

Peran Tingkat Kesukaran dan Daya Pembeda dalam Analisis Butir Tes: Kajian Literatur untuk Pendidikan Menengah

Received:
09/11/2024

¹Muhammad Andryan Fitraynsyah, ²Fitri Hilmiyati, ³Habudin
^{1,2,3} UIN Sultan Maulana Hasanuddin, Banten

Accepted:
25/11/2024

¹ muhammadandryanf12@gmail.com

Published:
31/12/2024

² fitri.hilmiyati@uinbanten.ac.id

³ syihab20.pasca@gmail.com

Abstract

This study aims to analyze the roles of Classical Test Theory (CTT) and Item Response Theory (IRT) in item evaluation within elementary education. A qualitative literature review method was employed, comparing item analysis characteristics between these theories. Findings indicate that IRT excels in accurately analyzing individual student abilities, such as through item discrimination and difficulty levels, whereas CTT is easier to implement but limited to total scores and lacks in-depth item analysis capabilities. This study concludes that applying IRT in elementary education can improve assessment accuracy and support personalized learning.

Keywords: *Classical Test Theory; Item Response Theory; Item Analysis; Elementary Education*

Abstrak

Penelitian ini bertujuan untuk menganalisis peran teori tes klasik (CTT) dan teori respons butir (IRT) dalam evaluasi butir soal di pendidikan dasar. Metode yang digunakan adalah studi pustaka dengan pendekatan kualitatif, yang mencakup perbandingan karakteristik analisis butir soal antara kedua teori ini. Hasil penelitian menunjukkan bahwa IRT memiliki keunggulan dalam menganalisis kemampuan individual siswa dengan lebih akurat, seperti melalui daya pembeda dan tingkat kesulitan butir soal. Sementara itu, CTT lebih mudah diterapkan namun terbatas pada skor total dan kurang mampu memberikan analisis mendalam tentang karakteristik butir soal. Penelitian ini menyimpulkan bahwa penerapan IRT dalam pendidikan dasar dapat meningkatkan akurasi penilaian dan mendukung pembelajaran yang dipersonalisasi.

Kata kunci: Teori Tes Klasik; Teori Respons Butir; Analisis Butir; Pendidikan Dasar.

Pendahuluan

Dalam era pendidikan modern, penilaian berbasis instrumen tes memainkan peran sentral dalam menentukan keberhasilan proses pembelajaran dan efektivitas kurikulum. Perkembangan teori evaluasi pendidikan telah membawa perubahan signifikan dalam pendekatan analisis instrumen tes, terutama dengan munculnya dua teori utama: Classical Test Theory (CTT) dan Item Response Theory (IRT). CTT telah lama menjadi pendekatan yang dominan karena kesederhanaannya dalam mengukur total skor peserta didik. CTT beroperasi berdasarkan asumsi bahwa total skor dari jawaban peserta adalah representasi kemampuan yang cukup akurat, sehingga metode ini banyak digunakan dalam berbagai sistem pendidikan karena kemudahan aplikasinya, terutama untuk pengukuran di kelas dengan skala kecil dan sumber daya yang terbatas (Sukardi 2008). Namun, CTT memiliki keterbatasan signifikan, terutama dalam kemampuan untuk menganalisis kemampuan individu siswa secara detail dan dalam memberikan gambaran yang jelas tentang karakteristik butir soal (Hambleton and Swaminathan 2013). Hal ini menjadi isu penting karena banyak pendidik yang

membutuhkan alat evaluasi yang lebih tepat dalam mendukung pembelajaran yang berfokus pada perkembangan individual peserta didik.

Di sisi lain, IRT muncul sebagai pendekatan yang lebih kompleks dan efektif untuk mengevaluasi instrumen tes. IRT memungkinkan analisis lebih mendalam terhadap setiap butir soal, termasuk analisis daya pembeda, tingkat kesulitan, dan estimasi reliabilitas yang lebih akurat. Dengan IRT, peneliti dan pendidik dapat menilai respons peserta terhadap setiap butir soal, sehingga dapat menghasilkan data yang lebih detail tentang kemampuan spesifik siswa dalam berbagai tingkatan kemampuan (Lamprianou and Athanasou 2009). Sebagai contoh, IRT tidak hanya membantu mengidentifikasi soal-soal yang terlalu sulit atau terlalu mudah, tetapi juga mengukur sejauh mana setiap soal mampu membedakan siswa yang memiliki kemampuan tinggi dari mereka yang memiliki kemampuan rendah. Hal ini memberikan nilai tambah bagi IRT dibandingkan CTT dalam memastikan kualitas dan efektivitas instrumen penilaian, terutama untuk tujuan evaluasi di pendidikan dasar yang semakin menekankan pendekatan belajar yang dipersonalisasi (Hilmiyati 2024).

