa melibatkan telaah kualitatif terhadap kesesuaian isi butir dengan indikator kompetensi guru biologi. Penelitian selanjutnya disarankan untuk melibatkan sampel yang lebih besar dan beragam dari berbagai institusi penyelenggara PPG serta melakukan analisis item response theory (IRT) untuk memperoleh estimasi parameter yang lebih akurat. Kajian kualitatif melalui expert judgment dan think-aloud protocols juga direkomendasikan untuk memperdalam pemahaman tentang bagaimana peserta menafsirkan dan menjawab butir soal, sehingga dapat meningkatkan validitas isi dan konstruksi tes secara komprehensif.

UCAPAN TERIMA KASIH

Penulis mengucapkan terima kasih yang sebesar-besarnya kepada Sekolah Pascasarjana Universitas Negeri Malang (UM) yang telah memberikan dukungan finansial untuk pelaksanaan penelitian ini.

DAFTAR RUJUKAN

Alhazmi, A. K. (2025, March). Bridging Theory and Practice in Outcome-Based Assessment:

Developing a Web Application for Learning Outcome Alignment. In 2025 14th International Conference on Educational and Information Technology (ICEIT) (pp. 356-367). IEEE.

Al-Muqbil, N. S. M. (2024). The impact of a nanotechnology-based training program on the development of digital competencies among high school biology teachers. Journal of Curriculum and Teaching, 13(2), 98-112.

Alyasin, A., Nasser, R., El Hajj, M., & Harb, H. (2023). Assessing Learning Outcomes in Higher Education: From Practice to Systematization. TEM Journal, 12(3).

Anderson, L. W., & Krathwohl, D. R. (Eds.). (2001). *A taxonomy for learning, teaching, and assessing: A revision of Bloom's taxonomy of educational objectives.* Longman.

Andra, Y., Syahputri, N. K., Nisa, S. F., Ananta, D., Amanda, M. D., Azizah, N., ... & Alfathir, M. I. (2024). Tes Standar dan Tes Non Standar. MUDABBIR Journal Research and Education Studies, 4(2), 476-488.

Arikunto, S. (2016). Prosedur Penelitian Suatu Pendekatan Praktik. Jakarta: Rineka Cipta.

Askın, A., Atar, E., Şengül, İ., Tosun, A., Demirdal, Ü., & Elmalı, F. (2020). Validity and reliability of the Turkish version of caregiver self-assessment questionnaire. Disability and Rehabilitation, 42(22), 3250-3255.

Astorga, O. J. M., & González-Carrasco, F. (2025). Evaluative Judgment: A Validation Process to Measure Teachers' Professional Competencies in Learning Assessments. Education Sciences, 15(5), 624.

Atmoko, A. (2020). Pengembangan tes potensi keberhasilan akademik pada mahasiswa program studi Bimbingan dan Konseling. Jurnal Konseling Gusjigang, 6(2), 72–84.

Caspersen, J., Smeby, J. C., & Olaf Aamodt, P. (2017). Measuring learning outcomes. *European journal of education*, *52*(1), 20-30.

Chou, C. L., Bell, J., Chou, C. M., & Chang, A. (2013). Remediation of interpersonal and communication skills. In *Remediation in medical education: A mid-course correction* (pp. 55-66). New York, NY: Springer New York.

Cortes, K. L., Reid, J. W., Fallin, R., Hao, J., Shah, L., Ray, H. E., & Rushton, G. T. (2022). A longitudinal study identifying the characteristics and content knowledge of those seeking certification to teach secondary biology in the United States. *CBE—Life Sciences Education*, *21*(4), ar63.

Destiana, D., Suchyadi, Y., & Anjaswuri, F. (2020). Pengembangan instrumen penilaian untuk meningkatkan kualitas pembelajaran produktif di sekolah dasar. Jurnal Pendidikan Dan Pengajaran Guru Sekolah Dasar (JPPGuseda), 3(2), 119-123.

Dodeen, H. (2004). The relationship between item parameters and item fit. *Journal of Educational Measurement*, *41*(3), 261-270.

Ebel, R. L. & Frisbie, D. A. (1991). Essentials of educational measurement, 5thEdn. Englewood Cliffs: Prentice-Hall.

Efendi, M., Zulhimmah, Z., & Harahap, H. A. (2024). Penerapan Asesmen Formatif Dan Sumatif Dalam Kurikulum Merdeka Di Madrasah Aliyah Swasta Darul Hadits Huta Baringin. Cognoscere: Jurnal Komunikasi Dan Media Pendidikan, 2(2), 64-72.

George, D., & Mallery, P. (2018). Reliability analysis. In IBM SPSS statistics 25 step by step (pp. 249-260). Routledge.

Goss, H. (2022). Student learning outcomes assessment in higher education and in academic libraries: A review of the literature. *The Journal of Academic Librarianship*, *48*(2), 102485.

Hrnjicic, A., Alihodžic, A., Cunjalo, F., & Kamber Hamzic, D. (2022). Development of an Item Bank for Measuring Students' Conceptual Understanding of Real Functions. *European Journal of Science and Mathematics Education*, *10*(4), 455-470.

