

ANALISIS BUTIR SOAL SECARA MODERN DALAM EVALUASI PENDIDIKAN

Achmad Rosyid Ridho¹, Muhamad Abdul Azis², Mudzakir Chabib³

^{1,2,3}Institut Islam Mamba'ul 'Ulum Surakarta

* Corresponding Email: ahmadrosyeed@gmail.com

A B S T R A K

Analisis butir soal memegang peran penting dalam evaluasi pendidikan modern, memastikan butir soal valid dan reliabel dalam mengukur kompetensi siswa. Makalah ini membahas metodologi, aplikasi, dan manfaat analisis butir soal modern, dengan memanfaatkan teknologi canggih seperti Item Response Theory (IRT), computerized adaptive testing (CAT), dan analisis big data. Fokus utama meliputi indeks kesulitan, indeks daya pembeda, dan efektivitas distraktor, didukung oleh pendapat para ahli dan penggunaan perangkat lunak. Selain itu, makalah ini juga menyoroti tantangan dalam penerapan metode analisis modern, seperti akses teknologi, keahlian pengguna, dan pengelolaan data. Kesimpulan menunjukkan bahwa integrasi analisis butir soal modern secara signifikan meningkatkan kualitas evaluasi pendidikan.

Kata Kunci: Analisis Butir Soal, Evaluasi Modern, Evaluasi Pendidikan.

A B S T R A C T

Item analysis plays a pivotal role in modern educational evaluation, ensuring test items are valid and reliable in measuring student competencies. This paper delves into the methodologies, applications, and benefits of modern item analysis, leveraging advanced technologies like Item Response Theory (IRT), computerized adaptive testing (CAT), and big data analytics. It discusses essential aspects such as difficulty index, discrimination index, and distractor effectiveness, supported by expert opinions and software applications. Additionally, it highlights the challenges in adopting modern analysis methods, such as access to technology, user expertise, and data management issues. The paper concludes that integrating modern item analysis significantly improves the quality of educational assessments.

Keywords: Item Analysis, Modern Evaluation, Educational Assessment

PENDAHULUAN

Dalam sistem pembelajaran, evaluasi merupakan salah satu komponen penting dan tahap yang harus ditempuh oleh guru untuk mengetahui keefektifan pembelajaran (Arifin, 2009: 41). Menurut PP Nomor 19 Tahun 2005 tentang Standar Nasional Pendidikan, penilaian adalah proses pengumpulan dan pengolahan informasi untuk mengukur pencapaian hasil belajar peserta didik. Penilaian hasil belajar dilakukan berkesinambungan untuk memantau proses, kemajuan, dan perbaikan hasil dalam bentuk ulangan harian, ulangan tengah semester, ulangan akhir semester dan ulangan kenaikan kelas. Penilaian digunakan untuk menilai pencapaian kompetensi peserta didik, bahan penyusunan laporan kemajuan hasil belajar dan memperbaiki proses pembelajaran.

Penilaian hasil belajar oleh satuan pendidikan bertujuan menilai pencapaian standar kompetensi lulusan untuk semua mata pelajaran. Penilaian hasil belajar oleh

pemerintah dalam bentuk ujian nasional bertujuan untuk menilai pencapaian kompetensi lulusan secara nasional pada mata pelajaran tertentu dalam kelompok mata pelajaran ilmu pengetahuan dan teknologi. Ujian nasional dilakukan secara objektif, berkeadilan, dan akuntabel.

Menurut Hambleton dalam classical test theory telah digunakan selama ini untuk menentukan tingkat kesukaran dan karakteristik lainnya dalam instrumen pengukuran. Adanya kelemahan dalam classical test theory yang menonjol adalah indeks butir soal seperti tingkat kesukaran dan daya pembeda yang didapatkan dengan menggunakan classical test theory bergantung pada kelompok peserta tes dan penilaian kemampuan peserta tes bergantung pada pemilihan butir soal instrumen. Pengembangan teori pengukuran modern yang disebut item response theory (IRT) dalam rangka mengatasi kekurangan tersebut (Nahadi, 2011: 109).