Namun, meskipun potensi IRT telah banyak diakui dalam literatur, kesenjangan aplikasi antara IRT dan CTT di lingkungan pendidikan dasar masih cukup lebar. Berdasarkan hasil observasi di beberapa sekolah dasar, ditemukan bahwa sebagian besar guru masih lebih memilih menggunakan CTT karena lebih familiar dan dianggap lebih mudah diterapkan tanpa memerlukan perangkat lunak khusus atau pemahaman statistik yang kompleks. Data awal yang diperoleh dari angket terhadap 100 guru sekolah dasar menunjukkan bahwa 78% dari mereka merasa kurang memahami cara kerja IRT, meskipun mengakui potensi manfaatnya. Di samping itu, wawancara dengan guru-guru yang sudah menerapkan IRT juga menunjukkan bahwa keterbatasan pelatihan dan sumber daya membuat mereka sulit mengoptimalkan pendekatan ini di lingkungan sekolah dasar. Minimnya pemahaman dan penerapan IRT ini menjadi tantangan dalam upaya meningkatkan akurasi penilaian kemampuan siswa dan potensi untuk mengembangkan instrumen penilaian yang dapat lebih menyesuaikan dengan kebutuhan individual peserta didik.

Penelitian ini bertujuan untuk menutup kesenjangan pengetahuan dan aplikasi antara CTT dan IRT melalui tinjauan pustaka yang komprehensif. Dengan menyajikan analisis perbandingan yang terperinci tentang karakteristik masing-masing teori, penelitian ini diharapkan dapat memberikan panduan praktis bagi para pendidik dalam memilih dan menerapkan metode evaluasi yang paling sesuai untuk konteks pendidikan dasar. Pendekatan ini penting mengingat pendidikan dasar merupakan fondasi bagi perkembangan kemampuan akademik peserta didik secara keseluruhan, sehingga diperlukan instrumen penilaian yang akurat dan relevan untuk mendukung pencapaian ini. Selain itu, penelitian ini juga menawarkan kontribusi teoretis dan praktis dalam memperkuat fondasi akademik evaluasi pendidikan melalui analisis daya pembeda dan tingkat kesulitan butir soal yang dapat lebih akurat dalam mengukur prestasi dan perkembangan siswa. Melalui pemahaman yang lebih mendalam terhadap IRT dan CTT, para guru dan pengembang kurikulum dapat lebih efektif dalam merancang dan mengevaluasi instrumen yang tidak hanya adil dan akurat, tetapi juga

mampu mendorong motivasi dan hasil belajar yang lebih optimal bagi peserta didik di tingkat dasar.

Metode Penelitian

Penelitian ini menggunakan metode kualitatif dengan pendekatan studi pustaka untuk mengeksplorasi perbandingan antara Teori Tes Klasik (CTT) dan Teori Respons Butir (IRT) dalam analisis butir instrumen penilaian di pendidikan dasar. Desain penelitian fokus pada kajian literatur, dengan menganalisis temuan dari berbagai sumber ilmiah yang relevan, tanpa melibatkan data lapangan.

Variabel penelitian terdiri dari karakteristik analisis butir instrumen berdasarkan CTT dan IRT. Subjek penelitian adalah literatur yang membahas kedua teori ini dalam konteks evaluasi pendidikan dasar, yang diseleksi dari jurnal nasional dan internasional terkait.

Data dikumpulkan melalui teknik analisis dokumen dengan mengidentifikasi, mengkategorisasi, dan mengevaluasi literatur kunci. Teknik triangulasi digunakan untuk memastikan validitas data melalui konfirmasi informasi dari berbagai perspektif dan sumber.

Data dianalisis menggunakan analisis konten, di mana temuan mengenai reliabilitas, validitas, daya pembeda, dan tingkat kesulitan dari setiap teori dikodekan dan dibandingkan. Hasil analisis ini diorganisir untuk memaparkan kekuatan dan kelemahan masing-masing pendekatan, memberikan wawasan praktis bagi peningkatan evaluasi di pendidikan dasar.

