Measurement in Educational Research
Volume 5, Issue 1, 2025, 16-31
Available online: https://ejournal.ressi.id/index.php/meter

Karakteristik butir soal materi penggabungan badan usaha pada Matakuliah Akuntansi Keuangan Lanjutan I

Natalina Premastuti Bratanigrum [1, a] *, Umi Farisiyah [2, b]
[1] Universitas Sanata Dharma. Jl. Affandi, Mrican, Sleman, 55281, Indonesia
[2] Universitas Negeri Yogyakarta. Jl. Colombo No. 1, Yogyakarta 55281, Indonesia
[a] premastuti@gmail.com; [b] umifarisiyah.2020@student.uny.ac.id
* Corresponding Author.

Received: 30 March 2025; Revised: 9 April 2025; Accepted: 23 April 2025; Published: 25 April 2025

Abstrak: Penelitian ini bertujuan untuk mengetahui karakteristik butir soal materi penggabungan badan usaha pada mata kuliah Akuntansi Keuangan Lanjutan I. Penelitian menggunakan pendekatan kuantitatif dengan desain deskriptif eksploratif, melibatkan 40 mahasiswa sebagai subjek. Instrumen berupa tes pilihan ganda terdiri dari 30 butir soal dianalisis menggunakan teori tes klasik dan model Rasch. Hasil validitas isi melalui indeks V Aiken menunjukkan 24 butir (80%) sangat valid, dan 6 butir (20%) validitas sedang. Secara kualitatif, butir soal memenuhi aspek materi (70%), konstruksi (79%), dan bahasa (97%). Berdasarkan teori tes klasik, 60% butir berada pada tingkat kesukaran sedang, 67% memiliki daya pembeda baik, dan 41,67% pengecoh berfungsi optimal. Analisis faktor menunjukkan 25 butir (83%) signifikan, dengan reliabilitas instrumen tinggi (Cronbach Alpha 0,835). Terdapat satu butir mengalami bias ($p < 0,05$). Dengan model Rasch, 28 butir (97%) sesuai model dan tingkat kesukaran berada dalam rentang sedang (b antara -2 hingga +2). Temuan menunjukkan sebagian besar butir mampu memberikan informasi akurat bagi mahasiswa dengan kemampuan menengah. Simpulan, instrumen soal menunjukkan kualitas baik baik dari sisi isi maupun konstruk.
Kata Kunci: Karakteristik Butir, Teori Tes Klasik, Model Rasch, Penggabungan Badan Usaha

Item characteristics of business entity merger material in Advanced Financial Accounting Course I

Abstract: *This study aims to examine the item characteristics of test questions on business combination material in the Advanced Financial Accounting I course. A quantitative approach with a descriptive-exploratory design was employed, involving 40 students as research subjects. The test instrument consisted of 30 multiple-choice items analyzed using Classical Test Theory (CTT) and the Rasch model. Content validity using Aiken's V index showed that 24 items (80%) were highly valid, while 6 items (20%) had moderate validity. Qualitative review revealed 70% compliance in material aspects, 79% in construction, and 97% in language. Based on CTT, 60% of items had moderate difficulty, 67% demonstrated good discrimination, and 41.67% of distractors functioned effectively. Factor loading analysis showed 25 items (83%) were significant, with a high reliability score (Cronbach's Alpha = 0.835). One item showed bias ($p < 0.05$). The Rasch model analysis indicated that 28 items (97%) fit the model and had difficulty levels within the moderate range (b between -2 and +2). Most items provided accurate information for students with average ability levels. In conclusion, the test items demonstrated good quality in terms of both content and construct validity.*
Keywords: Item Characteristic, Classical Test Teory, Rasch Model, Merger Of Business Entities

How to Cite: Bratanigrum, N. P., Farisiyah, U., & Hassan, A. B. (2025). Karakteristik butir soal materi penggabungan badan usaha pada Matakuliah Akuntansi Keuangan Lanjutan I. *Measurement In Educational Research*, *5*(1), 16-31. https://doi.org/10.33292/meter.v5i1.391

PENDAHULUAN

Undang-Undang Guru dan Dosen Nomor 14 tahun 2005 menggariskan bahwa seorang Dosen hendaknya memiliki kemampuan untuk menyusun rencana pembelajaran, melaksanakan proses pembelajaran, dan melakukan kegiatan penilaian dan evaluasi hasil belajar (Dewan

10.33292/meter.v5i1.391

Perwakilan Rakyat Indonesia, 2005). Dalam perencananaan akan ditetapkan tujuan pembelajaran dan rancangan skenario pembelajaran berdasarkan pada karakteristik materi dan karakteristik peserta didik sehingga skenario yang disusun mendukung ketercapaian kompetensi (Irwantoro & Suryanto, 2016). Pada akhirnya, untuk mengetahui ketercapaian tujuan pembelajaran dan kualitas pelaksanaan pembelajaran dilakukan serangkaian kegiatan pengukuran dan penilaian (Widharyanto & Prijowuntato, 2021). Alur tersebut menunjukkan bahwa dalam penyelenggaraan pembelajaran diperlukan kesesuaian antara tujuan pembelajaran, proses pembelajaran, dan penilaian (Arasian et al., 2015).

Dalam dunia pendidikan, pengukuran dan penilaian adalah kegiatan penting yang bertujuan untuk mengumpulkan dan menginterpretasi data tentang kinerja dan perkembangan peserta didik. Pengukuran biasanya merujuk pada proses sistematis untuk mengumpulkan data kuantitatif tentang kemampuan atau kinerja melalui instrumen yang telah distandarisasi (Allen & Yen, 1979). Sementara penilaian adalah proses menginterpretasi data tersebut untuk membuat keputusan tentang pencapaian, efektivitas, dan kualitas. Tujuan utama dari kegiatan ini adalah untuk memastikan bahwa informasi yang diperoleh benar-benar mencerminkan kemampuan pembelajar sehingga dapat digunakan untuk meningkatkan pengajaran, mengubah strategi pembelajaran, dan membantu pengambilan keputusan kebijakan pendidikan yang berbasis bukti. Secara konteks pengetahuan, struktur ilmu akuntansi termasuk pengetahuan prosedural, yakni pengetahuan yang berkait dengan proses (Airasian et al., 2015). Proses dalam konteks akuntansi dimulai dari pencatatan, penggolongan, dan peringkasan transaksi dalam bentuk laporan keuangan (Suwardjono, 2016). Sehingga dalam konteks pembelajaran akuntansi baik pada level pengantar-menengah-maupun lanjutan, konsep proses akan terjadi. Kenyataan tersebut memberi dampak pada jenis pengukuran yang dilakukan harus sampai mengukur kebenaran prosedur yang dilakukan, sehingga tipe tes yang cocok dan sering digunakan dosen adalah esai. Pada pihak lain, penggunaan tipe esai mengakibatkan belum terungkapnya pemahaman konsep dibalik prosedur yang dilakukan karena keluasan materi dan waktu yang terbatas. Mencermati hal tersebut, perlu dikembangkan tipe pilihan ganda karena selain mengungkap kemampuan prosedural juga akan terungkap pengetahuan konsepnya.

Terdapat berbagai jenis alat ukur yang dapat digunakan oleh dosen untuk mengukur kemampuan mahasiswa, sebagai contoh tugas terstruktur yang di kerjakan di luar kegiatan tatap muka, penyelenggaraan kuis, maupun ujian yang diadakan pada tengah semester dan di akhir semester. Sementara bentuk-bentuk instrumen yang bisa dikembangkan dalam kegiatan pengukuran adalah tes maupun non tes (Arikunto, 2018). Dengan melakukan kegiatan pengukuran secara komprehensif maka dosen dapat memperoleh skor yang merepresentasikan kemampuan mahasiswa. Selanjutnya skor ini akan dibandingan dengan standar tertentu dan menghasilkan nilai. Dengan demikian, adalah keniscayaan bagi dosen untuk menyusun instrumen tes yang baik, yakni terkonfirmasi validitas maupun reliabilitasnya (Bookhart & Nitko, 2019; Mardapi, 2008). Secara spesifik dapat diungkap instrumen yang baik adalah mengukur satu aspek, konteks kebahasaan yang baik, serta hasil pengukuran memiliki kesalahan sekecil mungkin. Penting diperhatikan bagi dosen untuk memitigasi kesalahan sistematis, yakni menyusun instrumen yang terlalu sulit atau justru terlalu mudah, karena hal ini akan menghasilkan skor yang tidak menunjukkan kemampuan mahasiswa sesungguhnya. Oleh sebab itu, penyusunan kisi-kisi dan telaah kolega serumpun akan membantu dosen dalam menghasilkan instrumen yang berkualitas.

