KUALITAS SOAL PENILAIAN AKHIR SEMESTER MAPEL BIOLOGI DENGAN PENDEKATAN ITEM RESPONSE THEORY DI SMA TRENSAINS MUHAMMADIYAH SRAGEN TA 2022/2023

Ifani Saskia Putri; Lina Agustina, Pendidikan Biologi, Fakultas Keguruan dan Ilmu Pendidikan, Universitas Muhammadiyah Surakarta

Abstrak

Penilaian akhir semester merupakan alat untuk mengukur kemampuan siswa setelah belajar. *Item Response Theory* (IRT) dapat mendeskripsikan interaksi antara responden dengan butir pernyataan. Penelitian ini bertujuan untuk mengetahui kualitas soal penilaian akhir semester (PAS) mata pelajaran biologi dengan pendekatan *Item Response Theory* (IRT) di SMA Trensains Muhammadiyah Sragen TA 2022/2023. Penelitian ini menggunakan jenis penelitian deskriptif kuantitatif dengan teknik pengumpulan data yang dilakukan yaitu dokumentasi dan wawancara. Penelitian ini dilakukan di SMA Trensains Muhammadiyah Sragen. Hasil penelitian ini menunjukkan bahwa keseluruhan soal valid, memiliki reliabilitas person bagus yaitu sebesar 0,84 serta reliabilitas item yang bagus sekali sebesar 0,92. Selain itu, pada tingkat kesukaran soal memiliki variasi yaitu kategori soal yang sangat sulit terdapat sebanyak 9 soal, soal sulit sebanyak 16 soal, soal mudah sebanyak 28 soal, dan soal dengan kategori sangat mudah sebanyak 7 soal. Keseluruhan soal memiliki daya pembeda yang bagus karena memiliki nilai indeks dibawah 0,5 dan potensi bias (DIF) berdasarkan kelompok jenis kelamin atau gender terdapat 8 soal yang bias..
Kata Kunci: *Evaluasi pembelajaran, Penilaian Akhir Semester (PAS), Item Response Theory (IRT)*

ABSTRACT

The end-of-semester assessment is a tool to measure students' abilities after studying. Item Response Theory (IRT) can describe interactions between respondents and statement items. This study aims to determine the quality of end-of-semester assessment questions (PAS) for biology subjects using the item response theory (IRT) approach at SMA Trensains Muhammadiyah Sragen TA 2022/2023. This research uses a type of quantitative descriptive research with data collection techniques, namely documentation and interviews. This research was conducted at SMA Trensains Muhammadiyah Sragen. The results of this study indicate that all questions are valid, have a good person reliability of 0.84, and excellent item reliability of 0.92. In addition, the difficulty level of the questions has variations; namely, in the category of very difficult questions, there are 9 questions, 16 difficult questions, 28 easy questions, and 7 questions in the very easy category. All questions have good discriminating power because they have an index value below 0.5 and the potential for bias (DIF) based on gender or gender groups. There are 8 items that are biased.

Keywords: *Final semester assessment, Item response theory, Learning evaluation*

1. PENDAHULUAN

Pendidikan merupakan komponen penting dalam sumber daya manusia yang ahli dalam bidangnya. Oleh karena itu, lembaga pendidikan sangat dituntut untuk meningkatkan kualitas pembelajaran agar menciptakan lulusan yang kompeten. Menurut peraturan menteri pendidikan nasional Republik Indonesia Nomor 16 Tahun 2007 tentang standar kualifikasi akademik dan kompetensi guru menyatakan bahwa kompetensi pendagogi yang harus dimiliki khususnya adalah kemampuan dalam menyelenggarakan penilaian proses dan hasil belajar yang terdiri dari: (a) memahami prinsip-prinsip penilaian hasil belajar sesuai dengan karakteristik mata pelajaran yang diampu, (b) menentukan aspek-aspek penilaian hasil belajar yang penting untuk dinilai, (c) menentukan prosedur penilaian hasil belajar, (d) mengembangkan instrumen penilaian hasil belajar, (e) mengadministrasikan penilaian proses dan hasil belajar secara berkesinambungan dengan menggunakan berbagai instrumen, (f) melakukan evaluasi proses dan hasil belajar.

