Jurnal Pendidikan Biologi
Universitas Negeri Malang, Jawa Timur, Indonesia
p-ISSN 2085-6873 | e-ISSN 2540-9271
Edisi Agustus 2025, Volume 16, Nomor 1, pp 101-109

ANALISIS PSIKOMETRI SOAL PILIHAN GANDA DALAM PROGRAM PENGEMBANGAN PROFESIONAL GURU BIDANG AGRIBISNIS TERNAK

Balqis[1*], Mardhatillah[1], Muh. Syahriridani[2]
[1] Sekolah Pascasarjana Universitas Negeri Malang, Jl. Semarang No. 5 Malang, Jawa Timur, Indonesia
[2] Departemen Biologi, Universitas Negeri Malang, Jl. Semarang No. 5 Malang, Jawa Timur, Indonesia
* corresponding author | email : balqis.fmipa@um.ac.id

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

ABSTRAK

Salah satu cara untuk mendukung pengembangan profesional berkelanjutan adalah dengan menyediakan instrumen yang dapat mengukur kompetensi secara tepat. Instrumen yang digunakan dalam program pengembangan profesional perlu dianalisis kembali untuk memastikan kualitas program. Tujuan dari penelitian ini adalah (1) mendeskripsikan hasil uji empiris instrumen pemahaman pedagogi guru SMK dalam program pengembangan profesional dan (2) mendeskripsikan hubungan antara validitas, kesukaran, dan daya beda soal. Responden yang digunakan adalah guru SMK yang tersebar di seluruh Indonesia sebanya 99 orang. Metode penelitian menggunakan analisis psikometri *clasical test theory* (CTT) yang terdiri dari validitas, daya beda, kesukaran, distraktor, dan reliabilitas. Hasil penelitian menunjukkan bahwa dari 35 butir soal, hanya 17 yang dapat digunakan dengan revisi, 9 di antaranya valid, 14 memiliki daya beda cukup, 10 memiliki tingkat kesukaran sedang, 3 dengan semua pilihan jawaban berfungsi sebagai pengecoh, dan nilai reliabilitas 0,711. Adapun koefisien korelasi antara validitas dan daya beda sebesar 0,747, sedangkan daya beda dan kesukaran sebesar -0,023. Rekomendasi penelitian ini adalah melakukan revisi serta menambah butir instrumen untuk meningkatkan kualitas analisis.

doi http://dx.doi.org/10.17977/um052v16i2p101-109

Kata Kunci : *Analisis item; Classical test theory; Kompetensi guru; Pengembangan profesional; Validitas psikometri*

One of the ways to support continuous professional development is by providing instruments that can accurately measure competence. The instruments used in professional development programs need to be reanalyzed to ensure the quality of the program. The objectives of this study are (1) to describe the empirical testing results of an instrument measuring vocational teachers' pedagogical understanding in a professional development program, and (2) to describe the relationships among item validity, difficulty level, and discrimination index. The respondents were 99 vocational school teachers from various regions across Indonesia. The study employed a psychometric analysis based on Classical Test Theory (CTT), consisting of item validity, discrimination index, difficulty level, distractor effectiveness, and reliability. The findings revealed that out of 35 items, only 17 could be used with revisions; among them, 9 were valid, 14 had adequate discrimination power, 10 were of moderate difficulty, 3 had all options functioning effectively as distractors, and the reliability coefficient was 0.711. The correlation coefficient between validity and discrimination was 0.747, while that between discrimination and difficulty level was -0.023. The study recommends revising and adding more items to improve the quality of the instrument analysis.

Keywords : *Item analysis; Classical Test Theory; Teacher competence; Professional development; Psychometric validity*

Pendidik yang profesional sangat dibutuhkan untuk mendorong kualitas pendidikan (Purwantiningsih & Suharso, 2019). Pendidik profesional berarti memiliki kualifikasi yang dibutuhkan untuk melaksanakan pembelajaran efektif. Profesionalitas pendidik dapat ditingkatkan melalui program pengembangan profesional (Abdallah & Abdallah, 2023). Salah satu kualifikasi yang

ditekankan dalam pendidikan profesional guru adalah pemahaman tentang pedagogi (Wahyuni et al., 2020a), sehingga mampu menyediakan pembelajaran yang efektif. Pemahaman pedagogi oleh guru perlu diamati dengan instrumen yang dapat mengukur kemampuan tersebut serta konsisten menunjukkan hasil pada kelompok setara