Dalam ilmu pengukuran telah berkembang beberapa jenis analisis untuk memastikan instrumen tes tersebut berkualitas, yakni menganalisis tingkat kesukaran item tes, kemampuan item tes dalam membedakan peserta tes, keberfungsian pengecoh, validitas butir soal serta reliabilitas (Elvira & Hadi, 2016). Indikator-indikator kualitas tersebut berkesesuaian dengan teori tes klasik (Sumaryanto, 2021). Asumsi yang harus dipenuhi dalam konsep teori tes klasik adalah bahwa skor nyata yang diperoleh peserta tes meliputi skor sebenarnya dan skor kesalahan dalam pengukuran kemudian asumsi kedua menjelaskan bahwa tidak ada

keterkaitan antara skor sebenarnya dan skor kesalahan (Mardapi, 2008). Dengan demikian, konteks kualitas dalam teori tes klasik bergantung pada karakteristik peserta tes dan fakta tersebut sekaligus menunjukkan kelemahan teori tes klasik (Azwar, 2015). Kelemahan yang lain adalah teori tes klasik tidak memberikan informasi yang memadai tentang bagaimana peserta tes menanggapi suatu butir. Dengan kata lain, teori ini tidak dapat menyajikan ramalan akurat mengenai seberapa baik peserta melakukan tes. Selain kelemahan yang dimiliki, menurut Hamleton dan Jones teori tes klasik memiliki kelebihan yakni dapat diimplementasikan pada sejumlah responden yang kecil(Bichi, 2016), penggunaan analisis matematis yang sederhana, serta tidak memerlukan uji kecocokan model sehingga konteks evaluasi terhadap pengembangan instrumen tes tetap dapat dilakukan (Sumaryanto, 2021). Kelebihan ini kemudian dimanfaatkan para pendidik untuk dapat mengembangkan tes sekalipun dalam skala kecil.

Memperhatikan berbagai kelemahan yang terdapat dalam teori tes klasik, maka para ahli pengukuran mengembangkan teori modern yang disebut Teori Respon Butir. Salah satu model yang berkembang dalam Teori Respon Butir adalah Model Rasch yakni model logistik dengan parameter tingkat kesukaran (b). Model Rasch pertama kali dikenalkan oleh Georg Rash, seorang matematikawan asal Denmark pada tahun 1960 (Tennant & Küçükdeveci, 2023). Asumsi yang harus dipenuhi dalam Teori Response Butir adalah bahwa setiap butir tes mengukur satu jenis kemampuan (unidimensi), jawaban benar peserta tes pada suatu butir tidak memberikan dampak pada jawaban butir lain (independensi lokal), dan butir tes berfungsi secara konsisten dan *fair* untuk semua kelompok peserta tes (Bichi & Talib, 2018). Beberapa keuntungan yang dapat dideskripsikan berdasarkan Teori Response Butir yaitu: subjek tidak mempengaruhi parameter butir namun membutuhkan sampel berukuran besar, estimasi kemampuan subjek yang diukur bersifat stabil artinya tidak dipengaruhi oleh materi yang diujikan atau bentuk tes, fungsi karakteristik butir dapat digunakan untuk mengestimasi kemampuan masing-masing subjek dengan kesalahan pengukuran yang berbeda untuk masing-masing subjek (Hu et al., 2021).

Sekalipun dosen telah mengupayakan penyusunan instrumen tes untuk mengukur kemampuan mahasiswa berdasar teori pengukuran yang relevan, namun masih perlu dipastikan kualitas instrumen tersebut. Hal ini penting sebagai langkah evaluatif untuk peningkatan sekaligus perbaikan kualitas instrumen tes. Dalam penelitian ini, instrumen tes yang akan dianalisis karakteristiknya adalah instrumen soal materi penggabungan badan usaha pada pada matakuliah Akuntansi Keuangan Lanjutan I (AKL I). Dengan demikian, pada penelitian ini akan mengungkap kualitas instrumen tes tipe pilihan ganda yang belum banyak dikembangan dalam pengukuran AKL I sehingga diketahui karakteristiknya dengan melakukan validasi isi dan konstruk. Validitas isi menggunakan Indek Aiken V dilanjutkan dengan telaah secara kualitatif (teoritis) karakteristik secara konstruk akan dibuktikan berdasarkan teori tes klasik dan model rasch. Pada akhirnya, artikel ini akan mengungkap apakah dua teori yang digunakan dalam analisi ini dapat saling menguatkan hasil satu sama lain.

METODE

Penelitian ini merupakan penelitian kuantitatif dengan desain deskriptif eksploratif. Penelitian ini akan mengungkap karakteristik instrumen butir soal materi penggabungan badan usaha pada matakuliah AKL I. Pelaksanaan penelitian ini pada semester genap 2024/2025.

Subjek penelitian ini adalah mahasiswa peserta matakuliah AKL I dengan jumlah 40 mahasiswa. Objek penelitian ini adalah perangkat tes materi penggabungan badan usaha pada matakuliah AKL I dengan jumlah butir pilihan ganda sebanyak 30 dalam 5 opsi pilihan dan respon mahasiswa berupa lembar jawab.

Dalam penelitian ini data dikumpulkan dengan teknik tes. Sementara itu, dokumen yang digunakan yakni spesifikasi butir soal materi penggabungan badan usaha dengan jumlah butir sebanyak 30, lembar jawab mahasiswa peserta tes materi penggabungan badan usaha pada matakuliah AKL I, dan kunci jawaban materi penggabungan badan usaha pada matakuliah

AKL I. Pada penelitian ini juga menggunakan kartu telaah untuk melakukan isi yang dilakukan oleh ahli materi.

Teknik analisis data dalam penelitian ini menggunakan pendekatan kualitatif dan pendekatan kuantitatif. Pendekatan kualitatif berupa telaah butir soal pada dimensi materi, konstruksi dan bahasa. Dimana rubrik yang digunakan berpedoman pada panduan penyusunan instrumen tes yang diterbitkan oleh Puspendik. Berdasarkan kesepakatan para ahli instrumen akan dibuktikan dengan validitas isi berdasar indeks V dari Aiken. Selanjutnya, pendekatan kuantitatif dilakukan melalui analisis empiris terhadap butir soal berdasarkan respon mahasiswa. Analisis butir soal akan dilakukan berdasarkan teori tes klasik untuk diketahui tingkat kesukaran, daya pembeda, keberfungsian distraktor, validitas dan reliabilitas serta dif. Analisis butir soal dengan model rasch dilakukan untuk mengetahui kecocokan butir soal dengan model dan tingkat kesukaran butir. Pada analisis kuantitaif ini, baik teori tes klasik maupun model rasch, akan dilakukan dengan bantuan Program R (Team, 2024), yaitu sebuah program yang diinisiasi oleh Ross Ihaka dan Robert Gentleman yang berkarier di Universitas Auckland, Selandia Baru dan berhasil dirilis pada tahun 1993, selanjutnya program ini dikembangkan oleh *Development Core Team* (Arlinwibowo et al., 2024) .

HASIL DAN PEMBAHASAN

Validitas Isi

Validitas isi terkait dengan analisis rasional terhadap domain yang diukur untuk menentukan keterwakilan instrumen dengan kemampuan yang diukur (Retnawati, 2017). Pada penelitian ini validitas isi dibuktikan berdasarkan indeks Aiken V. Tabel 1 ini ditampilkan deskripsi intepretasi validitas isi.

Tabel 1. Deskripsi Validitas Isi

Indeks V	Intepretasi	Jumlah Butir	Persentase
> 0,8	Sangat Valid	24	80
0,41 — 0,8	Validitas sedang	6	20
≤ 0,4	Validitas kurang	0	0
		30	100

Berdasarkan Tabel 1 dapat dideskripsikan bahwa sebagian besar butir, yaitu 24 butir atau 80% butir soal, termasuk dalam kategori sangat valid (Indeks V lebih dari 0,8). Butir-butir soal yang terbukti sangat valid memiliki nilai yang kuat untuk struktur yang dimaksud (Azwar, 2012). Sedangkan terdapat 6 butir atau 20% dari soal termasuk dalam kategori validitas sedang (Indeks V antara 0,41 dan 0,8) yakni butir 1, 2, 12, 13, 19, 21. Tidak ditemukannya butir dengan validitas rendah (≤ 0,4) juga merupakan indikator positif. Oleh karena itu, dapat disimpulkan bahwa sebagian besar butir soal di instrumen ini memiliki tingkat validitas yang tinggi dan layak digunakan dalam proses pengukuran atau evaluasi.