Hal senada Hambleton et al. yang dikutip Nahadi, Wiwi Siswaningsih mengemukakan item response theory (IRT) adalah teori yang menyatakan bahwa hasil tes dapat diprediksi atau dijelaskan melalui serangkaian faktor yang disebut dengan sifat atau karakter (trait), karakter terpendam (latent trait) atau kemampuan (abilities) dan hubungan antara jawaban peserta tes dengan kemampuannya dapat dijelaskan dengan grafik karakteristik butir atau item characteristic curve (ICC). Semakin tinggi tingkat kemampuan, semakin besar peluang jawaban benar dari suatu butir soal. Sehingga dengan IRT \square item based \square atau lebih berorientasi pada tiap butir soal.

METODE PENELITIAN

Penelitian ini menggunakan pendekatan kuantitatif dengan metode deskriptif analitik untuk menganalisis kualitas butir soal evaluasi pembelajaran. Data penelitian berupa respons peserta didik terhadap butir soal yang diperoleh melalui dokumentasi hasil tes. Analisis butir soal dilakukan menggunakan teori tes modern (Item Response Theory/IRT) karena mampu memberikan estimasi parameter butir yang lebih akurat dan tidak bergantung pada kelompok responden (Hambleton, Swaminathan, & Rogers, 1991; Embretson & Reise, 2000).

Analisis data diawali dengan pengujian asumsi IRT yang meliputi unidimensionalitas, independensi lokal, dan invariansi parameter (De Ayala, 2009). Selanjutnya dipilih model IRT yang sesuai, yaitu model logistik satu parameter (1PL/Rasch), dua parameter (2PL), atau tiga parameter (3PL). Parameter butir yang dianalisis mencakup tingkat kesukaran, daya pembeda, dan faktor tebakan, yang digunakan sebagai dasar penentuan kualitas dan kelayakan butir soal dalam evaluasi pendidikan.

HASIL DAN PEMBAHASAN

A.Pengertian Analisis Butir Soal Secara Modern

Analisis butir soal modern atau Item Response Theory (IRT) merupakan salah satu cara untuk menilai kelayakan butir dengan membandingkan rerata penampilan butir terhadap tampilan bukti kemampuan kelompok yang diramalkan oleh model. Menurut Hambleton dan Swaminathan dalam Depdiknas menyatakan bahwa tujuan IRT untuk memberikan kesamaan antara statistik soal dan estimasi kemampuan peserta didik

(Depdiknas, 2007: 19).

Bagian penting dari teori respons butir adalah probabilitas jawaban benar peserta tes, parameter butir dan parameter peserta tes dihubungkan melalui suatu fungsi matematik atau model formula matematik. Dalam formula ini, nilai kemungkinan peserta tes menjawab soal dipahami sebagai fungsi logistik perbedaan parameter yang dimasukkan ke dalam model. Nama lain IRT adalah Latent Trait Theory (LTT), atau Characteristics Curve Theory (ICC).

Dengan demikian dapat dipahami bahwa IRT berhubungan antara probabilitas jawaban suatu butir soal yang benar dan kemampuan peserta didik atau tingkatan prestasi peserta didik.

B.Teorи Responsi Butir (Item Response Theory)

Sebagaimana disebutkan di atas, pada hakikatnya IRT bertujuan untuk mengatasi kelemahan yang terdapat pada pengukuran klasik. Pada IRT menurut Hambleton & Rogers yang dikutip (Sudaryono, 2011:611), peluang jawaban benar yang diberikan siswa, ciri atau parameter butir, dan ciri atau parameter peserta tes dihubungkan melalui suatu model formula yang harus ditaati baik oleh kelompok butir tes maupun kelompok peserta tes.

Perbedaan mendasar antara pengukuran klasik dengan pengukuran modern terletak pada invariansi pensekoran, di mana pensekoran modern adalah invariansi (tidak berubah atau tetap) terhadap butir tes serta terhadap peserta tes. Menurut Lord (1990: 121) bahwa invariansi parameter-parameter butir tes melalui kelompok peserta tes merupakan karakteristik yang paling penting dari IRT.