Instrumen pengukuran kompetensi yang digunakan dalam program pengembangan profesional bagi guru maupun calon guru di Indonesia perlu dianalisis kembali untuk memastikan lulusan memiliki kompetensi yang sesuai. Perlunya analisis tersebut dikarenakan program pengembangan profesi guru belum dapat memastikan semua guru SMK memiliki kompetensi yang baik, sebagaimana temuan penelitian yang dilakukan oleh Apriyani et al., (2020) bahwa kompetensi pemahaman konten guru di Sekolah Menengah Kejuruan (SMK) masih belum cukup untuk menghasilkan pembelajaran yang efektif dan efisien. Pernyataan yang sejalan juga ditemukan oleh Wahyuni et al., (2020) yaitu terdapat kesenjangan kompetensi guru SMK yang berkaitan dengan penguasaan konten kejuruan dan penerapannya dalam pembelajaran.

Temuan penelitian oleh Estriyanto et al., (2017) semakin mendorong pentingnya analisis instrumen penilaian pemahaman konten guru untuk mendukung peningkatan profesional, yaitu konsepsi standar kompetensi guru SMK di Indonesia masih kurang jelas sehingga mengakibatkan sulitnya membuat instrumen pengukuran dalam uji kompetensi. Analisis terhadap instrumen penilaian dapat memberikan solusi terhadap kurangnya kompetensi guru melalui kegiatan evaluasi (Kang et al., 2018). Hasil analisis digunakan untuk menyeleksi butir dalam instrumen yang dapat mengukur kompetensi pemahaman konten guru maupun calon guru SMK dalam program pengembangan profesi.

Tujuan dari penelitian ini adalah untuk menganalisis secara empiris soal yang digunakan dalam mengukur pemahaman pedagogi guru SMK dalam pengembangan profesional, khususnya pada bidang Agribisnis Ternak. Pengembangan instrumen untuk mengukur kompetensi guru penting untuk memastikan pengembangan profesional yang berkelanjutan (Ramanan & Mohamad, 2021). Hasil dari penelitian ini akan dijadikan dasar untuk memberikan rekomendasi terhadap peningkatan kualitas instrumen dalam program pengembangan profesi guru di bidang kejuruan.

METODE

Penelitian ini dilakukan dengan analisis psikometri instrumen pengukuran pemahaman pedagogi guru SMK pada bidang Agribisnis Ternak se-Indonesia. Dua jenis analisis digunakan, yaitu pertama dengan analisis deskriptif dengan *classical test theory* (CTT), sebagaimana Sheng, (2019) meliputi validitas, reliabilitas, kesukaran, daya beda, dan distraktor, dan kedua dilakukan dengan menguji korelasi antara validitas, daya beda, dan kesukaran soal menggunakan bahasa pemrograman R melalui aplikasi RStudio.

Data dikumpulkan melalui tes menggunakan sistem *computer assisted test* (CAT). Soal yang digunakan berupa pilihan ganda yang terdiri dari lima pilihan jawaban, dengan jumlah butir soal sebanyak 35. Sampel guru yang mengikuti CAT sebanyak 99 orang dari berbagai SMK di seluruh Indonesia.

HASIL DAN PEMBAHASAN

Hasil uji validitas menunjukkan bahwa dari 35 soal, hanya 9 butir soal yang tergolong valid, sementara 26 butir soal lainnya tidak valid (Tabel 1). Analisis menggunakan uji korelasi *Pearson* yang dilakukan antara skor tiap butir dan skor total. Penentuan validitas soal ditentukan dari nilai koefisien korelasi, sebagaimana skema umum yang dijelaskan oleh Spencer, (1995) yaitu rendah jika kurang dari 0,30, sedang jika berada pada antara 0,30 – 0,70, tinggi jika 0,70 – 0,90, dan sangat tinggi jika diatas 0,90. Koefisien korelasi lebih kecil dari 0,30, memiliki kontribusi kurang dari 30% pada konstruk validitas soal (Atmoko, 2020). Hasil analisis menunjukkan bahwa sebagian besar soal belum mampu mengukur kemampuan yang diinginkan, sehingga perlu direvisi atau bahkan diganti agar instrumen dapat berfungsi sesuai tujuan pengukuran.