Hasil analisis validitas butir soal dengan Indeks V dapat menunjukkan seberapa baik alat tersebut digunakan untuk pengukuran. Validitas adalah bagian penting dari evaluasi kualitas instrumen karena menunjukkan sejauh mana suatu alat ukur mampu mengukur apa yang seharusnya diukur. Dalam hal ini, menggunakan Indeks V, yang juga dikenal sebagai validitas isi atau *content validity*, merupakan metode yang bergantung pada penilaian ahli tentang kesesuaian isi instrumen dengan indikator yang diukur (Aiken, 1985).

Supaya menjamin instrumen tes yang telah disusun memiliki kualitas yang baik, maka perlu dilakukan telaah pada aspek materi, konstruksi, dan bahasa. Tabel 2 disajikan ringkasan deskripsi telaah kualitatif 2 orang ahli di bidang materi akuntansi dan pembelajaran akuntansi.

Berdasarkan Tabel 2 dapat dijelaskan bahwa dalam setiap dimensi minimal 70% butir instrumen termasuk memenuhi kriteria, baik pada dimensi materi, dimensi konstruksi, dan dimensi bahasa. Namun demikian masih terdapat beberapa butir yang perlu menjadi perhati-

an untuk diperbaiki. Pada dimensi materi, terdapat 9 item yang tidak memenuhi kriteria pengecoh yang homogen dan berfungsi yakni butir: 2, 7, 9, 10, 11, 16, 17, 25 dan 26. Pada dimensi konstruksi terdapat 7 butir yang tidak memenuhi kriteria pilihan jawaban, seharusnya nominal diurutkan dari besar ke kecil atau sebaiknya (Penyusun, 2016), yakni butir: 13, 14, 20, 21, 22, 23, dan 29. Pada dimensi bahasa terdapat 1 butir yang belum menggunakan bahasa yang baik dan benar, yakni butir 19. Selanjutnya, instrumen diperbaiki sesuai saran ahli. Berdasarkan data ini, dosen semestinya terus mengembangkan kemampuan pedagogik utamanya dalam menyusun instrumen, sebagai contoh, institusi dapat mengupayakan pelatihan atau workshop untuk memberikan pengetahuan yang lebih *update* mengenai penyusunan instrumen, mengikuti diskusi atau berdiskusi dengan rekan sejawat terkait penyusunan instrumen yang bermutu (Hendriyaldi, 2023)

Tabel 2. Deskripsi Telaah Kualitatif

Dimensi	Telaah Kualitatif			
	Memenuhi		Tidak Memenuhi	
	Jumlah	%	Jumlah	%
Materi	21	70	9	30
Konstruksi	23	77	7	23
Bahasa	29	97	1	3

Validitas Konstruk

Karakteristik Butir Soal Materi Penggabungan Badan Usaha Berdasarkan Teori Tes Klasik

Sebanyak 30 butir soal telah dianalisis berdasar 40 respons mahasiswa untuk diketahui karakteristiknya meliputi indeks kesukaran, indeks pembeda, dan keberfungsian pengecoh serta keberfungsian butir soal pada kelompok peserta tes. Kriteria yang digunakan sesuai dengan standar yang dikemukan oleh Bichi (2016).

Tingkat Kesukaran Butir Soal

Salah satu karakteristik yang menunjukkan kualitas butir adalah tingkat kesukaran. Tabel 3 disajikan deskripsi tingkat kesukaran butir soal dengan kriteria yang dikembangkan oleh Bichi (Bichi, 2016).

Tabel 3. Tingkat Kesukaran Butir Soal Materi Penggabungan Badan Usaha

Indeks Kesukaran	Intepretasi	Jumlah Butir	Persentase (%)
$p \leq 0{,}3$	Sukar	12	40
$0{,}3 < p \leq 0{,}70$	Sedang	18	60
$p > 0{,}7$	Mudah	0	0
Jumlah		30	100

Berdasarkan Tabel 3 dapat diketahui bahwa sebanyak 60% butir soal terkategori sedang dan sebanyak 40% terkategori sulit. Idealnya dalam instrumen tes, indeks kesukaran pada *range* >0,3 dan ≤ 0,7 yang menunjuk pada intepretasi butir yang tidak sulit namun juga tidak mudah (sedang) (Arikunto, 2018), butir-butir soal tersebut adalah 3, 4, 5, 7, 8, 9, 10, 11, 12, 13, 19, 20, 23, 25, 26, 27, 28, dan 29. Butir-butir ini memiliki tingkat kesulitan yang seimbang dan sesuai dengan prinsip pengukuran yang baik. Dianggap ideal untuk instrumen tes, butir soal dengan tingkat kesulitan sedang mampu membedakan peserta didik dengan tingkat pemahaman yang berbeda tanpa membuat soal terlalu mudah atau sulit (Haladyna & Rodriguez, 2013).

Namun demikian, ditemukan bahwa 40% butir soal tergolong sulit ($p \leq 0, 3\ p \leq 0{,}3$), hal ini menunjukkan bahwa peserta tes mungkin menemukan soal-soal tersebut lebih sulit. Peserta tes dapat mengalami kesulitan dalam menjawab soal yang terlalu sulit, kondisi tersebut akan mengurangi dorongan mereka untuk merespon soal dengan benar sehingga mempengaruhi kredibilitas tes (DiBattista & Kurzawa, 2011). Perlu dilakukan analisis lebih lanjut terhadap bagian-bagian soal yang termasuk dalam kategori sulit untuk menentukan apakah kesulitan

tersebut disebabkan oleh masalah teknis atau memang merupakan karakteristik materi yang disyaratkan. Adapaun faktor-faktor yang menyebabkan butir soal sulit dijawab dengan benar oleh peserta tes adalah peserta tes mungkin tidak memahami materi dengan baik, redaksi soal yang tidak jelas, atau tuntutan tingkat berpikir yang terlalu tinggi dibandingkan dengan kompetensi yang diukur (Bookhart & Nitko, 2019). Pada satu sisi lain, butir soal dengan kategori sulit dapat diberikan jika ditentukan jumlah peserta tes yang dinyatakan lulus sangat terbatas sehingga akan mudah untuk menentukan.

Menariknya, analisis ini tidak menemukan butir soal dengan indeks kesukaran tinggi (p > 7), yang menunjukkan bahwa peserta tes tidak menghadapi soal yang terlalu mudah. Dalam tes ideal, sebagian besar soal terdiri dari kategori mudah untuk memberikan variasi dan mengimbangi tingkat kesulitan soal (Downing, 2018). Namun, tidak adanya soal dalam kategori mudah dapat menunjukkan bahwa soal-soal yang disusun cenderung lebih menantang, yang bermakna tes memang dimaksudkan untuk mengukur kemampuan tingkat tinggi.

Sangat penting untuk mengimbangi tingkat kesulitan pada kategori mudah, sedang, dan sulit saat menyusun butir soal. Komposisi yang ideal dari butir soal adalah dua puluh hingga tiga puluh persen butir soal kategori mudah, empat puluh hingga enam puluh persen butir soal kategori sedang, dan dua puluh hingga tiga puluh persen butir soal kategori sulit (Bookhart & Nitko, 2019) . Oleh karena itu, dalam konteks analisis ini, jumlah butir soal sulit yang cukup besar dapat dievaluasi kembali untuk memenuhi distribusi yang ideal. Adapun instrumen butir soal dapat diubah dengan menyederhanakan redaksi dan memberikan arahan yang lebih jelas (Gierl et al., 2017) .

Kemampuan Membedakan Butir Soal

Dalam penelitian ini kemampuan membedakan butir soal akan dideskripsikan dengan membagi menjadi 4 kelompok. Secara lengkap Tabel 4 menyajikan daya pembeda butir soal materi penggabungan badan usaha dengan 4 kategori yang dikembangkan oleh Bichi (Bichi, 2016).