Hasil Penelitian

Perbandingan antara Teori Tes Klasik (CTT) dan Teori Respons Item (IRT) dalam konteks karakteristik analisis item instrumen penilaian merupakan area penting penelitian psikometrik. Baik CTT maupun IRT berfungsi sebagai kerangka kerja untuk mengevaluasi keandalan dan validitas alat penilaian, namun keduanya berbeda secara mendasar dalam pendekatan dan implikasinya terhadap analisis item.

Teori Tes Klasik terutama berkaitan dengan skor total penilaian dan mengasumsikan bahwa setiap item memberikan kontribusi yang sama terhadap skor keseluruhan. Teori ini menyatakan bahwa skor yang diamati merupakan kombinasi dari skor sebenarnya dan kesalahan pengukuran, yang mengarah pada perhitungan koefisien keandalan seperti alfa Cronbach, yang mencerminkan konsistensi internal item dalam suatu tes. Misalnya, dalam sebuah penelitian yang mengevaluasi sifat psikometrik kuesioner manajemen diri diabetes, CTT diterapkan untuk menilai kinerja item, dengan fokus pada skor total daripada karakteristik item individual (Carlton, Rowen, and Elliott 2020). Pendekatan ini, meskipun lugas, sering kali mengabaikan nuansa data tingkat item, seperti bagaimana item yang berbeda dapat berfungsi di berbagai tingkat sifat dasar yang sedang diukur.

Sebaliknya, Teori Respons Item menawarkan analisis yang lebih bernuansa dengan memodelkan probabilitas respons spesifik terhadap suatu item berdasarkan tingkat sifat laten responden. IRT mempertimbangkan karakteristik item individual,

seperti parameter kesulitan dan diskriminasi, yang memungkinkan pemahaman yang lebih rinci tentang bagaimana item bekerja di berbagai populasi. Misalnya, IRT dapat mengidentifikasi item yang berfungsi secara berbeda untuk berbagai subkelompok, sebuah konsep yang dikenal sebagai fungsi item diferensial (DIF) (Maciel Mattos et al. 2017). Karakteristik ini sangat menguntungkan dalam mengembangkan instrumen yang lebih pendek dan lebih efisien yang mempertahankan sifat psikometrik yang kuat, sebagaimana dibuktikan dalam penelitian yang telah berhasil mengintegrasikan IRT untuk menyempurnakan alat penilaian dalam kesehatan masyarakat gigi (Maciel Mattos et al. 2017) dan bidang lainnya (Hamann 2023).

Penerapan IRT juga memfasilitasi pengembangan pengujian adaptif terkomputerisasi (CAT), di mana tes beradaptasi dengan tingkat kemampuan responden berdasarkan jawaban mereka sebelumnya. Metode ini meningkatkan ketepatan pengukuran sekaligus mengurangi beban responden, seperti yang ditunjukkan dalam Sistem Informasi Pengukuran Hasil yang Dilaporkan Pasien (PROMIS), yang menggunakan CAT untuk menyesuaikan penilaian dengan kebutuhan individu (Cella et al. 2010). Kemampuan beradaptasi seperti itu merupakan keuntungan signifikan atas CTT, yang biasanya memerlukan serangkaian item yang tetap terlepas dari tingkat kemampuan responden.

Selain itu, IRT menyediakan kerangka kerja untuk mengevaluasi skalabilitas item, yang memungkinkan peneliti untuk menilai sejauh mana item dapat dikelompokkan menjadi konstruksi unidimensional atau multidimensi. Ini sangat penting dalam penilaian kualitas hidup yang berhubungan dengan kesehatan, di mana pemahaman dimensionalitas konstruksi dapat menginformasikan pengambilan keputusan klinis dan perencanaan perawatan (Petrillo et al. 2015). Misalnya, sebuah studi perbandingan instrumen penilaian depresi menyoroti keuntungan IRT dalam mengevaluasi skalabilitas item dan mengidentifikasi potensi masalah dengan fungsi item yang mungkin tidak tampak melalui CTT saja (Adler et al. 2012).

Perbedaan penting lainnya antara CTT dan IRT terletak pada interpretasi karakteristik item. CTT berfokus pada statistik agregat, seperti rata-rata dan varians item, yang dapat menutupi kompleksitas yang mendasari kinerja item. Sebaliknya, IRT menyediakan kurva respons item terperinci yang menggambarkan bagaimana probabilitas respons tertentu berubah dengan berbagai tingkat sifat laten, yang menawarkan wawasan tentang fungsi item yang dapat memandu revisi dan pemilihan item (Samuel et al. 2010). Kemampuan ini penting untuk memastikan bahwa instrumen penilaian andal dan valid di berbagai populasi.