Tabel 1 . Validitas Butir Tes

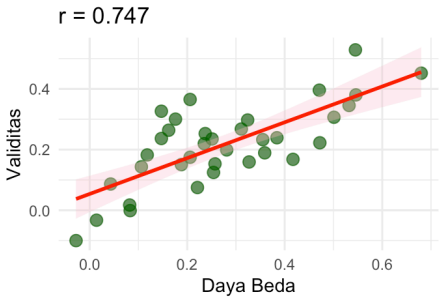
Koefisien Korelasi	Keterangan	Kriteria	Jumlah Butir
< 0,30	Rendah	Tidak valid	26
0,30 – 0,70	Sedang	Valid	9
0,71 – 0,90	Tinggi	Valid	0
> 0,90	Sangat tinggi	Valid	0

Validitas merupakan dasar fungsional bagi sebuah instrumen. Penelitian ini berfokus validasi empiris digunakan untuk memastikan bahwa instrumen tidak hanya relevan secara teori melainkan juga dalam penggunaan di situasi nyata (Hernández-Vicente et al., 2017).Validitas instrumen tes untuk guru dalam program pengembangan profesional sangat penting untuk memastikan kompetensi terukur dengan jelas. Uji validitas instrumen akan menunjukkan beberapa item yang tidak valid untuk ditindaklanjuti melalui revisi.

Penelitian yang dilakukan oleh Atmoko, (2020) pada bidang studi Bimbingan dan Konseling menunjukkan 38 dari 105 soal yang tergolong valid atau memiliki koefisien korelasi ≥ 0,30. Penyebab tingginya soal yang dikembangkan untuk mengukur pemahaman profesi guru tidak valid atau memiliki validitas rendah dijelaskan oleh Wantoro et al., (2019) dalam penelitiannya yaitu adanya intensitas pengerjaan yang sama serta pemahaman guru yang sudah baik terhadap materi. Selain itu, jawaban yang ditebak secara acak dapat mempengaruhi validitas instrumen (Yasuda & Taniguchi, 2013). Karena itu, soal perlu ditingkatkan di bagian daya beda, kesukaran, dan pengecoh agar dapat mengukur kemampuan guru secara akurat.

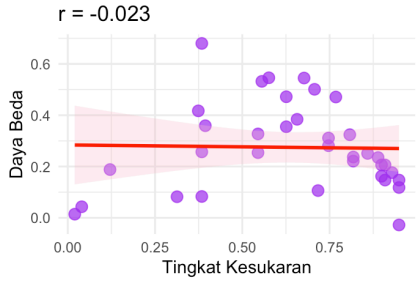
Daya beda merupakan kemampuan suatu instrumen dalam membedakan kemampuan responden (Quaigrain & Arhin, 2017). Berdasarkan kategori daya beda soal oleh Atmoko, (2020), soal dengan selisih responden menjawab benar di kelompok bawah dan kelompok atas berada pada rentang 0 – 29% memiliki daya beda rendah, 30% - 69,9% ideal atau cukup, dan 70% ke atas memiliki daya beda tinggi. Karena itu, hasil analisis daya beda menunjukkan bahwa terdapat 14 butir soal yang memiliki daya beda ideal atau cukup, sedangkan 21 butir soal lainnya termasuk kategori jelek. Kondisi ini mengindikasikan bahwa sebagian besar soal belum efektif dalam membedakan peserta berkemampuan tinggi dengan yang berkemampuan rendah. Daya beda memiliki hubungan dengan validitas soal.

Penelitian yang dilakukan oleh Quaigrain & Arhin, (2017) mengatakan bahwa daya beda merupakan salah satu indikator kualitas soal, sehingga memiliki hubungan yang positif dengan validitas. Panjaitan et al., (2018) juga mengatakan bahwa skor validitas butir soal umumnya memiliki pola yang sama dengan daya beda. Hasil analisis korelasi dengan menggunakan bahasa pemrograman R melalui aplikasi RStudio, menunjukkan bahwa daya beda soal dan validitas memiliki hubungan yang kuat secara positif (Gambar 1).