Tabel 4. Tingkat Daya Pembeda Butir Soal Materi Penggabungan Badan Usaha

Indeks Pembeda	Intepretasi Butir Soal	Jumlah Butir	Persentase (%)
$D \geq 0,4$	Sangat membedakan	12	40
$0,3 \leq D \leq 0,39$	Mampu membedakan, memerlukan sedikit revisi	8	27
$0,2 \leq D < 0,29$	Kurang membedakan, memerlukan revisi	8	27
$D \leq 0,19$	Tidak membedakan, dihapus, atau direvisi keseluruhan	2	6

Berdasarkan Tabel 4, dapat dideskripsikan bahwa daya beda butir soal tentang kemampuan peserta tes terdistribusi menyebar di semua kategori. Terdapat 20 butir soal (67%) yang termasuk kualifikasi mampu membedakan yakni butir nomor 1, 2, 4, 6, 9, 12, 13, 15, 16, 17, 18, 19, 21, 22, 23, 24, 26, 27, 28, dan 30. Artinya, butir soal tersebut mampu membedakan peserta tes yang menguasai materi dan yang belum menguasai materi. Peserta tes yang memiliki kemampuan tinggi tentu dapat mengerjakan butir soal tersebut, dan sebaliknya dengan peserta tes dengan kemampuan rendah tidak dapat menjawab butir soal tersebut dengan benar. Pada dasarnya, ada banyak pendapat mengenai keberterimaan indeks daya pembeda, Ebel & Frisbie, mengungkapkan bahwa indeks daya pembeda tidak perlu memiliki rincian kategori yang mendetail, artinya, selagi nilainya positif, maka butir soal tersebut sudah dapat dapat membedakaan kemampuan peserta tes (Suseno, 2017).

Daya pembeda adalah komponen penting dalam analisis butir soal karena menentukan sejauh mana suatu soal dapat membedakan antara peserta yang berkemampuan tinggi dan rendah (Gierl et al., 2017). Soal dengan daya pembeda tinggi cenderung dijawab dengan benar oleh peserta yang berkemampuan tinggi, dan soal dengan daya pembeda rendah cenderung dijawab secara acak, sehingga tidak memberikan informasi apa pun. Namun demikian terdapat beberapa faktor yang menyebabkan indeks pembeda tidak ideal, salah satunya adalah sulitnya karakteristik materi yang diujikan, sehingga peserta tes menentukan jawaban tidak

didahului proses berpikir yang runtut namun hanya sekedar menjawab (Ambarwati & Ismiyati, 2021).

Terdapat beberapa saran untuk meningkatkan kualitas soal agar lebih efektif dalam membedakan siswa berdasarkan pemahaman mereka berdasarkan analisis daya pembeda butir soal. Pertama, soal dengan indeks pembeda rendah harus direvisi untuk menjadi lebih optimal. Ini dapat dilakukan dengan memperjelas redaksi soal, meningkatkan tingkat kesulitan sesuai dengan kompetensi yang diukur, atau memperbaiki pilihan jawaban (pengecoh) agar lebih efektif (Downing, 2018). Pengecoh yang baik harus mampu menarik peserta tes untuk memilih tanpa memberi mereka arahan yang terlalu jelas tentang jawaban yang benar (Bookhart & Nitko, 2019) .

Tingkat Keberfungsian Pengecoh

Berikut ini adalah deskripsi fungsi pengecoh butir soal materi penggabungan badan usaha. Pengecoh disimpulkan berfungsi dengan baik jika dipilih oleh minimal 5% dari peserta tes (Sajjad et al., 2020). Setelah masing-masing distraktor dianalisis kemudian diringkas berdasarkan jumlah pengecoh yang berfungsi dalam setiap butirnya (Ambarwati & Ismiyati, 2021).

Tabel 5. Analisis Fungsi Pengecoh Butir Soal Materi Penggabungan Badan Usaha

Fungsi Pengecoh		Jumlah Item	%
Jumlah	%		
4	100	1	3
3	75	8	27
2	50	6	20
1	25	10	33
0	0	5	17
Jumlah		30	100

Dalam instrumen ini terdiri dari 30 butir dan setiap butir memiliki 4 pengecoh serta 1 kunci jawaban. Dengan demikian total alternatif pilihan jawaban sebanyak 5 x 30 = 150; serta pengecoh keseluruhan sebanyak 4 x 30 = 120. Mencermati Tabel 5, dapat dijelaskan bahwa terdapat 1 butir item (3%) dimana semua pengecohnya berfungsi dengan baik yakni nomor 19. Selanjutnya terdapat 8 butir soal (27%) dimana 3 butir pengecohnya berfungsi efektif, 6 butir soal (20%) yang memiliki 2 distraktor berfungsi efektif, 10 butir soal (33%) yang memiliki 1 disktraktor yang berfungsi efektif, serta 5 item butir soal (17%) yang semua pengecohnya tidak berfungsi dengan baik. Sehingga secara keseluruhan terdapat 50 butir (41,67%) pengecoh yang berfungsi dengan baik, serta 70 butir (58,33%) pengecoh yang belum berfungsi dengan baik.

Pengecoh yang efektif dapat menarik bagi peserta yang kurang memahami materi untuk memungkinkan pengukuran pemahaman yang lebih akurat (Haladyna & Rodriguez, 2013). Namun demikian, menyusun pengecoh tidaklah mudah dan perlu waktu lama (Ihda & Heri, 2023). Unsur-unsur yang semestinya terpenuhi dalam pengecoh adalah pertama homogen, artinya pengecoh harus tetap relevan sebagai jawaban sehingga tidak mudah dihindari oleh peserta tes (Raina et al., 2024). Menurut penelitian yang dilakukan oleh Gierl, pengecoh yang berkesusuaian dengan konteks dapat meningkatkan kualitas butir soal (Gierl et al., 2017). Unsur kedua yaitu pengecoh mengandung kesalahan persepsi secara umum dan dituliskan sesuai tata bahasa yang baku serta memiliki makna yang nalar sehingga mampu mengecoh peserta tes yang tidak menguasai materi (Qiu et al., 2020).

Dengan demikian, terhadap pengecoh yang belum berfungsi optimal, dapat dilakukan beberapa hal yakni tetap menyajikan dalam tes dengan melakukan revisi atau menghapusnya dan mengganti dengan alternatif pilihan yang memenuhi persyaratan (Rezigalla et al., 2024). Hal lain yang perlu dilakukan adalah pengujian awal dan analisis butir. Hal ini sesuai dengan hasil penelitian yang dilakukan oleh Downing yang menunjukkan bahwa analisis butir dan pengujian awal dapat membantu menemukan pengecoh yang tidak berfungsi (Downing, 2018). Terakhir kegiatan pelatihan sangat penting untuk meningkatkan kualitas tes (Haladyna &

Rodriguez, 2013), dengan demikian perlu dilakukan pelatihan kepada penulis soal tentang cara-cara yang efektif untuk membuat pengecoh sehingga meningkatkan kualitas butir soal.

Validitas dan Reliabilitas

Pada penelitian ini, analisis validitas butir menggunakan nilai *loading factor*, yakni koefisien yang menunjukkan seberapa besar hubungan antara variabel pengamatan (butir) dan faktor laten (konstruk). Nilai *loading factor* menunjukkan seberapa baik suatu item mewakili konstruk yang diukur. Secara umum, nilai 0,5, menunjukkan bahwa item tersebut memberi kontribusi yang signifikan dan valid dalam mengukur konstruk tersebut, sedangkan nilai yang rendah menunjukkan bahwa item tersebut kurang sesuai dan mungkin perlu dipertimbangkan untuk diubah atau dihapus. Konsep ini penting untuk memastikan keandalan alat ukur dan validitas konstruk dalam penelitian. Namun demikian dalam penelitian ini nilai *loading factor* ditetapkan sebesar 0,3 yang berkesesuaian dengan jenis penelitian eksploratori (Hair et al., 2010). Pada Tabel 6 ini akan ditampilkan deskripsi analisis validitas butir berdasarkan nilai *loading factor*.

Tabel 6. Analisis Validitas Butir

Kriteria	Intepretasi	Jumlah Butir	Persentasi
Loading ≥ 0,7	Ideal	10	33
Loading ≥ 0,5	Signifikan	15	50
Loading ≥ 0,3	Minimal Signifikan	3	10
Loading < 0,3	Tidak Signifikan	2	7
Jumlah		30	100

Sebagian besar butir (83%) termasuk dalam kategori "signifikan" hingga "ideal", menurut hasil analisis faktor pengisi yang ditunjukkan dalam Tabel 6. Karena mayoritas butir memiliki korelasi yang kuat dengan faktor yang diukur, kondisi ini menunjukkan bahwa instrumen secara umum memiliki kualitas pengukuran yang baik. Sepuluh item dalam kategori "ideal" (nilai *loading factor* lebih dari 0,7) menunjukkan kontribusi mereka terhadap pengukuran konstruk yang benar-benar optimal (Hair et al., 2014). Nilai *loading factor* lebih dari 0,5 sudah dianggap memadai, dan nilai *loading factor* lebih dari 0,7 dianggap sebagai standar ideal.