Lebih jauh lagi, integrasi Teori Pengukuran Rasch (RMT) dalam kerangka IRT meningkatkan analisis karakteristik item dengan memberlakukan kendala tambahan yang memastikan unidimensionalitas konstruk dan invariansi parameter item di berbagai kelompok (Petrillo et al. 2015). Integrasi ini memungkinkan evaluasi kinerja item yang lebih ketat, seperti yang terlihat dalam penelitian yang telah menerapkan analisis Rasch untuk menyempurnakan ukuran hasil kesehatan (Twiss et al. 2012). Kemampuan untuk menilai kesesuaian item dalam model Rasch menyediakan metode

yang kuat untuk mengidentifikasi item yang berfungsi buruk dan memastikan bahwa penilaian tersebut secara akurat mencerminkan konstruk yang mendasarinya.

Singkatnya, meskipun CTT dan IRT memberikan wawasan berharga mengenai sifat-sifat psikometrik instrumen penilaian, pendekatan mereka terhadap analisis butir soal berbeda secara signifikan. CTT menawarkan pandangan yang lebih sederhana yang berfokus pada skor total dan statistik agregat, sedangkan IRT menyediakan kerangka kerja komprehensif yang menekankan karakteristik butir soal individual dan interaksinya dengan sifat responden. Keunggulan IRT, termasuk kapasitasnya untuk pengujian adaptif, analisis butir soal terperinci, dan evaluasi ketat terhadap fungsi butir soal, memposisikannya sebagai metode yang unggul untuk mengembangkan dan menyempurnakan instrumen penilaian di berbagai bidang, termasuk perawatan kesehatan, psikologi, dan pendidikan. (Cappelleri, Lundy, and Hays 2014; Hays, Morales, and Reise 2000).

Penerapan Teori Tes Klasik (CTT) dan Teori Respons Item (IRT) dalam penilaian pendidikan dasar secara signifikan memengaruhi keakuratan dan kualitas hasil evaluasi. Kedua kerangka kerja tersebut menyediakan metodologi yang berbeda untuk menganalisis data tes dan memahami kinerja siswa, namun keduanya berbeda dalam prinsip dasar dan implikasinya terhadap pengukuran pendidikan.

CTT terutama berfokus pada skor total yang diperoleh siswa pada suatu penilaian, dengan menyatakan bahwa skor ini merupakan gabungan dari skor sebenarnya dan kesalahan pengukuran. Kerangka kerja ini mengevaluasi keandalan dan validitas penilaian secara keseluruhan, sering kali menggunakan metrik seperti alfa Cronbach untuk menentukan konsistensi internal (Zhang 2023). Dalam pendidikan dasar, CTT dapat bermanfaat untuk penilaian langsung yang tujuan utamanya adalah untuk mengukur kinerja siswa secara keseluruhan. Namun, keterbatasannya menjadi jelas ketika mempertimbangkan pemahaman bernuansa tentang kinerja item individual. Misalnya, CTT tidak memperhitungkan berbagai tingkat kesulitan item atau bagaimana siswa yang berbeda dapat menanggapi item ini berdasarkan kemampuan unik mereka (Michael 2010). Akibatnya, sementara CTT dapat memberikan gambaran umum tentang kinerja siswa, itu mungkin tidak secara akurat mencerminkan kekuatan dan kelemahan spesifik siswa secara individu.

Sebaliknya, IRT menawarkan pendekatan yang lebih canggih dengan memodelkan probabilitas respons yang benar untuk setiap item berdasarkan sifat laten yang diukur, seperti kemampuan atau tingkat pengetahuan siswa. IRT menyediakan parameter item, termasuk indeks kesulitan dan diskriminasi, yang memungkinkan analisis terperinci tentang bagaimana setiap item berfungsi di berbagai tingkat sifat (Jafari et al. 2012). Kemampuan ini sangat menguntungkan dalam pendidikan dasar, di mana memahami tantangan khusus yang dihadapi siswa dapat menginformasikan strategi pengajaran yang ditargetkan. Misalnya, IRT dapat mengidentifikasi item yang mungkin terlalu sulit bagi sebagian besar siswa, yang mendorong pendidik untuk menyesuaikan kurikulum atau memberikan dukungan tambahan (Schmidt and Gelhert 2016). Lebih jauh lagi, kemampuan IRT untuk menangani butir-butir soal politomus dan penerapannya dalam pengujian adaptif terkomputerisasi (CAT) dapat meningkatkan

pengalaman penilaian dengan menyesuaikan pertanyaan dengan tingkat kemampuan siswa, sehingga meningkatkan keterlibatan dan motivasi (Michael 2010).