Gambar 1. Hasil Uji Korelasi antara Validitas dan Daya Beda Soal

Berdasarkan Gambar 1, dapat dikatakan bahwa semakin baik validitas suatu soal, maka semakin baik juga daya bedanya. Berbeda dengan validitas, daya beda soal justru memiliki hubungan yang negatif dengan tingkat kesukaran. Sweeney et al., (2022) mengatakan bahwa semakin tinggi kesukaran soal, maka semakin rendah kemampuannya dalam membedakan kemampuan responden. Pernyataan tersebut didukung oleh temuan penelitian ini, yaitu terdapat korelasi negatif antara daya beda dan kesukaran soal, yaitu dengan koefisien korelasi -0,023 (Gambar 2), yang berarti semakin tinggi tingkat kesukaran maka semakin rendah daya beda, atau sebaliknya.



Gambar 2. Hasil Uji Korelasi antara Daya Beda dan Tingkat Kesukaran

Berdasarkan analisis deskriptif terhadap hasil uji tingkat kesukaran, terdapat 14 butir soal yang tergolong sedang, 3 butir tergolong sulit, dan 18 butir tergolong mudah. Distribusi soal pada hasil analisis kesukaran menunjukkan bahwa sebagian besar soal yaitu sebanyak 21 yang terdiri dari kategori mudah dan sulit sehingga tidak ideal untuk digunakan. Tingkat kesukaran item ditentukan dari jumlah responden yang menjawab banar. Soal yang ideal memiliki jumlah jawaban benar tidak terlalu sedikit yaitu kurang 30% dan tidak terlalu banyak yaitu lebih dari 70%. Tingkat kesukaran instrumen yang baik adalah tidak terlalu mudah dan tidak terlalu sulit (Muchtar et al., 2020). Selain kesukaran, distraktor adalah bagian yang penting dari kualitas soal.

Distraktor merupakan pilihan jawaban yang tampak seperti benar (Kumar et al., 2023), sehingga dapat mengungkap miskonsepsi siswa (Özdemir & Toker, 2025). Berdasarkan hasil analisis terhadap distraktor, ditemukan bahwa terdapat empat kategori soal berdasarkan kemampuan pilihan jawaban dalam mengecoh. Dari 35 butir soal, sebanyak 14 butir yang semua pilihan jawabannya tidak berfungsi sebagai mengecoh, 8 butir memiliki dua pilihan jawaban yang berfungsi sebagai pengecoh, 8 butir memiliki 3 pilihan jawaban yang berfungsi sebagai pengecoh, 5 butir yang empat pilihan jawabannya berfungsi sebagai pengecoh.

Berdasarkan hasil analisis, dilakukan pemilihan terhadap butir soal yang akan digunakan berdasarkan validitas, daya beda, kesukaran, dan distraktornya. Jika butir soal masih tergolong tidak valid namun memiliki daya beda, kesukaran, dan distraktor yang baik, maka akan direvisi sesuai kekurangan yang ditemukan sebelum digunakan. Begitu juga sebaliknya, jika ditemukan soal yang kurang dalam hal daya beda, kesukaran, dan distraktor namun valid, maka akan digunakan setelah revisi. Adapun jika butir soal sudah memenuhi semua kriteria yaitu valid, memiliki daya beda cukup, kesukaran sedang, dan semua pilihan jawaban berfungsi sebagai distraktor, butir soal akan digunakan tanpa revisi.

Item yang tidak valid masih dipertimbangkan untuk digunakan. Validitas instrumen bisa dipengaruhi oleh responden yang memberikan jawaban secara acak (van Laar & Braeken, 2022). Karena itu, kualitas instrumen dapat dipertimbangkan dari daya beda, kesukaran, dan distraktor, sebalum membuang butir dalam instrumen. Daya beda suatu instrumen merupakan kemampuan untuk membedakan kemampuan tiap individu (Metsämuuronen, 2020). (Kim et al., 2024) melakukan pengukuran terhadap daya beda untuk memastikan validitas instrumen. Begitu juga dengan tingkat kesulitan, dapat menjadi pendorong kualitas instrumen yaitu dengan mengakomodasi setiap kemampuan responden yang berbeda (Muchtar et al., 2020). Selain itu distraktor dapat menjadi pendorong validitas dan keandalan (Deepak et al., 2015) serta kualitas instrumen (Alhazmi et al., 2024; Shakurnia et al., 2022).