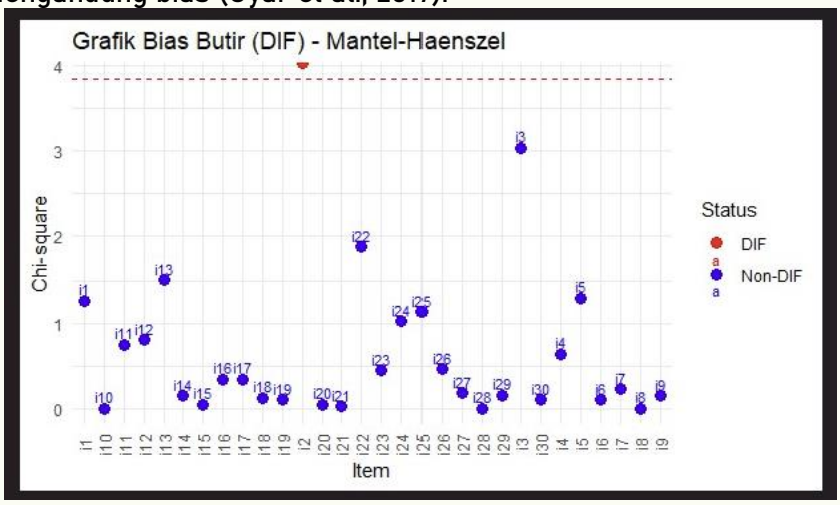
Dalam penelitian ini, dua item dikategorikan sebagai "tidak Signifikan" dengan nilai *loading factor* di bawah 0,3, yakni item 7 dan 29, serta tiga item dikategorikan sebagai "minimal Signifikan" dengan nilai *loading factor* di atas 0,3 tetapi di bawah 0,5. Komponen ini memerlukan perhatian khusus, meskipun jumlahnya kecil. Peneliti perlu meninjau ulang komponen dengan *loading factor* yang rendah karena dapat mengaburkan intepretasi konstruk (Hair et al., 2014).

Dengan demikian, mengingat sedemikian penting kontribusi butir pada konstruk dalam penyusunan instrumen maka diperlukan ahli untuk mencermati secara kontruksi dan isi (Bademci, 2022). Beberapa alternatif untuk meningkatkan nilai *loading factor* butir adalah memperbaiki redaksi atau substansi butir jika elemen ini dianggap penting secara konseptual (Lane et al., 2016). Namun, penghapusan butir dapat dilakukan untuk menjaga kejelasan dan konsistensi struktur yang diukur jika butir tersebut tidak terlalu penting (Lane et al., 2016). Kemudian, pada butir yang direvisi kemudian dilakukan pengujian ulang secara empiris (Haladyna & Rodriguez, 2013). Hal lain adalah dalam konteks ukuran sampel maka penambahan ukuran sampel sangat disarankan supaya validitas semakin tinggi (Datau et al., 2022).

Reliabilitas dimaknai sebagai tingkat konsistensi suatu hasil pengukuran dalam mengukur konstruk yang sama bahkan ketika diulangi pada waktu yang berbeda (Haladyna & Rodriguez, 2013). Nilai *Cronbach's alpha* di atas 0,70 dianggap memadai untuk penelitian eksploratori, dan nilai *Cronbach's alpha* di atas 0,80 dianggap stabil dan reliabel (Hair et al., 2010; Lee & Garg, 2020). Dalam penelitian ini, nilai reliabilitas diketahui melalui *output cronbach Alpha* sebesar 0,835 dan bermakna alat ukur dalam instrumen ini memiliki tingkat konsistensi internal yang baik, item-item di dalamnya secara keseluruhan mampu menghasilkan skor yang stabil dan konsisten dalam kondisi yang sebanding.

Bias Butir

Untuk mendeteksi bias butir dalam penelitian ini menggunakan metode Mantel Haenzel melalui Program R.Terdapat dua kelompok yang diuji yaitu kelompok jenis kelamin meliputi laki-laki dan perempuan dan kelompok program studi yakni kelompok IPA, kelompok IPS dan SMK jurusan Akuntansi, serta kelompok selain kedua kelompok terdahulu. Bias butir dapat dimaknai sebagai kondisi dimana sebuah butir direspon berbeda oleh peserta tes yang memiliki kemampuan relatif sama namun berasal dari kelompok yang berbeda (Osadebe & Agbure, 2019). Dengan arti kata yang lain, semestinya sebuah butir soal akan direspon dengan jawaban yang sama (jawaban benar) oleh peserta tes yang memiliki kemampuan relatif sama sekalipun berasal dari kelompok berbeda. Namun jika hal tersebut tidak terjadi, maka butir soal tersebut mengandung bias (Uyar et al., 2017).



Gambar 1. Grafik Bias Butir

Mencermati Gambar 1, dapat dideskripsikan bahwa butir nomor 2 berwarna merah dengan status DIF yang bermakna butir tersebut mengandung bias jenis kelamin dengan nilai *p-value* sebesar 0,00. Gambar 2 adalah narasi butir nomor 2:

2. Pilihlah salah satu pernyataan berikut ini yang merupakan pernyataan yang benar mengenai bentuk-bentuk persekutuan
   a. Pada persekutuan terbatas, apabila persekutuan tidak dapat membayar utang, maka kreditur dapat menagih utang pada anggota manapun tanpa memandang porsi kepemilikan.
   b. Pada persekutuan umum, anggota dapat bertindak atas nama perusahaan dan dapat diminta tanggungjawab dengan mempertimbangkan status anggota dan jumlah kepemilikan.
   c. Pada *joint stock companies*, tanggungjawab masing-masing anggota terbatas pada jumlah kepemilikan di dalam persekutuan.
   d. Persekutuan perdagangan merupakan terbentuk dengan tujuan membeli barang dagangan dan menjualnya kembali tanpa memproduksi.
   e. Tidak terdapat pilihan jawaban yang tepat.

Gambar 2. Narasi butir nomor 2

Berdasarkan analisis, butir tersebut mengandung bias *gender* karena menimbulkan perbedaan pemahaman atau respons antara siswa laki-laki dan perempuan yang tidak terkait dengan kompetensi yang hendak diukur. Dalam situasi seperti ini, pemilihan kata, contoh, dan ilustrasi yang secara implisit menunjukkan peran gender tertentu dapat menyebabkan bias. Akibatnya, siswa yang merasa tidak terwakili atau tidak akrab dengan konteks tersebut dapat menghadapi tantangan tambahan yang tidak terkait dengan kemampuan materi mereka. Variansi hasil tes yang disebabkan oleh faktor eksternal, seperti stereotip sosial, disebut sebagai variasi konstruktif yang tidak relevan (Bordbar, 2020). Pada sisi lain, butir soal yang mengandung bias gender tersebut secara konteks memerlukan pemahaman verbal yang baik dan dalam hal ini menurut penelitian peserta tes dengan jenis kelamin perempuan lebih diuntungkan (Amalia et al., 2022) .

Langkah mitigasi agar instrumen tidak memiliki bias adalah dilakukannya telaah instrumen yang melibatkan ahli materi, ahli bahasa, dan perwakilan kedua gender untuk mengurangi bias gender. Tujuan dari proses ini adalah untuk memastikan bahwa contoh, konteks, dan redaksi soal yang digunakan inklusif. Jika uji coba lapangan atau analisis statistik telah menunjukkan bias pada bagian soal, revisi dan penggantian bagian dapat dilakukan untuk memastikan bahwa ujian benar-benar mengukur kemampuan yang diinginkan.

Karakteristik butir soal Materi Penggabungan Badan Usaha Berdasarkan Model Rasch

Tiga puluh butir telah dianalisis dengan menggunakan Program R untuk diketahui kecocokan butir dengan Model Rasch dan Indeks Kesukaan. Tabel 7 berturut-turut akan dideskripsikan hasilnya.