Teori responsi butir perlu menentukan model karakteristik butir yang digunakan. Model karakteristik butir dapat berbentuk satu parameter (1P), dua parameter (2P), tiga parameter (3P), atau model lain. Di sini pembahasan dibatasi pada satu sampai tiga parameter serta pada sekor dikotomi, yaitu: 1P : $P(q) = f(b, q)$ 2P : $P(q) = f(a, b, q)$ dan 3P : $P(q) = (a, b, c, q)$. Satu, dua, dan tiga adalah banyaknya parameter butir. Parameter q adalah parameter kemampuan responden. Parameter b adalah parameter taraf sukar butir. Pada 1P dan 2P, $b = q$ ketika $P(q) = 0,5$. Pada 3P, $b = q$ ketika $P(q) = 0,5(1 + c)$. Parameter a adalah parameter daya beda butir. Parameter c adalah parameter terkaan betul jawaban butir.

1. Tujuan Responsi Butir

Teori responsi butir membebaskan responden dan butir dari interdependensi, sehingga taraf butir sukar tidak lagi bergantung kepada kemampuan responden. Kemampuan responden tidak bergantung pada taraf butir sukar. Melalui independensi diantara taraf butir sukar dan kemampuan responden, dapat dipilih butir yang cocok dengan responden.

Proporsi jawaban benar di dalam sebuah kelompok peserta tes tidak secara nyata mengukur kesulitan tes tersebut. Proporsi tersebut tidak hanya menjelaskan butir tes tetapi juga kelompok peserta yang dites. Ini merupakan suatu tujuan dasar untuk kesepakatan analisis statistik butir tes, yang dikenal dengan istilah invariansi. Yang menjadi dasar invariansi adalah taraf butir sukar tidak langsung dikaitkan dengan kemampuan responden melainkan dikaitkan dengan lengkungan karakteristik butir pada persamaan : $P(q) = P_{\min} + (1 - P_{\min})$

2. Asumsi Teori Responsi Butir

Dalam teori responsi butir taraf butir sukar dan daya beda butir tes tetap sama, walaupun butir tes tersebut diselesaikan oleh kelompok peserta tes yang berbeda. Untuk itu, teori responsi butir mengembangkan model yang menghubungkan parameter butir dengan kemampuan peserta tes. Menurut Hambleton yang dikutip Sudaryono (2011:611) asumsi untuk model teori responsi butir secara mendalam digunakan, sehingga hanya satu kemampuan yang diukur dengan butir-butir tes tersebut. Hal ini dinamakan unidimensi. Suatu konsep yang menghubungkan keunidimensian adalah apa yang disebut dengan independensi lokal (local independence) dalam bahasan berikut ini.

1) Unidimensi

Asumsi unidimensi terpenuhi apabila butir-butir di dalam perangkat tes hanya mengukur satu kemampuan peserta tes. Misalnya butir-butir yang termuat di dalam perangkat tes bertujuan untuk mengukur kemampuan peserta tes dalam mata pelajaran.

Menurut Dali S. Naga (1992:1640) bahwa persyaratan unidimensi ditujukan untuk mempertahankan invariansi pada teori responsi butir. Jika butir tes sampai mengukur lebih dari satu dimensi, maka jawaban terhadap butir itu merupakan kombinasi dari berbagai kemampuan peserta tes, akibatnya, tidak diketahui kontribusi dari setiap kemampuan terhadap jawaban peserta tes tersebut.

2) Independensi Lokal

Asumsi independensi lokal menurut James J. Allen & Yen dibagi menjadi dua yaitu independensi lokal terhadap respons peserta tes dan independensi lokal terhadap butir tes. Independensi lokal terhadap respons peserta tes, memiliki arti bahwa betul salahnya peserta tes menjawab sebuah butir tidak terpengaruh oleh betul salahnya peserta tes yang lain dalam menjawab butir tersebut. Sedangkan indepensi lokal terhadap butir, memiliki arti bahwa betul salahnya seorang peserta tes menjawab sebuah butir tidak terpengaruh oleh betul salahnya peserta tes dalam menjawab butir yang lain.

Menurut Crocker dan Algina (1986: 342), dalam teori responsi butir secara bersama-sama digunakan konsep-konsep yang lebih umum terhadap keterikatan dan kebebasan statistik untuk menyatakan tentang hubungan antara varaiabel-variabel.