Berdasarkan semua pertimbangan yang telah disesuaikan dari temuan penelitian, maka dari 35 butir soal, hanya sebanyak 1 butir soal yang digunakan tanpa revisi, 16 butir yang dipakai dengan revisi, dan 18 butir lainnya digugurkan (Tabel 2).

Tabel 2. Butir Soal yang Digunakan

Nomor Item	Validitas	Daya Beda	Tingkat Kesukaran	Distraktor		Keterangan
				Jumlah yang Berfungsi	Pilihan yang Direvisi	
5	Valid	Jelek	Mudah	1	A, C, D	Dipakai dengan revisi
8	Tidak valid	Cukup	Sedang	3	D	Dipakai dengan revisi
10	Tidak valid	Cukup	Sedang	1	B, D, E	Dipakai dengan revisi
11	Valid	Cukup	Mudah	1	A, B, E	Dipakai dengan revisi
12	Tidak valid	Cukup	Sedang	4	-	Dipakai dengan revisi
14	Valid	Cukup	Sedang	4	-	Dipakai tanpa revisi
15	Valid	Jelek	Mudah	-	A, C, D, E	Dipakai dengan revisi
16	Tidak Valid	Cukup	Sedang	1	C, D, E	Dipakai dengan revisi
17	Valid	Cukup	Mudah	2	A, E	Dipakai dengan revisi
19	Tidak valid	Cukup	Mudah	1	A, C, D	Dipakai dengan revisi
20	Valid	Cukup	Sedang	2	A, D	Dipakai dengan revisi
22	Valid	Cukup	Sedang	2	A, C	Dipakai dengan revisi
27	Tidak valid	Cukup	Sedang	4	-	Dipakai dengan revisi
28	Tidak valid	Cukup	Sedang	3	A	Dipakai dengan revisi
31	Valid	Jelek	Mudah	-	A, C, D, E	Dipakai dengan revisi
34	Tidak valid	Cukup	Mudah	3	D	Dipakai dengan revisi
35	Valid	Cukup	Sedang	2	C, D	Dipakai dengan revisi

Revisi yang harus dilakukan sesuai Tabel 2, disesuaikan dengan kekurangan yang ditemukan pada bagian daya beda, tingkat kesukaran, dan distraktor. Selain itu, terdapat 8 butir soal yang tidak valid namun tetap digunakan, dikarenakan memiliki daya beda dan tingkat kesukaran yang baik, serta beberapa pilihan jawaban yang berfungsi sebagai pengecoh. Setelah mengetahui kualitas soal dari validitas, daya beda, kesukaran dan distraktor, selanjutnya dilakukan uji reliabilitas untuk mengetahui konsistensi soal dalam mengukur kemampuan yang diinginkan.

Uji reliabilitas dilakukan dengan dua tahap. Hasil uji reliabilitas tahap pertama, ditemukan bahwa 17 butir soal dapat dilakukan revisi karena pertimbangan kondisi daya beda, kesukaran, dan pengecoh yang baik. Selain itu, nilai *Cronbach's Alpha if Item Deleted* setiap item tidak lebih tinggi daripada nilai *Cronbach's* total, yaitu 0,711 atau berada pada kategori dapat diterima jika mengacu pada (George & Mallery, 2003). Hasil uji reliabilitas tersebut menunjukkan bahwa sebagian besar item mengganggu reliabilitas soal secara keseluruhan. Dari 17 item soal yang dipilih, hanya 8 butir yang memiliki *Corrected Item-Total Correlation* > 0,3 , yang berarti 9 item yang lain tidak memiliki kualitas konstruk yang baik (Magdy et al., 2024), sehingga butuh pertimbangan untuk perbaikan.