Kecocokan Model

Tabel 7. Kecocokan Butir Soal dengan Model Rasch

Kriteria	Intepretasi	Jumlah Butir	Persentase (%)
pvalue ≥ 0,5	Cocok dengan Model	28	93
pvalue < 0,5	Tidak cocok dengan Model	2	7
	Jumlah	30	100

Hasil analisis menunjukkan bahwa 28 dari 30 butir (93%) memiliki nilai *p-value* lebih dari 0,5. Hasil ini menunjukkan bahwa sebagian besar butir sesuai dengan model pengukuran yang digunakan, yaitu model Rasch. Jumlah butir yang memenuhi kriteria menunjukkan bahwa instrumen secara keseluruhan dirancang dengan baik dan sesuai dengan konstruk teoretis yang ingin diukur. Hasil ini sejalan dengan rekomendasi ahli yang menyatakan bahwa *p-value* ≥ 0,5 pada analisis *goodness-of-fit* menunjukkan item cocok dengan model dan dapat dipertahankan tanpa perubahan yang signifikan (Boone & Staver, 2020) .

Namun demikian, terdapat 2 butir soal (7%) yakni butir nomor 2 dan 17 dengan nilai *p-value* < 0,5, yang bermakna adanya ketidakcocokan dengan model rasch. Beberapa saran yang dapat dilakukan untuk memperbaiki butir soal yang tidak cocok dengan model Rasch adalah melakukan revisi stem soal utama yang menimbulkan keraguan atau justru terlalu kompleks, hal lain adalah mengurangi multitafsir dengan menambah konteks khusus, terakhir untuk memastikan bahwa masalah bukan berasal dari karakteristik sampel awal, maka diperlukan pengujian ulang pada kelompok sampel yang berbeda (Bond & Fox, 2020) .

Tingkat Kesukaran Butir Soal

Tabel 8. Tingkat Kesukaran Butir Soal Yang Cocok Dengan Model Rasch

Kriteria	Intepretasi	Jumlah Butir	Persentase (%)
>2	Sukar	0	0
-2 sampai dengan 2	Tidak sukar dan tidak mudah	28	100
<-2	Mudah	0	0
	Jumlah	28	100

Tabel 8 menunjukkan bahwa keseluruhan butir soal yang cocok dengan model rasch, yakni sebesar 28 butir soal (100%) berada dalam rentang kesulitan -2 hingga 2. Butir-butir tersebut adalah: 1, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30. Hal ini berarti bahwa tidak ada satu pun butir soal yang dikategorikan sebagai "sukar" (> 2) atau "mudah" (< 2). Dengan kata lain, setiap komponen soal memiliki tingkat kesulitan yang sedang. Hasil ini menunjukkan bahwa alat tes memiliki keseimbangan kesulitan yang baik, sehingga peserta tidak menemukan item yang terlalu mudah atau terlalu sulit. Butir yang terlalu mudah atau terlalu sulit cenderung tidak informatif dalam membedakan kemampuan peserta didik sehingga kondisi ini biasanya tidak diharapkan dalam ujian.

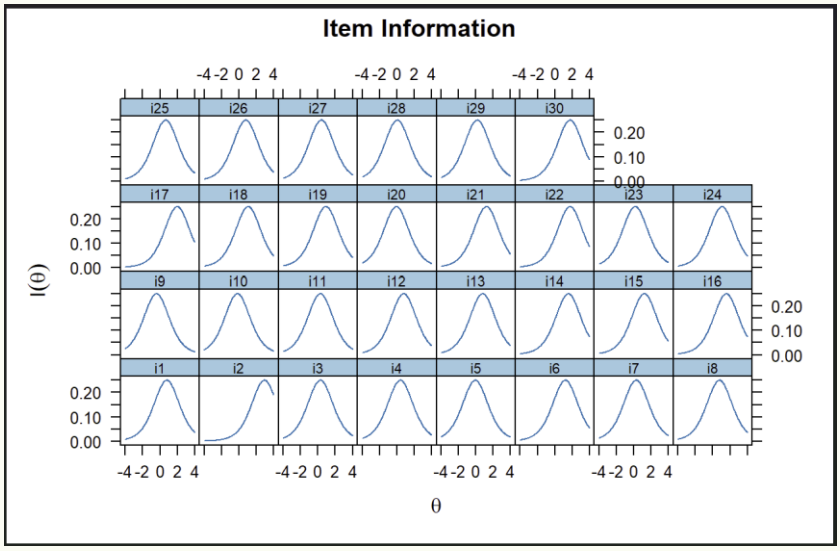
Temuan ini memenuhi prinsip pengukuran yang efektif dalam Teori Respons Butir (*Item Response Theory/Rasch Model*), yakni menunjukkan keseimbangan yang baik dalam desain instrumen. Menurut Hamleton dan Swaminathan butir ideal terlatak pada rentang -2 hingga

+2 dalam skala logit (Retnawati, 2017). Butir ideal memungkinkan instrumen membedakan kemampuan peserta secara optimal (Van der linden, 2016). Butir dalam rentang ini dikategorikan sebagai moderat, yang berarti peserta dengan kemampuan rendah hingga tinggi dapat menjawab dengan benar, tetapi masih memiliki daya diskriminasi yang cukup (Bond & Fox, 2020). Sebagai misal, butir dengan nilai −2 logit dapat dijawab dengan benar oleh sekitar 85% peserta karena memiliki tingkat kesukaran relatif rendah. Sebaliknya, butir dengan nilai +2 logit hanya dapat dijawab dengan benar oleh sekitar 15% peserta karena memiliki tingkat kesukaran butir tinggi. Ahli lain menyatakan bahwa instrumen yang baik harus memiliki distribusi kesukaran butir yang simetris di sekitar nilai 0 logistik, sehingga mencakup spektrum kemampuan peserta secara luas (Boone & Staver, 2020) . Pada sisi lain, Andrich dan Marais dalam studinya menekankan bahwa butir moderat lebih informatif dalam penilaian formatif dan sumatif karena dapat mengidentifikasi kemajuan belajar peserta tes di berbagai tingkat kemampuan (Andrich & Maris, 2019) .

Sekalipun temuan ini positif, ada beberapa hal yang perlu diperhatikan. Pertama, alat tes dengan semua butir moderat mungkin tidak cocok untuk populasi heterogen dengan variasi kemampuan yang sangat beragam, seperti peserta yang dikaruniai atau memiliki kebutuhan khusus (Boone & Staver, 2020). Kedua, jika semua butir terkonsentrasi di sekitar nol logistik, mungkin ada redundansi informasi, yang berdampak tes menjadi kurang efektif (Bond & Fox, 2020) .

Informasi Butir

Sebagai tambahan informasi terkait karakteristik butir menurut Model Rasch. Gambar 3 disajikan informasi butir.



Gambar 3. Informasi Kemampuan Butir

Grafik fungsi informasi butir (*Item Information Function*/IIF) pada Gambar 3 menunjukkan bahwa puncak kurva informasi setiap butir secara rerata berada di rentang kemampuan (θ) yang tidak terlalu ekstrem, yaitu antara −2 hingga +2. Ini menunjukkan bahwa sebagian besar butir soal memberikan informasi paling akurat bagi siswa yang berada di rentang kemampuan menengah. Namun di area kemampuan sangat tinggi (>2) dan sangat rendah (<2), kurva informasi cenderung rendah, sehingga pengukuran menjadi tidak akurat pada tingkat ini. Hal ini sesuai dengan prinsip *targeting* model rasch, di mana kesesuaian antara kemampuan peserta tes (*item-person match*) dan tingkat kesulitan butir sangat penting untuk mengukur presisi (Bond & Fox, 2020).

Namun demikian, penurunan kurva IIF di luar rentang tersebut (θ <-2 atau θ>2), misalnya pada butir 2 dan butir 17 menunjukkan keterbatasan instrumen dalam menjangkau peserta tes dengan kemampuan ekstrim. Hal ini disebabkan oleh kurangnya butir dengan tingkat

kesulitan ekstrem dan distribusi populasi yang homogen (Van der linden, 2016). Maka, penambahan butir pada tingkat kesulitan yang lebih rendah atau lebih tinggi dapat meningkatkan presisi pengukuran dalam situasi di mana cakupan informasi cenderung rendah pada kemampuan ekstrim (Baker & Kim, 2017). Langkah perbaikan ini sejalan dengan tujuan utama model rasch, yakni menitikberatkan pada keselarasan antara distribusi kemampuan peserta didik dan karakteristik butir.