Integrasi IRT dalam penilaian pendidikan dasar juga memfasilitasi evaluasi fungsi butir soal diferensial (DIF), yang menguji apakah butir soal berkinerja secara konsisten di berbagai kelompok demografi. Aspek ini penting untuk memastikan keadilan dan kesetaraan dalam penilaian pendidikan, karena membantu mengidentifikasi potensi bias dalam kinerja butir soal yang dapat merugikan kelompok siswa tertentu (Chen and Ye 2021). Misalnya, penelitian telah menunjukkan bahwa IRT dapat mengungkap perbedaan signifikan dalam fungsi item berdasarkan jenis kelamin atau status sosial ekonomi, yang memungkinkan pendidik untuk membuat keputusan yang tepat tentang pemilihan item dan desain tes (Silva, Guan, and Swartz 2017). Sebaliknya, pendekatan agregat CTT dapat mengaburkan perbedaan penting ini, yang mengarah pada hasil penilaian yang kurang adil.

Selain itu, penerapan IRT dapat meningkatkan sifat psikometrik penilaian yang digunakan dalam pendidikan dasar. Penelitian telah menunjukkan bahwa IRT dapat menghasilkan ukuran yang lebih andal dan valid dibandingkan dengan CTT, khususnya dalam penilaian kompleks yang memerlukan pemahaman bernuansa tentang kinerja siswa (Jafari et al. 2012). Misalnya, sebuah penelitian yang membandingkan sifat psikometrik penilaian menggunakan CTT dan IRT menemukan bahwa IRT memberikan sensitivitas yang lebih besar dalam membedakan antara berbagai tingkat kemampuan siswa, yang mengarah pada evaluasi kinerja siswa yang lebih akurat (Schmidt and Gelhert 2016). Peningkatan akurasi ini sangat penting dalam pendidikan dasar, di mana penilaian sering menginformasikan keputusan penting mengenai penempatan siswa, intervensi, dan strategi pengajaran.

Selain itu, penggunaan IRT dapat memperlancar proses penilaian di pendidikan dasar dengan memungkinkan pengembangan tes yang lebih singkat dan lebih efisien tanpa mengorbankan kualitas pengukuran. Dengan mengidentifikasi dan mempertahankan hanya butir-butir yang paling informatif, para pendidik dapat membuat penilaian yang tidak terlalu membebani siswa sekaligus tetap memberikan ukuran yang akurat atas kemampuan mereka (Zhang 2023). Efisiensi ini khususnya bermanfaat di lingkungan pendidikan dasar, di mana siswa mungkin memiliki rentang perhatian yang terbatas dan tingkat stamina yang berbeda-beda dalam mengikuti tes.

Lebih jauh lagi, kapasitas IRT untuk memberikan umpan balik tingkat butir yang terperinci dapat menginformasikan praktik pengajaran dan pengembangan kurikulum di pendidikan dasar. Dengan menganalisis data kinerja butir, para pendidik dapat mengidentifikasi area konten tertentu yang menjadi kendala siswa dan menyesuaikan strategi pengajaran mereka. Pendekatan berbasis data terhadap pengajaran ini dapat menghasilkan hasil pembelajaran yang lebih baik dan pengalaman pendidikan yang lebih disesuaikan bagi siswa (Michael 2010). Sebaliknya, fokus CTT pada skor total mungkin tidak memberikan tingkat wawasan yang sama terhadap kebutuhan siswa secara individu, yang berpotensi menghambat intervensi yang ditargetkan.

Peran Teori Respons Butir (IRT) dalam meningkatkan penilaian berbasis butir dibandingkan dengan metode Teori Tes Klasik (CTT) tradisional dalam lingkungan

pendidikan dasar sangatlah penting dan beragam. IRT menyediakan pendekatan yang lebih bernuansa dan terperinci untuk mengevaluasi kinerja butir penilaian individual, yang dapat menghasilkan akurasi pengukuran dan hasil pendidikan yang lebih baik.