Setelah dilakukan penghapusan soal sebanyak 18 butir, kembali dilakukan uji reliabilitas tahap dua. Temuan yang diperoleh adalah nilai *Cronbach's Alpha if Item Deleted* untuk semua item lebih rendah dari nilai reliabilitas total yaitu 0,702 atau pada kategori dapat diterima, yang berarti bahwa setiap butir tidak menjadi penganggu pada reliabilitas soal. Dari 17 butir yang telah dianalisis kembali, menunjukkan bahwa hanya terdapat 5 butir yang memiliki kualitas baik dan selaras dalam mengukur item lain dalam skala. Hal ini dibuktikan dari nilai *Corrected Item-Total Correlation* lebih besar dari 0,3. Adapun 12 item lainnya, perlu dipertimbangkan untuk revisi. Perbandingan antara kualitas item berdasarkan validitas dan reliabilitasnya dapat dilihat pada Tabel 3.

Tabel 3 . Perbandingan Reliabilitas Item yang Digunakan pada Uji pertama dan Kedua

Nomor Butir	*Corrected Item-Total Correlation* (Uji Pertama)	*Cronbach's Alpha if Item Deleted* (Uji Perama)	*Corrected Item-Total Correlation* (Uji Kedua)	*Cronbach's Alpha if Item Deleted* (Uji Kedua)
5	0,365	0,699	0,290	0,691
8	0,168	0,709	0,249	0,694
10	0,233	0,704	0,207	0,699
11	0,307	0,699	0,280	0,690
12	0,222	0,705	0,258	0,693
14	0,452	0,687	0,394	0,676
15	0,300	0,703	0,275	0,692
16	0,159	0,710	0,197	0,700
17	0,396	0,693	0,373	0,680
19	0,297	0,700	0,276	0,690
20	0,529	0,682	0,498	0,664
22	0,380	0,693	0,371	0,679
27	0,189	0,707	0,225	0,696
28	0,239	0,704	0,294	0,688
31	0,326	0,703	0,248	0,695
34	0,268	0,702	0,276	0,690
35	0,345	0,695	0,304	0,687
Cronbach's Alpha		0,711		0,702

Penyebab adanya penurunan nilai *Chronbach's Alpha* dari analisis reliabilitas dikarenakan jumlah butir yang terbatas. Semakin tinggi jumlah item yang diuji secara reliabilitas akan menghasilkan temuan analisis yang lebih baik (Hayashi & Yuan, 2023). Kekurangan jumlah butir yang dianalisis akan mengakibatkan terjadinya ketidakstabilan analisis reliabilitas (Aubin et al., 2020). Hal yang sama juga terjadi dalam penelitian ini, yaitu semakin sedikit jumlah item yang dianalisis, item dengan total korelasi as 0,3 semakin menurun, serta seolah tidak sejalan dengan keadaan nilai *Cronbach's Alpha if Item Deleted* yang lebih rendah dari reliabilitas total item.

KESIMPULAN DAN SARAN

Kesimpulan

Pengembangan kompetensi profesional guru yang berkelanjutan dapat dilakukan dengan menyediakan instrumen yang berkualitas. Berdasarkan temuan penelitian, kurang dari setengah jumlah butir instrumen yang sebelumnya dibuat untuk dapat digunakan dalam mengukur kompetensi guru. Selain itu, butir yang dipertimbangkan dapat digunakan belum sepenuhnya berkualitas karena nilai dalam korelasi total kurang konsisten. Perlu dilakukan perbaikan untuk semua item dan menambah butir soal agar hasil analisis menjadi lebih terpercaya.

Saran

Penyusunan instrumen untuk mengukur kemampuan guru dalam program PPG dapat dilakukan

dengan melibatkan akademisi dan praktisi lapangan. Keterlibatan berbagai pihak akan meningkatkan fungsi dan kualitas instrumen.

UCAPAN TERIMA KASIH

Penelitian ini didanai oleh dana **PPG Universitas Negeri Malang**. Penulis mengucapkan terima kasih kepada seluruh guru SMK yang bersedia menjadi responden, kepada tim pengelola program PPG Universitas Negeri Malang atas dukungan administrasi dan pendanaan, serta kepada keluarga yang memberikan dukungan moral selama pelaksanaan penelitian.