Evaluasi merupakan kegiatan untuk mengumpulkan data-data untuk mengukur sejauh mana tujuan pembelajaran telah tercapai. Proses evaluasi itu harus diarahkan ke tujuan tertentu, untuk mendapatkan berbagai jawaban tentang bagaimana memperbaiki pembelajaran serta evaluasi mengharuskan penggunaan berbagai alat ukur yang akurat dan bermakna untuk mengumpulkan informasi yang dibutuhkan dalam membuat keputusan (Febriana 2012). Oleh karena itu, pendidik harus mampu mengevaluasi proses pembelajaran guna untuk menentukan kemajuan dari suatu pembelajaran.

Tantangan dalam proses evaluasi adalah kesulitan untuk menghindari kesalahan dalam pengukuran. Alat ukur atau intrumen dapat berupa ter tertulis. Pendidik memberikan nilai berdasarkan jumlah jawaban benar atua kualitas jawaban yang diberikan oleh siswa. Beberapa factor dapat menyebabkan kesalahan dalam penilaian hasil belajar. Salah satunya adalah alat ukur/instrument. Alat ukur yang digunakan tidak dapat mengukur apa yang seharusnya diukur (Marera, 2020).

Alat ukur yang baik seharusnya dapat mencerminkan kemampuan dan kompetensi siswa. Dalam proses evaluasi hasil belajar siswa sehingga dapat mengetahui kemampuan dan kompetensi siswa maka perlu dilakukan penilaian atau ujian (Amelia and Kriswantoro 2017). Penilaian merupakan komponen yang selalu melekat pada proses belajar mengajar dalam pendidikan. Hasil penilaian bisa digunakan sebagai acuan seorang guru untuk mengetahui keberhasilan dan meningkatkan

kualitas pengajaran (Mardapi 2016). Menurut Santoso (2019), penilaian adalah suatu kegiatan menafsirkan hasil pengukuran. Penilaian akhir semester merupakan alat untuk mengukur kemampuan siswa setelah belajar dan menjadi tolak ukur seorang guru sebagai fasilitator dalam merencanakan proses pembelajaran yang sesuai dengan kemampuan siswa sehingga pembelajaran lebih efektif dan dapat mencapai tujuan pendidikan. Tes untuk mengukur penguasaan siswa terhadap metode saintifik dapat berupa tes tertulis dan tes unjuk kerja. Untuk dapat mengukur keterampilan siswa maka perlu adanya pengembangan instrumen tes yang ideal dan sesuai dengan standar (Subali, Kumaidi, and Aminah 2020).

Menurut (Subali, Kumaidi, and Aminah 2020) tes untuk mengukur penguasaan siswa terhadap metode saintifik dapat berupa tes tertulis dan tes unjuk kerja. Untuk dapat mengukur keterampilan siswa maka perlu adanya pengembangan instrument tes yang ideal dan sesuai dengan standar. Kualitas butir soal dapat dianalisis dengan menggunakan *Classical Test Theory* (CTT) dan *Item Response Theory* (IRT). *Item Responsse Theory* (IRT) dapat menggambarkan perbandingan tingkat kesulitan soal dan tingkat kemampuan siswa dalam satu plot garis yang menggunakan skala logit. Dengan demikian penilaian akan lebih bermakna dan mudah untuk mengetahui kemampuan siswa dan kualitas item soal secara bersamaan (Subali, Kumaidi, and Aminah 2020). Hal ini, memperlihatkan ciri dari *Item Response Theory* (IRT) sebagai penilaian yang lebih unggul dalam menskor kemampuan personal, di mana hubungan peserta tes dengan nilai dari estimasi kemampuan yang dimiliki terukur dengan baik. Menurut penelitian dari Salsabila et al., (2023) menyatakan bahwa analisis instrumen manajemen diri mempergunakan pendekatan *item response theory* (IRT) atau rasch model dapat memastikan data yang diperoleh tepat, objektif, dan konsisten dikarenakan pengukuran menggunakan pendekatan tersebut mampu mendeskripsikan interaksi antara responden dengan butir pernyataan.