Iñarrairaegui, M., Fernández-Ros, N., Lucena, F., Landecho, M. F., García, N., Quiroga, J., & Herrero, J. I. (2022). Evaluation of the quality of multiple-choice questions according to the students' academic level. *BMC Medical Education*, *22*(1), 779.

Judijanto, L., Haryani, H., Sari, N., Pranata, A., Mutoharoh, M., Lumbu, A., ... & Wiradika, I. N. I. (2025). *Assessment, Testing dan Evaluasi*. PT. Sonpedia Publishing Indonesia.

Kabieva, S. Z., Mukataeva, Z. M., Toktarbaeva, A. S., Syzdykova, G. K., & Korogod, N. P. (2014). Role of biological disciplines in formation of professional competence of future teacher of biology. Life Science Journal, 11(SPEC. ISSUE 5), 280-284.

Kaper, N. M., Swennen, M. H., van Wijk, A. J., Kalkman, C. J., van Rheenen, N., van der Graaf, Y., & van der Heijden, G. J. (2015). The "evidence-based practice inventory": reliability and validity was demonstrated for a novel instrument to identify barriers and facilitators for Evidence Based Practice in health care. Journal of clinical epidemiology, 68(11), 1261-1269.

Malahayati, E. N., Corebima, A. D., & Zubaidah, S. (2015). Hubungan keterampilan metakognitif dan kemampuan berpikir kritis dengan hasil belajar biologi peserta didik sma dalam pembelajaran problem based learning (PBL). Jurnal Pendidikan Sains Universitas Negeri Malang, 3(4), 178-185.

Matondang, Z. (2009). Validitas dan reliabilitas suatu instrumen penelitian. Jurnal tabularasa, 6(1), 87-97.

Mulvidatin, O., & Kurniawati, L. (2024). Pentingnya Pengetahuan Pedagogis Umum dalam Praktik Instruksional STEM Pada Guru Matematika Dasar Agar Terciptanya Pembelajaran Bermakna. EduSpirit: Jurnal Pendidikan Kolaboratif, 1(1), 307-313.

Nurdianti, R. R. S. (2017). Pengaruh kompetensi profesional Dan kompetensi pedagogik terhadap kinerja guru ekonomi SMA Negeri di Kota Bandung. Jurnal Ilmiah Manajemen dan Bisnis, 18(2), 177-188.

Nurgiyantoro, B. 2010. Penilaian Pembelajaran Bahasa Berbasis Kompetensi. Yogyakarta: BPFE-Yogyakarta

Pratama, D. N., Pratiwi, O. N., & Sutoyo, E. (2021, September). Classification of Questions Based on Difficulty Levels using Support Vector Machine and Naïve Bayes Algorithms for Imbalanced Class. In *2021 4th International Conference of Computer and Informatics Engineering (IC2IE)* (pp. 40-45). IEEE.

Rachmaningtyas, N. A., Firdaus, N., Afendi, A. R., Ramadhanti, D., Halim, A., Raprap, W. P., ... & Fitriana, T. R. (2025). Menjadi Guru Profesional: Strategi Pembelajaran dan Teknik Evaluasi yang Efektif. Star Digital Publishing.

Rahma, N. A., Shamad, M. M., Idris, M. E., Elfaki, O. A., Elfakey, W. E., & Salih, K. M. (2017). Comparison in the quality of distractors in three and four options type of multiple choice questions. *Advances in Medical Education and Practice*, 287-291.

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.

Rohayati, E., Diana, S. W., & Priyandoko, D. (2018, May). Lesson plan profile of senior high school biology teachers in Subang. In Journal of Physics: Conference Series (Vol. 1013, No. 1, p. 012003). IOP Publishing.

Santen, S. A., Ryan, M., Helou, M. A., Richards, A., Perera, R. A., Haley, K., ... & Park, Y. S. (2021). Building reliable and generalizable clerkship competency assessments: Impact of 'hawk-dove'correction. *Medical Teacher*, *43*(12), 1374-1380.

Schoepp, K. (2019). The state of course learning outcomes at leading universities. Studies in Higher Education, 44(4), 615-627.

Son, A. L. (2019). Instrumentasi kemampuan pemecahan masalah matematis: analisis reliabilitas, validitas, tingkat kesukaran dan daya beda butir soal. *Gema wiralodra*, *10*(1), 41-52.

Sudarso, S., Suroto, S., Hartoto, S., & Dinata, V. C. (2023). Kompetensi guru pendidikan jasmani olahraga dan kesehatan dari perspektif masa kerja. Multilateral: Jurnal Pendidikan Jasmani dan Olahraga, 22(1), 41-50.

Suseno, M. N. M. (2014). Pengembangan pengujian validitas isi dan validitas konstrak: Interpretasi hasil pengujian validitas. In *Seminar Nasional Psikometri* (Vol. 1, No. 5).

Tighe, J., McManus, I. C., Dewhurst, N. G., Chis, L., & Mucklow, J. (2010). The standard error of measurement is a more appropriate measure of quality for postgraduate medical assessments

than is reliability: an analysis of MRCP (UK) examinations. *BMC medical education*, *10*(1), 40.

Zijlmans, E. A., Tijmstra, J., Van der Ark, L. A., & Sijtsma, K. (2019). Item-score reliability as a selection tool in test construction. *Frontiers in psychology*, *9*, 2298. https://doi.org/10.3389/fpsyg.2018.02298