DAFTAR RUJUKAN

Abdallah, A. K., & Abdallah, R. K. (2023). Achieving academic excellence: The intersection of teacher development, quality education, and entrepreneurship. In *Innovations in Teacher Development, Personalized Learning, and Upskilling the Workforce* (pp. 136–158). IGI Global. https://doi.org/10.4018/978-1-6684-5518-0.ch007

Alhazmi, E., Sheng, Q. Z., Zhang, W. E., Zaib, M., & Alhazmi, A. (2024). Distractor Generation in Multiple-Choice Tasks: A Survey of Methods, Datasets, and Evaluation. *EMNLP 2024 - 2024 Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference*, 14437–14458. https://doi.org/10.18653/v1/2024.emnlp-main.799

Apriyani, D., Krisnawati, M., Nurmasitah, S., Sudana, M. M., & Fajri, S. N. (2020). Standard competency gaps of vocational teachers. *International Journal of Innovation and Learning*, *28*(1), 40–46. https://doi.org/10.1504/IJIL.2020.28.ISSUE-1/ASSET/14C55DFB-C814-55DF-EC81-C55DFBEC814C/COVER.JPG

Atmoko, A. (2020). Pengembangan Tes Potensi Keberhasilan Akademik pada Mahasiswa Program Studi Bimbingan dan Konseling. *Jurnal Konseling Gusjigang*, *6*(2), 72–84. https://doi.org/10.24176/jkg.v6i2.2972

Aubin, A. S., Young, M., Eva, K., & St-Onge, C. (2020). Examinee Cohort Size and Item Analysis Guidelines for Health Professions Education Programs: A Monte Carlo Simulation Study. *Academic Medicine*, *95*(1), 151–156. https://doi.org/10.1097/ACM.0000000000002888

Deepak, K. K., Al–Umran, K. U., Al–Sheikh, M. H., Adkoli, B. V., & Al -Rubaish, A. (2015). Psychometrics of multiple choice questions with non-functioning distracters: Implications to medical education. *Indian Journal of Physiology and Pharmacology*, *59*(4), 428–435.

Estriyanto, Y., Pardjono, P., & Sofyan, H. (2017). *The missing productive vocational high school teacher competency standard in the Indonesian education system*. https://www.researchgate.net/publication/318225537

George, D., & Mallery, P. (2003). *SPSS for Windows Step by Step A Simple Guide and Reference. 11.0 Update* (4th ed.). Boston Allyn & Bacon. https://www.scirp.org/reference/ReferencesPapers?ReferenceID=1457632

Hayashi, K., & Yuan, K. H. (2023). On the Relationship Between Coefficient Alpha and Closeness Between Factors and Principal Components for the Multi-factor Model. *Springer Proceedings in Mathematics and Statistics*, *422*, 173–185. https://doi.org/10.1007/978-3-031-27781-8_16

Hernández-Vicente, I. A., Lumbreras-Guzmán, M., Méndez-Hernández, P., Rojas-Lima, E., Cervantes-Rodríguez, M., & Juárez-Flores, C. A. (2017). Validation of a scale to assess the labour quality of life in public hospitals from Tlaxcala[Validación de una escala para medir la calidad de vida laboral en hospitales públicos de Tlaxcala]. *Salud Publica de Mexico*, *59*(2), 183–192. https://doi.org/10.21149/7758

Kang, M., Kim, K., & Cho, B. (2018). Development of a competency assessment instrument for early childhood teachers. *Asia Life Sciences*.

Kim, Y. H., Kim, B. H., Kim, J., Jung, B., & Bae, S. (2024). Item difficulty index, discrimination index, and reliability of the 26 health professions licensing examinations in 2023, Korea: a psychometric study. *Journal of Educational Evaluation for Health Professions*, *21*, 1–4.

https://doi.org/10.3352/jeehp.2024.21.40

Kumar, A. P., Nayak, A., Manjula Shenoy, K., Goyal, S., & Chaitanya. (2023). A novel approach to generate distractors for Multiple Choice Questions. *Expert Systems with Applications*, *225*. https://doi.org/10.1016/j.eswa.2023.120022

Magdy, R., Mohammed, Z., Hassan, A., Ali, M., Ibrahim, A., Adel, S., & Hussein, M. (2024). Validation of the Arabic version of Parkinson's Disease Sleep Scale-Revised Version (PDSS-2). *Revue Neurologique*, *180*(3), 195–201. https://doi.org/10.1016/j.neurol.2023.08.018

Metsämuuronen, J. (2020). Generalized Discrimination Index. *International Journal of Educational Methodology*, *6*(2), 237–258. https://doi.org/10.12973/ijem.6.2.237