Aspek yang akan diukur adalah validitas butir soal atau sering disebut dengan validitas item yaitu suatu item dapat dikatakan valid apabila memiliki dukungan yang besar terhadap skor total (Arikunto 2021). Validitas digunakan untuk mengetahui tingkat keakuratan tes dalam menjalankan fungsinya sebagai alat ukur, hal ini berkaitan dengan suatu instrumen tes apakah sudah dapat mengukur apa yang seharusnya diukur (Himelfarb 2019). Kemudian reliabilitas juga digunakan untuk mengetahui tentang konsistensi suatu tes dalam menilai apa yang seharusnya dinilai. Suatu instrumen dapat dikatakan konsisten jika menghasilkan respon siswa yang relatif sama apabila diujikan secara berulang kali pada kondisi yang sama dengan peserta dan waktu yang berbeda

(Huda and Wahyuni 2020). Tingkat kesukaran adalah mengukur kemampuan peserta didik dalam menjawab soal dengan benar dan tepat. Jika semakin sulit suatu item maka semakin tinggi kemampuan siswa (Eaton et al., 2019). Sedangkan daya pembeda pada soal digunakan untuk mengetahui kemampuan peserta didik yang telah menguasai kompetensi atau belum menguasai kompetensi. Daya pembeda mencerminkan seberapa baik item dapat membedakan antara berbagai tingkat kemampuan siswa yang memiliki kemampuan tinggi dan kemampuan rendah (Nima et al. 2020). Pada potensi bias atau Analisis DIF mencakup serangkaian proses untuk membandingkan kinerja kelompok pada item sambil mempertimbangkan potensi peserta didik dalam menjawab soal dengan baik. Sehingga analisis dapat mengidentifikasi kesenjangan pencapaian yang tidak terungkap saat membandingkan nilai total (Martinková et al. 2017).

Berdasarkan aspek tersebut maka fokus pada penelitian ini adalah untuk mengetahui kualitas soal penilaian akhir semester (PAS) mata pelajaran biologi dengan pendekatan *item response theory* (IRT) di SMA Trensains Muhammadiyah Sragen TA 2022/2023 yang akan dianalisis dengan bantuan software winstep.

2. Metode

Penelitian ini dilaksanakan di SMA Trensains Muhammadiyah Sragen pada semester genap tahun ajaran 2022/2023. Populasi penelitian ini adalah seluruh siswa kelas XI SMA Trensains Muhammadiyah Sragen yang berjumlah 84 siswa terdiri dari 4 kelas yaitu MIPA 1 sampai MIPA 4. Pengambilan subjek penelitian ini menggunakan penelitian populasi yang menggunakan teknik pengambilan sampel yaitu Purposive Sampling. Sampel yang digunakan yaitu seluruh siswa kelas XI SMA Trensains Muhammadiyah Sragen yang berjumlah 84 siswa, data dalam penelitian ini meliputi soal penilaian akhir semester, kunci jawaban penilaian akhir semester, dan hasil jawaban siswa yang dianalisis secara deskriptif kuantitatif menggunakan bantuan software winstep.

3. HASIL DAN PEMBAHASAN

3.1 Validitas

Hasil analisis instrument tes menggunakan *Item response theory* (IRT) didapati semua soal dinyatakan valid. Penentuan validitas soal *Item response theory* (IRT) ditentukan oleh 3 aspek yaitu berdasarkan nilai *Outfit MNSQ, Outfit ZSTD* dan *PT Measure Correlation*. Menurut (Cordier et al. 2019; Nuryanti, Masykuri, and Susilowati 2018) soal dikatakan valid berdasarkan IRT apabila memiliki nilai sebagai berikut :