Perbandingan Hasil Penelitian Berdasar Teori Tes Klasik dan Model Rasch

Berdasarkan berbagai hasil penelitian yang telah dianalisis baik menggunakan Teori Tes Klasik maupun Model Rasch menunjukkan hasil yang tidak saling bertentangan. Bahkan dapat saling melengkapi dalam proses pengembangan dan evaluasi instrumen. Menurut Boone dan Staver sekalipun model Rasch menawarkan pendekatan yang lebih modern dan berbasis probabilistik, banyak hasil dari Teori Tes Klasik yang menunjukkan konsistensi, khususnya terkait tingkat kesukaran dan daya pembeda butir (Boone & Staver, 2020). Sebagai misal hasil analisis menggunakan teori tes klasik diperoleh 18 butir (2, 4, 5, 7, 8, 9, 10, 11, 12, 13, 19, 20, 23, 25, 26, 27, 28, 29) diintepretasikan memiliki tingkat kesukaran sedang, dan kedelapan belas item tersebut juga diintepretasikan pada tingkat kesukaran yang sama menurut model Rasch. Temuan yang lain, adalah butir item nomor 2 menurut teori tes klasik teridentifikasi mengandung bias butir, artinya butir ini tidak ideal menjadi alat ukur dan pada analisis kecocokan model butir 2 tersebut tidak cocok dengan model Rasch. Dengan mencermati hasil tersebut dapat dinyatakan bahwa kedua teori ini saling mengkonfirmasi sehingga karakteristik butir soal yang diteliti akan menunjukkan deskripsi kualitas yang komprehensif.

Selain itu, seperti yang ditekankan oleh Sumintono dan Widhiarso, teori tes klasik dan model Rasch bukan dua pendekatan yang saling meniadakan, sebaliknya, keduanya dapat digunakan secara strategis pada tahapan analisis yang berbeda (Sumintono & Widhiarso, 2019). Model Rasch digunakan untuk pengukuran dan pemurnian instrumen yang lebih presisi, sementara teori tes klasik cocok untuk tahap awal penyusunan dan penyaringan butir. Oleh karena itu, kualitas pengukuran dapat ditingkatkan dengan menggunakan keduanya secara bersamaan tanpa adanya pembedaan yang signifikan di antara keduanya. Ini menunjukkan bahwa menggabungkan teori klasik dengan model Rasch adalah metode yang sah dan direkomendasikan untuk praktik evaluasi kontemporer.

SIMPULAN

Berdasarkan temuan dan pembahasan yang telah dilakukan, berikut ini adalah hal-hal yang menjadi kesimpulan. *Pertama*, berdasarkan analisis validitas isi terdapat 24 butir (80%) terkategori valid, hasil telaah kualitatif instrumen belum sepenuhnya menunjukkan kualitas baik, karena hanya aspek bahasa yang memenuhi 97% kriteria. Sementara pada aspek materi dan konstruksi masing-masing memenuhi 70% dan 77%. *Kedua*, validitas konstruk dengan analisis teori tes klasik disimpulkan karakteristik butir soal dikategorikan cukup baik, dikarenakan komposisi tingkat kesukaran belum ideal. Sekalipun butir soal dengan intepretasi sedang sudah memenui (60%) namun kriteria butir soal sulit terlalu banyak (40%). Berdasarkan daya pembeda masih terdapat 10 butir (33%) yang harus direvisi karena kurang membedakan kemampuan peserta tes. Selanjutnya berdasarkan fungsi pengecoh, hanya terdapat 1 butir yang memenuhi fungsi pengecoh dengan sempurna (3%), dan secara keseluruhan pengecoh yang berfungsi dengan baik sebanyak 50 butir (41,67%) sementara 70 pengecoh (58,33%) tidak berfungsi dengan baik. Berdasarkan kriteria validitas, terdapat 25 butir (83%) yang memiliki nilai *loading factor* terkategori signfikan hingga ideal, sementara 5 item (17%) memiliki nilai *loading factor* terkategori minimal signifikan hingga tidak signifikan. Nilai reliabilitas instrumen pada *cronbach alpha* sebesar 0,835 yang dimaknai memiliki tingkat konsistensi internal yang baik. Pada instrumen ini teridentifikasi terjadi bias butir jenis kelamin pada item nomor 2 dengan *p-value* 0,00. *Ketiga*, validitas konstruk berdasarkan model Rasch dapat disimpulkan karakteristik butir soal terkategori baik, hal ini berdasarkan analisis sebanyak 93% atau

28 butir soal cocok dengan model Rasch. Kemudian, dari 28 butir soal yang cocok dengan model Rasch tersebut kesemuanya memiliki tingkat kesukaran sedang dengan nilai b berkisar di antara -2 sampai dengan 2. Berdasarkan fungsi informasi menunjukkan sebagian besar butir soal mampu memberikan informasi aktual bagi siswa yang berkemampuan sedang dikarenakan puncak kurva informasi setiap butir secara rerata berada pada rentanf logit -2 hingga +2.

Berdasarkan kesimpulan tersebut, beberapa saran yang dapat disampaikan adalah sebagai berikut. *Pertama*, instrumen hendaknya diperbaiki sacara isi, mengingat berdasar telaah kualitatif pada aspek materi dan dan konstruksi masih terdapat beberapa item yang belum memenuhi kriteria. *Kedua*, hendaknya perbaikan soal dapat dilakukan secara menyeluruh utamanya pada kemampuan alternatif jawaban mengecoh peserta tes. Hal ini akan berdampak pada meningkatnya butir soal dalam kemampuan membedakan peserta tes. *Ketiga*, dosen atau para penyusun tes hendaknya meningkatkan kemampuan dalam menyusun tes dengan memperbanyak literatur sebagai sarana pengembangan diri atau mengikuti *workshop* dan diskusi-diskusi terkait penyusunan tes.

Kontribusi Artikel terhadap Bidang Ilmu Terkait

Penelitian ini diharapkan memberikan kontribusi di bidang akuntansi utamanya dalam melakukan analisis karakteristik butir soal pilihan ganda baik secara kualitatif maupun kuantitatif. Hasil penelitian ini diharapkan mampu memberikan referensi yang bermakna mengenai berbagai alternatif analisis karakteristik butir soal di bidang pengukuran akuntansi dengan menggunakan teori tes klasik dan model Rasch.

DAFTAR REFERENSI

Aiken, L, R. (1985). Three coefficients for analyzing the reliability and validity of ratings. *Educational Dan Psicological Measurement*, *45*, 131–142. https://doi.org/10.1177/0013164485451012

Airasian, P. W., Cruikshank, K. A., Mayer, R. E., Pintrich, P. R., Raths, J., & Wittrock, M. C. (2015). *Pembelajaran, Pengajaran, Dan Asesmen* (L. W. Anderson & D. R. Krathwohl, Eds.; A. Prihantoro, Trans.; 1st ed.). Pustaka Pelajar.

Allen, M. J., & Yen, W. M. (1979). *Introduction to Measurement Theory*. Brooks/Cole Publishing Company.

Amalia, R., Astuti, S., & Sari, A. (2022). Deteksi bias gender pada instrumen evaluasi belajar kimia dengan metode mantel-haezel. *Tarbiyah*, *29*(2), 243–256. https://doi.org/http://dx.doi.org/10.30829/tar.v29i2.1781

Ambarwati, Y., & Ismiyati, I. (2021). Analisis butir soal pilihan ganda ulangan akhir semester genap mata pelajaran kearsipan. *Measurement in Educational Research*, *1*(2), 64–75.

Andrich, D., & Maris, I. (2019). *A course in rasch model measurement theory*. Spinger.

Arikunto, S. (2018). *Dasar-dasar evaluasi pendidikan* (3rd ed.). PT Bumi Aksara.

Arlinwibowo, J., Retnawati, H., & Hadi, S. (2024). *Aplikasi teori respon butir dengan R dan R Studio*. Cahaya Harapan.

Azwar, S. (2012). *Reliabilitas dan validitas*. Pustaka Pelajar.

Azwar, S. (2015). *Dasar-dasar psikometri*. Pustaka Pelajar.