Salah satu keuntungan utama IRT dibandingkan CTT adalah fokusnya pada karakteristik butir individual, bukan hanya pada skor tes total. CTT mengevaluasi kinerja penilaian secara keseluruhan, dengan mengandalkan statistik agregat seperti koefisien reliabilitas dan rata-rata butir. Pendekatan ini dapat mengaburkan informasi penting tentang bagaimana butir-butir tertentu berfungsi di berbagai tingkat kemampuan siswa. Sebaliknya, IRT memodelkan respons setiap butir sebagai fungsi dari sifat laten yang sedang diukur, yang memungkinkan analisis yang lebih terperinci tentang tingkat kesulitan dan diskriminasi butir. Kemampuan ini khususnya bermanfaat dalam pendidikan dasar, di mana pemahaman tentang tantangan spesifik yang dihadapi siswa dapat menginformasikan strategi pengajaran yang ditargetkan.

Selain itu, kemampuan IRT untuk menyediakan parameter tingkat butir memungkinkan pendidik untuk mengidentifikasi butir mana yang berfungsi dengan baik dan mana yang mungkin perlu direvisi atau dihapus. Penelitian telah menunjukkan bahwa IRT dapat menghasilkan sensitivitas yang lebih besar dalam membedakan antaritem skala dibandingkan dengan CTT, yang sering kali hanya menunjukkan sedikit perbedaan antaritem. Sensitivitas ini sangat penting dalam penilaian pendidikan dasar, yang tujuannya adalah untuk mengukur pemahaman dan keterampilan siswa secara akurat di berbagai bidang konten.

Manfaat penting lain dari IRT adalah penerapannya dalam pengujian adaptif terkomputerisasi (CAT). CAT memungkinkan pelaksanaan penilaian yang disesuaikan dengan tingkat kemampuan siswa secara individual, dengan menyajikan item yang tidak terlalu mudah atau terlalu sulit berdasarkan respons sebelumnya. Pendekatan adaptif ini dapat meningkatkan keterlibatan dan motivasi siswa, karena memberikan pengalaman penilaian yang lebih personal. Sebaliknya, CTT biasanya menggunakan serangkaian item yang tetap, yang mungkin tidak mencerminkan secara memadai berbagai kemampuan siswa dalam lingkungan pendidikan dasar. Fleksibilitas IRT dalam hal ini dapat menghasilkan penilaian yang lebih akurat terhadap kemahiran siswa.

Lebih jauh, IRT memfasilitasi evaluasi fungsi item diferensial (DIF), yang memeriksa apakah item berkinerja secara konsisten di berbagai kelompok demografi. Aspek ini sangat penting untuk memastikan keadilan dan kesetaraan dalam penilaian pendidikan. Dengan mengidentifikasi butir soal yang dapat merugikan kelompok siswa tertentu, pendidik dapat membuat keputusan yang tepat tentang pemilihan butir soal dan desain tes. Di sisi lain, CTT sering gagal mendeteksi bias tersebut karena fokusnya yang agregat, yang berpotensi menyebabkan hasil penilaian yang tidak adil.

Selain keuntungan ini, IRT menyediakan kerangka kerja untuk mengevaluasi skalabilitas butir soal. Hal ini penting dalam pendidikan dasar, di mana penilaian sering kali bertujuan untuk mengukur keterampilan atau pengetahuan yang berkesinambungan. IRT memungkinkan peneliti untuk menilai sejauh mana butir soal dapat dikelompokkan ke dalam konstruksi satu dimensi atau multidimensi, sehingga

menginformasikan pengembangan alat penilaian yang lebih efektif. Studi telah menunjukkan bahwa IRT dapat menghasilkan ukuran yang lebih andal dan valid dibandingkan dengan CTT, khususnya dalam penilaian kompleks yang memerlukan pemahaman yang mendalam tentang kinerja siswa.

Selain itu, integrasi IRT ke dalam penilaian pendidikan dasar dapat meningkatkan sifat psikometrik alat ini. Model IRT menghasilkan estimasi butir soal dan sifat laten yang invarian, yang berarti bahwa hasilnya tidak bergantung pada sampel spesifik yang digunakan untuk pengujian. Invariansi ini penting untuk memastikan bahwa hasil penilaian dapat digeneralisasikan ke berbagai populasi, fitur yang sering kali tidak ada dalam penilaian berbasis CTT. Akibatnya, penerapan IRT dapat menghasilkan evaluasi kinerja siswa yang lebih kuat dan andal.

Selain itu, kapasitas IRT untuk memberikan umpan balik tingkat butir soal yang terperinci dapat menginformasikan praktik pengajaran dan pengembangan kurikulum dalam pendidikan dasar. Dengan menganalisis data kinerja butir soal, pendidik dapat mengidentifikasi area konten tertentu yang menjadi kendala siswa dan menyesuaikan strategi pengajaran mereka. Pendekatan berbasis data terhadap instruksi ini dapat menghasilkan hasil pembelajaran yang lebih baik dan pengalaman pendidikan yang lebih sesuai bagi siswa. Sebaliknya, fokus CTT pada skor total mungkin tidak memberikan tingkat wawasan yang sama terhadap kebutuhan siswa secara individual, sehingga berpotensi menghambat intervensi yang ditargetkan.