- Nilai Outfit MNSQ (Mean Square) yang diterima adalah: 0,5 < Outfit – MNSQ < 1,5
- Nilai Outfit ZSTD (Z – Standard) yang diterima adalah: –2,0 < ZSTD < +2,0
- Nilai Pt Measure Corr (Point Measure Correlation): 0,4 < Point Measure Corr < 0,85
    - Soal dinyatakan valid jika memenuhi minimal 2 aspek dalam analisis *Item response theory* (IRT).
    - **Tabel 1** Hasil analisis berdasarkan validitas

Kategori	**Nomor Soal**
Valid	1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50, 51, 52, 53, 54, 55, 56, 57, 58, 59, 60
Tidak Valid	-

Pada uji validitas, hal yang dapat dilihat adalah berdasarkan tingkat kesesuaian (Item fit). Tingkat kesesuian butir dilihat dari 3 aspek yaitu nilai MNSQ 0,5-1,5. Nilai MNSQ idealnya adalah 1,0 yang artinya soal tersebut dipahami seluruhnya oleh responden, jika nilai MNSQ semakin menjauhi 1,0 maka artinya soal tersebut semakin tidak dipahami oleh responden, nilai ZSTD memiliki rentang indeks nilai dari -2 sampai 2, dan nilai *Point* Pt Mean Corr memiliki rentang indeks nilai dari 0,4 sampai 0,85. Jika soal telah memenuhi setidaknya dua indikator di atas maka soal tersebut dapat disimpulkan valid atau fit.

Item dapat dinyatakan valid atau fit setidaknya dapat memenuhi dua aspek diatas. Uji validitas yang diujikan terhadap 60 item pertanyaan semuanya menunjukkan valid atau fit. 60 item tersebut setidaknya telah memenuhi dua aspek seperti pada item nomor 38 telah memenuhi tiga aspek yaitu nilai MNSQ (Mean Square) 1,28 dan nilai ZSTD (Z-Standard) 0,8 serta nilai Point Measure Corelation yaitu 0,6 sehingga dapat disimpulkan bahwa soal tersebut valid atau fit, begitu pula dengan 59 item yang lain. Salsabila et al., (2023) menyatakan bahwa seluruh butir item peserta didik dikatakan valid atau fit yang berarti berfungsi normal dan bisa dimengerti dengan benar oleh peserta didik, sehingga dapat diartikan bisa mengukur manajemen diri dari peserta didik.

3.2 Reliabilitas

Nilai reliabilitas yang rendah dapat dipengaruhi oleh jumlah siswa sebagai responden kurang atau jumlah soalnya yang terlalu sedikit (Cordier et al. 2018). Nilai ketentuan kategori soal reliabilitas pada item dan respon :

Tabel 2 Ketentuan kategori Soal Reliabilitas pada Item dan Person

Reliabilitas Item Dan Person	
Indeks	**Kategori**
Nilai <0,67	Lemah
Nilai 0,67-0,8	Cukup
Nilai 0,81-0,9	Bagus
Nilai 0,91-0,94	Bagus Sekali
Nilai >0,94	Istimewa

Hasil analisis instrument tes menggunakan *Item response theory* (IRT) menerangkan tentang reliabilitas item dan person.

Tabel 3 Hasil Analisis Reliabilitas Item dan Person

Reliabilitas Item dan Person		
	Nilai	**Kategori**
Item	0,92	Bagus sekali
Person	0,84	Bagus

Reliabilitas item menunjukkan konsistensi soal yang digunakan untuk mengukur kemampuan siswa, didapati nilai 0,92 yang termasuk kategori bagus sekali. Sedangkan reliabilitas person menunjukkan konsistensi jawaban siswa dalam menjawab soal, didapati nilai 0,84 termasuk pada kategori bagus.

3.3 Tingkat Kesukaran

Distribusi kemampuan siswa dan tingkat kesukaran item dapat juga dilihat pada peta dimensi person-item menggunakan skala logit. Semakin ke atas semakin sulit soal tersebut, sedangkan semakin ke bawah maka dapat dikatakan soalnya semakin mudah (Cordier et al. 2018).