3) Invarian

Item Response Theory (IRT) bertujuan untuk mengatasi kelemahan pengukuran klasik, perbedaan mendasar antara pengukuran klasik dengan pengukuran modern terletak pada invariansi pensekoran, di mana pensekoran modern adalah invarians (tidak berubah) terhadap butir tes serta terhadap peserta tes. Menurut Lord (1990:126) bahwa invariansi parameter-parameter butir tes melalui kelompok peserta tes merupakan karakteristik yang paling penting dari IRT. Dapat dikatakan bahwa indeks kesukaran butir tes sebagai proporsi jawaban yang benar sehingga sukar untuk membayangkan bagaimana indeks kesukaran tes dapat menjadi invarian terhadap kelompok peserta tes dari tingkat kemampuan yang berbeda.

3. Karakteristik butir

1) Taraf Butir Sukar

Pada umumnya makin mudah butir atau makin kecil b , maka makin besar probabilitas responden untuk menjawab butir itu dengan benar sehingga nilai $P()$ menjadi besar. Sebaliknya makin sukar butir atau makin besar b , maka makin kecil probabilitas

responden menjawab butir itu dengan benar sehingga nilai menjadi kecil. Dengan demikian mudah sukaranya suatu butir menurut Dali S. Naga (1992: 34) sering dikaitkan dengan kemampuan responden dengan taraf butir sukar yakni dengan.

2) Daya Beda Butir

Ada butir yang memiliki ciri: dapat dijawab dengan betul oleh kebanyakan responden yang berkemampuan tinggi, tidak dapat dijawab dengan betul oleh kebanyakan responden yang berkemampuan rendah. Butir demikian memiliki daya untuk membedakan responden berdasarkan kemampuan mereka. Butir memiliki parameter berupa daya beda butir. Daya beda butir adalah kemampuan suatu butir soal dapat membedakan antara peserta didik atau warga belajar yang telah menguasai materi yang ditanyakan dan warga belajar atau peserta didik yang belum menguasai materi yang ditanyakan.

3) Tingkat Kebetulan Betul pada Butir

Dalam bentuk probabilitas, bahwa tingkat kebetulan pada jawaban benar adalah c , maka untuk butir ini, probabilitas jawaban benar karena kebetulan. Kalau jumlah pilihan ganda itu adalah empat (misalkan A, B, C, D), maka melalui terkaan saja terdapat 1 di antara 4 kemungkinan bahwa jawaban itu benar.

Dalam hal ini probabilitas jawaban benar karena kebetulan adalah $\frac{1}{4}$ atau 0,25 sehingga $c = 0,25$. Pada butir pilihan ganda dapat saja terjadi bahwa jawaban betul dicapai melalui terkaan. Jawaban betul ini adalah kebetulan betul. Tingkat kebetulan menjawab betul pada butir ke- i dinyatakan dengan parameter butir c_i dan merupakan probabilitas jawaban betul minimum (Dali, S., 1992: 37).

C. Penyusunan Hasil Tes Belajar

Menurut Kumaidi (2000: 102) untuk mengembangkan suatu tes dan sejumlah butir soal yang \square defensible \square maka prosedur pengembangan perlu ditradisikan, dalam arti proses pengembangan tes (dan penulisan butir soal) dimulai dengan pengembangan rancangan atau kisi-kisi tes, yang didahului oleh pembedahan kurikulum yang memuat segala informasi tentang tes tersebut. Rancangan tes ini memuat tujuan penilaian yang akan dilakukan, tempo (waktu yang ditempuh) untuk pelaksanaan pengujian, pesan utama kurikulum (sasaran pembelajaran dan garis besar topik materi uji), indikator butir soal (ciri-ciri penguasaan materi uji dan pencapaian sasaran pembelajaran), serta jumlah dan bentuk butir soal (per-indikator, per topik, dan keseluruhan tes). Sebaran butir soal dalam tes seharusnya memperhatikan keseimbangan tuntutan penguasaan sesuai dengan pesan kurikulum, sehingga memberi nuansa keterwakilan topik bahasan.