Bademci, V. (2022). Correcting fallacies about validity as the most fundamental concept in educational and psychological measurement. *International E-Journal of Educational Studies*, *6*(12), 148–154. https://doi.org/10.31458/iejes.1140672

Baker, F. B., & Kim, S. (2017). *The information function. in: the basic of item response theory using R. statistics for social and behavioral sciences.* Springer, Cham. https://doi.org/https://doi.org/10.1007/978-3-319-54205-8_6

Bichi, A. A. (2016). Classical test theory: an introduction to linear modeling approach to test and item analysis. *International Journal for Social Studies* , *2*(9), 27–33. https://edupediapublications.org/journals

Bichi, A. A., & Talib, R. (2018). Item response theory : an introduction to latent trait models to test and item development. *International Journal of Evaluation and Research in Education (IJERE)*, *7*(2), 142–151. https://doi.org/10.11591/ijere.v7.i2.12900

Bond, T. G., & Fox, C. M. (2020). *Applying the rasch model; fundamental measurement in the human sciences*. Raoutledge.

Bookhart, S. M., & Nitko, A. J. (2019). *Educatonal assessment of student* (6th ed.). Pearson.

Boone, W. J., & Staver, J. R. (2020). *Rasch measurement estimation prosedure. In: Advance In Rasch Analyses In Human Sciences*. Spinger. https://doi.org/https://doi.org/10.1007/978-3-030-43420-5_14

Bordbar, S. (2020). Gender differential item functioning (GDIF) analysis in iran ' s university entrance exam. *English Language in Focus (ELIF), 3*(1), 49-68, https://doi.org/10.24853/elif.3.1.49-68

Datau, R., Putrawan, I. M., & Rahayu, W. (2022). The effects of sample size and options number on the validity item of students' environmental personality score. *Proceedings of the Eighth Southeast Asia Design Research (SEA-DR) & the Second Science, Technology, Education, Arts, Culture, and Humanity (STEACH) International Conference (SEADR-STEACH 2021)*, *627*, 165–170. https://doi.org/10.2991/assehr.k.211229.026

Dewan Perwakilan Rakyat Indonesia. (2005). Undang-Undang (UU) tentang guru dan dosen nomor 14. *Dewan Perwakilan Rakyat Indonesia*, 2.

DiBattista, D., & Kurzawa, L. (2011). Examination of the quality of multiple-choice items on classroom tests. *The Canadian Journal for the Scholarship of Teaching and Learning*, *2*(2). https://doi.org/10.5206/cjsotl-rcacea.2011.2.4

Downing, S. M. (2018). *Twelve step for effective test devolopment. In Handbook of Test Development*. Routledge.

Elvira, M., & Hadi, S. (2016). Karakteristik butir soal ujian semester dan kemampuan siswa sma di kabupaten muaro jambi. *Jurnal Evaluasi Pendidikan*, *4*(1), 1–23. http://journal.student.uny.ac.id/ojs/index.php/jep%0AKARAKTERISTIK

Gierl, M. J., Bulut, O., & Zhang, X. (2017). Developing, analyzing, and using distractors for multiple-choice testes in education: A Comprehensive Review. *Review of Educational Research*, *87*(6), 1082–1116. https://doi.org/https://doi.org/10.3102/0034654317726529

Hair, J. F., Black, W. C., Babin, B. J., & Anderson, B. J. (2010). *Multivariat data analisis a global perspertive* (7th ed.). Pearson.

Hair, J. F., Sarstedt, M., Hopkins, L., & Kuppelwieser, V. G. (2014). Partial least squares structural equation modeling (PLS-SEM): An emerging tool in business research. *European Business Review*, *26*(2), 106–121. https://doi.org/10.1108/EBR-10-2013-0128

Haladyna, T., & Rodriguez, M. (2013). *Developing and validating test item* (1st ed.). Routledge, Taylor & Francis Group.

Hendriyaldi. (2023). Strategi inovasi pengembangan kompetensi dosen menuju sumber daya manusia unggul pada pasca pandemi Covid-19 di Universitas Jambi. *Jurnal Paradigma Ekonomika*, *18*(4), 28–42.

Hu, Z., Lin, L., Wang, Y., & Li, J. (2021). The integration of classical testing theory and item response theory. *Psychology*, *12*, 1397–1409. https://doi.org/10.4236/psych.2021.129088

Ihda, M. F., & Heri, R. (2023). *Analysis of the distractor of the multiple-choice test using classical test theory (CTT) and item response theory (IRT)*. *Vi*, 196–203. https://doi.org/10.31643/2023.23

Irwantoro, N., & Suryanto, Y. (2016). *Kompetensi padagogik*. Genta Group Production.

Lane, S., Raymond, M., & Haladnya, T. (2016). *Handbook of test development* (2nd ed.). Raoutledge.

Lee, E., & Garg, N. (2020). Reliability of multiple-choice versus problem-solving student exam scores in higher education: Empirical tests. *International Conference on Higher Education Advances*, *2020-June*, 1399–1407. https://doi.org/10.4995/HEAd20.2020.11303

Mardapi, D. (2008). *Teknik penyusunan instrumen tes dan non tes*. Mitra Cendekia.

Osadebe, P. U., & Agbure, B. (2019). Assessment of differential item functioning in social studies multiple choice questions. *European Journal of Education Studies*, 312–344. https://doi.org/10.5281/zenodo.3674732

PUSPENDIK. (2016). *Panduan penulisan soal*. Kementrian pendidikan dan kebudayaan RI

Qiu, Z., Wu, X., & Fan, W. (2020). Automatic distractor generation for multiple choice questions in standard tests. *COLING 2020 - 28th International Conference on Computational Linguistics, Proceedings of the Conference*, 2096–2106. https://doi.org/10.18653/v1/2020.coling-main.189

Raina, V., Liusie, A., & Gales, M. (2024). *Assessing distractors in multiple-choice tests*. 12–22. https://doi.org/10.18653/v1/2023.eval4nlp-1.2

Retnawati, H. (2017). *Validitas reliabilitas & karakteristik butir*. Parama Publisher.

Rezigalla, A. A., Eleragi, A. M. E. S. A., Elhussein, A. B., Alfaifi, J., ALGhamdi, M. A., Al Ameer, A. Y., Yahia, A. I. O., Mohammed, O. A., & Adam, M. I. E. (2024). Item analysis: the impact of distractor efficiency on the difficulty index and discrimination power of multiple-choice items. *BMC Medical Education*, *24*(1), 1–7. https://doi.org/10.1186/s12909-024-05433-y

Sajjad, M., Iltaf, S., & Khan, R. A. (2020). Nonfunctional distractor analysis: An indicator for quality of multiple choice questions. *Pakistan Journal of Medical Sciences*, *36*(5), 982–986. https://doi.org/10.12669/pjms.36.5.2439

Sumaryanto. (2021). *Teori tes klasik & teori respon butir konsep & penerapannya* (1st ed.). CV. Confident.

Sumintono, B., & Widhiarso, W. (2019). *Aplikasi Model Rasch untuk penelitian ilmu-ilmu sosial*. Trikom Publishing House.

Suseno, I. (2017). Komparasi karakteristik butir tes pilihan ganda ditinjau dari teori tes klasik. *Jurnal Ilmiah Kependidikan*, *4*(1), 1–8.

Suwardjono. (2016). *Akuntansi pengantar*. Fakultas Ekonomika dan Bisnis UGM.

Team, R. C. (2024). *R: A language and environment for statistical computing. R Foundation for Statistical Computing*. https://www.r-project.org/

Tennant, A., & Küçükdeveci, A. A. (2023). Application of the rasch measurement model in rehabilitation research and practice: early developments, current practice, and future challenges. *Frontiers in Rehabilitation Sciences*, *4*(July). https://doi.org/10.3389/fresc.2023.1208670

Uyar, Ş., Kelecioğlu, H., & Doğan, N. (2017). Comparing differential item functioning based on manifest groups and latent classes. *Kuram ve Uygulamada Egitim Bilimleri*, *17*(6), 1977–2000. https://doi.org/10.12738/estp.2017.6.0526

Van der linden, W. J. (2016). *Handbook of item response theori*. CRC Press. https://doi.org/https://doi.org/10.1201/9781315374512

Widharyanto, B., & Prijowuntato, S. W. (2021). *Menilai peserta didik (Penyusunan instrumen penilaian)* (N. Premastuti Brataningrum, Ed.; 1st ed.). Sanata Dharma University Press.

Conflict of Interest Statement: The Author(s) declares that the research was conducted in the absence of any commercial or financial relationship that could be construed as a potential conflict of interest.

Measurement in Educational Research is an open access and peer-reviewed journal published bt Research and Social Study Institute, Indonesia

Open Access