Tabel 4 Ketentuan Kategori Tingkat Kesukaran Secara IRT

Indeks	**Kategori**	**Standar Deviasi = 0,96**
Nilai <-SD	Sangat Mudah	Nilai < -0,96
Nilai -SD-0	Mudah	Nilai -0,96-0
Nilai 0-SD	Sulit	Nilai 0-0,96
Nilai >SD	Sangat Sulit	Nilai >0,96

Berdasarkan hasil analisis instrument tes menggunakan *Item response theory* (IRT) didapati hasil urutan tingkat kesukaran soal yang sangat sulit hingga soal yang mudah. Dengan ketentuan kategori dari *Item response theory* (IRT) yang terbagi menjadi 4 kategori yaitu soal sangat sulit, sulit, mudah, dan sangat mudah.

Tabel 5 Hasil Analisis Berdasarkan Tingkat Kesukaran

Tingkat Kesukaran		
Kategori	**Jumlah**	**Nomor soal**
Sangat Mudah	7	16, 21, 22, 30, 31, 42, 50
Mudah	28	1, 2, 4, 5, 6, 7, 11, 12, 13, 15, 18, 19, 20, 23, 25, 26, 27, 34, 35, 37, 39, 44, 47, 49, 52, 54, 55, 57
Sulit	16	3, 8, 9, 17, 24, 29, 32, 33, 36, 40, 41, 43, 45, 48, 53, 60

Sangat Sulit	9	10, 14, 28, 38, 46, 51, 56, 58, 59

Soal sangat sulit terdapat 9 soal, soal yang sulit terdapat 16 soal, soal yang tergolong mudah terdapat 28 soal, dan 7 soal dengan kategori sangat mudah. Nilai measure yang menunjukkan tingkat kesukaran soal dapat digunakan untuk melihat apakah soal yang diberikan tersebut dapat mewakili variasi peserta didik atau tidak. Tak hanya itu, pada nilai measure yang berupa *logit* juga dapat digunakan untuk membedakan kemampuan dari peserta didik atau siswa dengan lebih teliti.

Hal tersebut sejalan dengan penelitian dari Salsabila et al., (2023) menyatakan bahwa butir item dalam sebuah instrumen manajemen diri menunjukkan tingkat kesukaran yang beragam dengan kategori sukar sekali, sukar, mudah, dan mudah sekali. Sehingga dapat dikatakan bahwa seluruh item dalam instrumen tersebut berfungsi secara normal dalam mengukur manajemen diri.

3.4 Daya Pembeda

Daya pembeda pada soal digunakan untuk mengetahui kemampuan siswa yang telah menguasai kompetensi atau belum menguasai kompetensi. Daya pembeda mencerminkan seberapa baik item dapat membedakan antara berbagai tingkat kemampuan siswa, yang memiliki kemampuan tinggi dan kemampuan rendah (Eaton et al. 2019b; Nima et al. 2020). Soal yang mempunyai daya pembeda dapat membedakan siswa yang memiliki kemampuan yang tinggi dengan siswa yang memiliki kemampuan yang rendah (Fatimah, Laela Umi: Alfath 2019). Nilai standar error (SE) mengindikasikan daya beda:

Tabel 6 Ketentuan Kategori Daya Pembeda Secara IRT

Indeks Diskriminasi Item	Kategori
Nilai Model SE < 0,5	Bagus
Nilai Model SE 0,5-1	Cukup
Nilai Model SE > 1	Jelek

Berdasarkan anlisis menggunakan pendekatan *Item response theory* (IRT) yang dibantu dengan software Winstep, probabilitas jawaban benar dalam tes tidak langsung dikaitkan dengan kemampuan tes, namun dihubungkan melalui formula matematika yang dikenal dengan istilah logit.