Muchtar, H. K., Nahadi, & Hernani. (2020). Evaluation of chemical literacy assessment instruments in solution materials. *Journal of Physics: Conference Series*, *1521*(4). https://doi.org/10.1088/1742-6596/1521/4/042061

Özdemir, A. Z., & Toker, Z. (2025). Analysis of distractors in mathematics questions and their potential to lead misconceptions. *Thinking Skills and Creativity*, *56*. https://doi.org/10.1016/j.tsc.2024.101730

Panjaitan, R. L., Irawati, R., Sujana, A., Hanifah, N., & Djuanda, D. (2018). Item validity vs. item discrimination index: A redundancy? *Journal of Physics: Conference Series*, *983*(1). https://doi.org/10.1088/1742-6596/983/1/012101

Purwantiningsih, A., & Suharso, P. (2019). Improving Teacher Professionalism Toward Education Quality in Digital Era. *Journal of Physics: Conference Series*, *1254*(1). https://doi.org/10.1088/1742-6596/1254/1/012019

Quaigrain, K., & Arhin, A. K. (2017). Using reliability and item analysis to evaluate a teacher-developed test in educational measurement and evaluation. *Cogent Education*, *4*(1). https://doi.org/10.1080/2331186X.2017.1301013

Ramanan, B., & Mohamad, M. Bin. (2021). How Does Teacher Continues Professional Development Practices Help Teachers to Become Competent? Validating Models: A Structural Equation Modeling Approach. *Journal of Physics: Conference Series*, *1793*(1). https://doi.org/10.1088/1742-6596/1793/1/012003

Shakurnia, A., Ghafourian, M., Khodadadi, A., Ghadiri, A., Amari, A., & Shariffat, M. (2022). Evaluating Functional and Non-Functional Distractors and Their Relationship with Difficulty and Discrimination Indices in Four-Option Multiple-Choice Questions. *Education in Medicine Journal*, *14*(4), 55–62. https://doi.org/10.21315/eimj2022.14.4.5

Sheng, Y. (2019). CTT Package in R. *Measurement*, *17*(4), 211–219. https://doi.org/10.1080/15366367.2019.1600839

Spencer, B. (1995). Correlations, sample size, and practical significance: A comparison of selected psychological and medical investigations. *Journal of Psychology: Interdisciplinary and Applied*, *129*(4), 469–475. https://doi.org/10.1080/00223980.1995.9914982

Sweeney, S. M., Sinharay, S., Johnson, M. S., & Steinhauer, E. W. (2022). An Investigation of the Nature and Consequence of the Relationship between IRT Difficulty and Discrimination. *Educational Measurement: Issues and Practice*, *41*(4), 50–67. https://doi.org/10.1111/emip.12522

van Laar, S., & Braeken, J. (2022). Random Responders in the TIMSS 2015 Student Questionnaire: A Threat to Validity? *Journal of Educational Measurement*, *59*(4), 470–501. https://doi.org/10.1111/jedm.12317

Wahyuni, D. S., Agustini, K., Sindu, I. G. P., & Sugihartini, N. (2020a). Analysis on vocational high school teacher competency gaps: Implication for VHS teacher training needs. *Journal of Physics: Conference Series*, *1516*(1). https://doi.org/10.1088/1742-6596/1516/1/012051

Wahyuni, D. S., Agustini, K., Sindu, I. G. P., & Sugihartini, N. (2020b). Analysis on vocational high school teacher competency gaps: Implication for VHS teacher training needs. *Journal of Physics: Conference Series*, *1516*(1). https://doi.org/10.1088/1742-6596/1516/1/012051

Wantoro, J., Sutama, S., Zuhriah, S., & Hafida, S. H. N. (2019). PENGEMBANGAN INSTRUMEN

PENILAIAN PENDIDIKAN PROFESI GURU SEKOLAH DASAR BEBASIS HOTS. *Profesi Pendidikan Dasar*, *1*(1), 11–20. https://doi.org/10.23917/ppd.v1i1.8453

Yasuda, J. I., & Taniguchi, M. A. (2013). Validating two questions in the Force Concept Inventory with subquestions. *Physical Review Special Topics - Physics Education Research*, *9*(1). https://doi.org/10.1103/PhysRevSTPER.9.010113