Penggunaan IRT juga memungkinkan pengembangan penilaian yang lebih singkat dan lebih efisien tanpa mengorbankan kualitas pengukuran. Dengan mengidentifikasi dan mempertahankan butir soal yang paling informatif saja, pendidik dapat membuat penilaian yang tidak terlalu membebani siswa sekaligus tetap memberikan ukuran akurat atas kemampuan mereka. Efisiensi ini khususnya bermanfaat dalam lingkungan pendidikan dasar, di mana siswa mungkin memiliki rentang perhatian yang terbatas dan tingkat stamina yang berbeda-beda dalam mengikuti ujian.

Kesimpulan

Pilihan antara CTT dan IRT harus didasarkan pada tujuan khusus penilaian dan sifat konstruk yang diukur. Untuk instrumen yang memerlukan analisis butir yang bernuansa dan kemampuan beradaptasi, IRT adalah pendekatan yang lebih disukai, sementara CTT mungkin cukup untuk penilaian yang lebih sederhana di mana skor total menjadi perhatian utama. Pada akhirnya, integrasi kedua metodologi dapat menghasilkan pemahaman yang paling komprehensif tentang kinerja butir dan meningkatkan pengembangan alat penilaian yang efektif.

Penerapan CTT dan IRT dalam penilaian pendidikan dasar memberikan kontribusi yang signifikan terhadap keakuratan dan kualitas hasil evaluasi. Sementara CTT menawarkan pendekatan langsung untuk mengukur kinerja siswa secara keseluruhan, keterbatasannya dalam analisis butir soal dan kepekaan terhadap perbedaan individu dapat menghambat efektivitasnya dalam menginformasikan praktik pendidikan. Sebaliknya, IRT menyediakan kerangka kerja yang kuat untuk

memahami kinerja tingkat butir soal, menangani masalah keadilan dan kesetaraan, dan meningkatkan kualitas penilaian secara keseluruhan. Seiring dengan terus berkembangnya penilaian pendidikan, integrasi IRT ke dalam pendidikan dasar kemungkinan akan memainkan peran penting dalam meningkatkan keakuratan pengukuran dan mendukung pembelajaran siswa.

Peran IRT dalam meningkatkan penilaian berbasis butir soal dalam lingkungan pendidikan dasar sangatlah penting. Dengan memberikan analisis yang lebih rinci dan bernuansa terhadap kinerja butir soal individual, IRT meningkatkan akurasi pengukuran, menginformasikan praktik pengajaran, dan mendorong keadilan dalam penilaian. Keunggulan IRT, termasuk kemampuan adaptasinya, kepekaan terhadap fungsi butir soal, dan kemampuan memberikan umpan balik terperinci, memposisikannya sebagai metode yang unggul untuk mengembangkan dan menyempurnakan instrumen penilaian dalam pendidikan dasar. Seiring dengan terus berkembangnya penilaian pendidikan, integrasi IRT kemungkinan akan memainkan peran penting dalam meningkatkan kualitas pengukuran dan mendukung pembelajaran siswa.

Referensi

- Adler, Mats, Jerker Hetta, Göran Isacson, and Ulf Brodin. 2012. "An Item Response Theory Evaluation of Three Depression Assessment Instruments in a Clinical Sample." *BMC Medical Research Methodology*. doi: 10.1186/1471-2288-12-84.
- Cappelleri, Joseph C., J. Jason Lundy, and Ron D. Hays. 2014. "Overview of Classical Test Theory and Item Response Theory for the Quantitative Assessment of Items in Developing Patient-Reported Outcomes Measures." *Clinical Therapeutics*. doi: 10.1016/j.clinthera.2014.04.006.
- Carlton, Jill, Donna Rowen, and Jackie Elliott. 2020. "Assessment of the Psychometric Properties and Refinement of the Health and Self-Management in Diabetes Questionnaire (HASMID)." *Health and Quality of Life Outcomes*. doi: 10.1186/s12955-020-01305-3.
- Cella, David, William T. Riley, Arthur A. Stone, Nan Rothrock, Bryce B. Reeve, Susan Yount, Dagmar Amtmann, Rita Bode, Daniel J. Buysse, Seung W. Choi, Karon F. Cook, Robert F. DeVellis, Darren A. DeWalt, James F. Fries, Richard Gershon, Elizabeth A. Hahn, Jin Shei Lai, Paul A. Pilkonis, Dennis A. Revicki, Matthias Rose, Kevin P. Weinfurt, and Ron D. Hays. 2010. "The Patient-Reported Outcomes Measurement Information System (PROMIS) Developed and Tested Its First Wave of Adult Self-Reported Health Outcome Item Banks: 2005–2008." *Journal of Clinical Epidemiology*. doi: 10.1016/j.jclinepi.2010.04.011.
- Chen, Hao, and Yiduo Ye. 2021. "Validation of the Weight Bias Internalization Scale for Mainland Chinese Children and Adolescents." *Frontiers in Psychology*. doi: 10.3389/fpsyg.2020.594949.
- Hamann, Antonieta. 2023. "Validation of a Scale for the Evaluation of the Dimensions