Menurut Jihad, ada sembilan langkah yang harus ditempuh dalam mengembangkan tes hasil atau prestasi belajar, yaitu: 1) menyusun spesifikasi tes; 2) menulis soal tes; 3) menelaah soal tes; 4) melakukan uji coba tes; 5) menganalisis butir soal; 6) memperbaiki tes; 7) merakit tes; 8) melaksanakan tes; dan 9) menafsirkan hasil tes. Khusus mengenai uji coba tes, dalam penyusunan tes untuk mengukur prestasi hasil pembelajaran yang diselenggarakan oleh guru di kelas seperti ulangan harian, ulangan umum, dan ulangan kenaikan kelas, tidak harus dilakukan secara tersendiri. Pembuatan tes dilakukan melalui beberapa kali ujicoba. Sedangkan Djaali (2011: 187) menjelaskan bahwa, penyusunan dan pengembangan tes dimaksudkan untuk memperoleh tes yang valid, sehingga hasil ukurnya dapat mencerminkan secara tepat hasil belajar yang dicapai

oleh masing-masing individu peserta tes setelah selesai mengikuti pembelajaran. Adapun langkah-langkah kontruksi tes yang ditempuh adalah sebagai berikut: 1) menetapkan tujuan tes; 2) analisis kurikulum; 3) analisis buku pelajaran dan sumber materi belajar lainnya; 4) membuat kisi-kisi; 5) penulisan tujuan instruksional khusus; 6) penulisan soal; 7) telaah soal (face validity); 8) reproduksi tes terbatas; 9) uji coba tes; 10) analisis hasil uji coba; 11) revisi soal, dan 12) merakit soal menjadi tes.

Langkah awal dalam mengembangkan tes adalah menetapkan spesifikasi tes, yaitu berisi uraian yang menunjukkan keseluruhan karakteristik yang harus dimiliki suatu tes. Spesifikasi yang jelas akan mempermudah dalam menulis soal, dan siapa saja yang menulis soal akan menghasilkan tingkat kesulitan yang relatif sama. Penyusunan spesifikasi tes mencakup kegiatan berikut ini: 1) menentukan tujuan tes; 2) menyusun kisi-kisi tes; 3) memilih bentuk tes; dan 4) menentukan panjang tes.

Pengembangan rancangan tes ini melibatkan spesialis (termasuk guru) bidang studi, sehingga bila rancangan tes telah selesai disusun maka rancangan tes tersebut harus divalidasi, melalui penelaahan pakar dan teman sejawat, sehingga benar-benar sesuai dengan pesan kurikulum. Untuk mengatasi variasi butir soal yang berlebihan, dengan pemahaman indikator butir soal, ada baiknya dikembangkan apa yang disebut oleh Nitko sebagai spesifikasi butir soal (item specification). Spesifikasi ini menyangkut uraian tentang batasan dan rambu-rambu yang harus dipatuhi oleh penulis butir soal.

Gronlund dalam Sudaryono (1992: 615) menyarankan beberapa hal dalam pengkonstruksian tes, diantaranya: 1) stem item tersebut sebaiknya memaknai butir itu sendiri dan menampilkan masalah tertentu; 2) stem butir tes melibatkan banyak kemungkinan jawaban dan bebas dari materi yang tidak relevan; 3) gunakan pernyataan stem butir yang bersifat negatif hanya ketika hasil belajar yang dikehendaki cukup berarti (signifikan); 4) Semua alternatif jawaban secara gramatikal konsisten dengan stem butir tersebut; 5) sebuah butir secara jelas hanya mengandung satu jawaban benar terbaik; 6) butir-butir tes digunakan untuk mengukur pemahaman yang mengandung beberapa hal baru, tetapi harus berhati-hati; 7) semua pengecahan harus masuk akal; 8) assosiasi verbal antara stem dan jawaban yang benar harus dihindarkan; 9) secara relatif, panjang pilihan jawaban tidak menunjukkan suatu petunjuk untuk jawaban tersebut; 10) jawaban benar sebaiknya muncul pada masing-masing posisi pilihan atas beberapa kesamaan pendekatan, tetapi dalam urutan random; 11) gunakan dengan hemat pilihan-pilihan khusus seperti tidak satu pun jawaban di atas benaratau semua jawaban di atas benar; dan 12) jangan gunakan butir-butir pilihan berganda ketika butir yang lainnya lebih tepat.

Suatu tes harus mengukur hasil belajar dalam skala yang sama dan pendekatan yang mungkin dilakukan antara lain: 1) pemakaian butir soal penjangkar (common items) untuk beberapa set tes; 2) pemakaian butir soal yang telah terkalibrasi (butir soal yang diketahui karakteristiknya pada satu skala umum); dan 3) kombinasi kedua pendekatan itu yakni soal penjangkar dipilih dari butir yang terkalibrasi. Dalam hal ini peranan IRT cukup berguna untuk menyamakan skala tersebut. Setelah soal-soal berkualitas terpilih berdasarkan professional adjustment dari para ahli bidang studi dan ahli pengukuran (measurement specialist) dan juga didukung data empirik hasil uji coba soal, maka kegiatan berikutnya adalah membuat skala dan menentukan di mana setiap soal terletak dalam skala tersebut (Kumaidi, 2000: 105).

Dalam penelitian digunakan teori skor modern membentuk perangkat soal yang baik dibutuhkan banyak hal, terutama dari aspek esensial yang membutuhkan pengkajian lebih mendasar dan mendetail.

SIMPULAN DAN SARAN

Dari uraian makalah di atas dapat pemakalah simpulkan sebagai berikut:

1. Bahwa analisis butir soal modern pada hakekatnya proses pengukuran yang memahami perbedaan dari responsi tes dan butir tes sukar sehingga pengukuran dapat dilakukan berdasarkan kelompok tertentu
2. Bahwa penskoran klasik sebagai pembanding karena dalam pengukuran klasik responsi tes diperlakukan sama tanpa ada perbedaan, maka dengan analisis modern IRT mengubah keterkaitan responsi butir sehingga tidak terlalu mengikat.
3. Bahwa teori responsi butir Item Response Theory bertujuan untuk mengatasi kelemahan yang terdapat pada pengukuran klasik, peluang jawaban benar yang diberikan siswa, ciri atau parameter butir, dan ciri atau parameter peserta tes dihubungkan melalui suatu model formula yang harus ditaati baik oleh kelompok butir tes maupun kelompok peserta tes, pengukuran modern, taraf butir sukar dikaitkan langsung dengan karakteristik butir, model karakteristik butir dapat berbentuk satu parameter,
4. Bahwa penyusunan hasil tes belajar yang sangat beragam di, seharusnya digunakan juga ujian atau tes yang berbeda tingkat kesukaran soalnya, supaya adil dan juga akurat hasilnya. Peserta tes atau ujian (seperti Ujian Nasional) yang mengerjakan tes atau ujian yang berbeda tingkat kesukaran soalnya, tetap bisa dibandingkan kemampuannya, asalkan soal-soal dalam ujian tersebut berasal atau diambil dari bank soal yang sudah dikalibrasi dengan konsep item response theory.

DAFTAR PUSTAKA

- Arifin. Z., Evaluasi Pembelajaran, Bandung : PT Remaja Rosdakarya, 2009.
- Crocker. Linda, & Algina, James. Introduction to classical and modern test theory. New York: Holt, Rinehart and Winston, Inc. 1986.
- Dali, S. Naga. Pengantar Teori Sekor Pada Pengukuran Pendidikan. Jakarta: Besbats, 1992.
- Depdiknas. Standar Proses untuk Satuan Pendidikan Dasar dan Menengah, Jakarta : Badan Standar Nasional Pendidikan, 2007.
- Frederick, M. Lord. Applications of Item Response Theory to Practical Testing Problems. New Jersey: Lawrence Erlbaum Associates, Publishers, 1990.
- Jihad. Asep, Abdul Haris. Evaluasi Pembelajaran. Multi Pressindo: Yogyakarta, 2011.
- Kumaidi. Standardisasi Butir Soal. Jurnal Pendidikan dan Kebudayaan. V(5), 2000.
- Nahadi, Wiwi Siswaningsih & Ana Rofiat, Pengembangan dan Analisis Soal Ulangan Kenaikan Kelas Kimia SMA Kelas X Berdasarkan Clasical Test Theory dan Item Response Theory, Jurnal Pengajaran IPA, Volume 16, Nomor 2, Oktober 2011.
- Peraturan Pemerintah nomor 19 tahun 2005 tentang Standar Nasional Pendidikan.
- Sudaryono, Implementasi Teori Responsi Butir (Item Response Theory) Pada Penilaian Hasil Belajar Akhir di Sekolah, Jurnal Pendidikan dan Kebudayaan, Vol. 17, Nomor 6, Nopember 2011.