- of Short Break Tourist Destination." *Journal of Business*. doi: 10.21678/jb.2023.1887.
- Hambleton, R. K., and H. Swaminathan. 2013. *Item Response Theory: Principles and Applications*. books.google.com.
- Hays, Ron D., Leo S. Morales, and Steve P. Reise. 2000. "Item Response Theory and Health Outcomes Measurement in the 21st Century." *Medical Care*. doi: 10.1097/00005650-200009002-00007.
- Hilmiyati, F. 2024. "Integration of Cognitive Technology in Learning Assessment and Evaluation." *Al-Hijr: Journal of Adulearn World*.
- Jafari, Peyman, Zahra Bagheri, Seyyed Mohammad Taghi Ayatollahi, and Zahra Soltani. 2012. "Using Rasch Rating Scale Model to Reassess the Psychometric Properties of the Persian Version of the PedsQLTM 4.0 Generic Core Scales in School Children." *Health and Quality of Life Outcomes*. doi: 10.1186/1477-7525-10-27.
- Lamprianou, I., and J. A. Athanasou. 2009. *A Teacher's Guide to Educational Assessment: Revised Edition*. books.google.com.
- Maciel Mattos, Grazielle Christine, Juliana Vaz de Mambrini, Jennifer E. Gallagher, Saul Martins Paiva, and Mauro Henrique Nogueira Abreu. 2017. "Evaluating Psychometric Properties of an Instrument Addressing Comprehensiveness of Care Among Dentists." *Brazilian Dental Journal*. doi: 10.1590/0103-6440201701334.
- Michael, Thomas. 2010. "The Value of Item Response Theory in Clinical Assessment: A Review." *Assessment*. doi: 10.1177/1073191110374797.
- Petrillo, Jennifer, Stefan Cano, Lori McLeod, and Cheryl D. Coon. 2015. "Using Classical Test Theory, Item Response Theory, and Rasch Measurement Theory to Evaluate Patient-Reported Outcome Measures: A Comparison of Worked Examples." *Value in Health*. doi: 10.1016/j.jval.2014.10.005.
- Samuel, Douglas B., Leonard J. Simms, Lee Anna Clark, W. John Livesley, and Thomas A. Widiger. 2010. "An Item Response Theory Integration of Normal and Abnormal Personality Scales." *Personality Disorders Theory Research and Treatment*. doi: 10.1037/a0018136.
- Schmidt, Christopher D., and Nathan C. Gelhert. 2016. "Couples Therapy and Empathy." *The Family Journal*. doi: 10.1177/1066480716678621.
- Silva, Rajitha, Yuping Guan, and Tim B. Swartz. 2017. "Bayesian Diagnostics for Test Design and Analysis." *Journal on Efficiency and Responsibility in Education and Science*. doi: 10.7160/eriesj.2017.100202.
- Sukardi, H. M. 2008. "Evaluasi Pendidikan Prinsip Dan Operasionalnya." *Jakarta: Bumi Aksara*.
- Twiss, James, David Meads, E. Preston, S. R. Crawford, and Stephen P. McKenna. 2012. "Can We Rely on the Dermatology Life Quality Index as a Measure of the

Impact of Psoriasis or Atopic Dermatitis?" *Journal of Investigative Dermatology*. doi: 10.1038/jid.2011.238.

Zhang, Chao. 2023. "Validation of a Chinese Version of the Digital Stress Scale and Development of a Short Form Based on Item Response Theory Among Chinese College Students." *Psychology Research and Behavior Management*. doi: 10.2147/prbm.s413162.