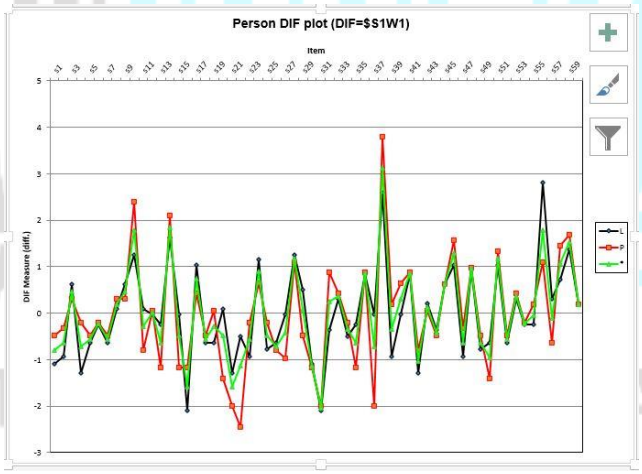
Nilai daya pembeda soal dapat dilihat melalui model nilai S.E (standar eror). Nilai tersebut menunjukkan seberapa presisi atau seberapa tepat suatu soal tersebut dalam membedakan tingkat kemampuan dan pemahaman peserta didik terhadap suatu materi yang diujikan. Prinsip dasarnya masih sama yaitu probabilitas, yang mana item yang sulit akan mengakibatkan peluang peserta didik menjawab item tersebut dengan benar akan semakin kecil dan peserta didik yang memiliki

tingkat abilitas tinggi memiliki peluang menjawab soal dengan benar lebih tinggi jika dibandingkan dengan peserta didik yang mmeiliki tingkat abilitas yang rendah. Nilai model S.E (standart eror) dari penelitian ini seluruhnya menghasilkan nilai kurang dari 0,5 dimana daya pembeda semua item tergolong bagus. Sehingga dapat disimpulkan bahwa kemampuan soal untuk membedakan kemampuan peserta didik bagus, peserta didik yang memiliki tingkat abilitas tinggi mempunyai peluang menjawab soal dengan benar semakin tinggi.

3.5 Potensi Bias (DIF)

Analisis *Differential Item Functioning* (DIF) digunakan untuk menguji apakah item yang digunakan dapat berfungsi sama pada semua kelompok. DIF terjadi ketika karakteristik dari responden bukan hanya kemampuan saja yang mempengaruhi hasil kerja tetapi terdapat sifat yang lainnya (Cordier et al. 2018). Dalam menentukan DIF dapat dilakukan dengan analisisi pada nilai probalilitas, jika kurang dari 0,05 maka item terbilang memiliki potensi bias.

Hasil analisis DIF yang dilihat dari nilai probabilitas terdapat 8 item tergolong soal yang bias. Untuk mengetahui pihak mana yang diuntungkan dapat dilihat pada grafik berikut :



Gambar 1 Grafik Pengukuran DIF

Menunjukkan tingkat kesulitan item relative bagi masing-masing person. Jadi semakin tinggi titik grafik, semakin sulit item tersebut bagi person. Terdapat tiga buah kurva berdasarkan jenis kelamin, yaitu L (laki-laki), P (perempuan), dan tanda * (bintang) menunjukkan dari nilai rata-rata. Dari grafik tersebut terlihat bahwa jarak nilai DIF per soal pada jenis kelamin kelompok laki-laki dan perempuan. Penilaian dapat ditentukan dengan semakin jauh jaraknya maka semakin besar potensi biasnya. Berdasarkan grafik dapat diketahui pada nomor 20, 22, 37, 56 lebih menguntungkan kelompok laki-laki, sedangkan pada nomor 10, 15, 32, 39 lebih menguntungkan kelompok perempuan. Instrument tidak boleh mendikriminasi hasil kerja siswa berdasarkan

kelompok, jenis kelamin, etnis, bahasa dan sebagainya, sehingga harus disusun dengan menjaga keadilan untuk menilai kemampuan yang sama (Lim, Choe, and Han 2022).

4. PENUTUP

Berdasarkan hasil penelitian dan pembahasan dapat disimpulkan bahwa kualitas soal penilaian akhir semester (PAS) mata pelajaran biologi dengan pendekatan *item response theory* (IRT) di SMA Trensains Muhammadiyah Sragen TA 2022/2023 menghasilkan keseluruhan soal valid, memiliki reliabilitas person bagus yaitu sebesar 0,84 serta reliabilitas item yang bagus sekali sebesar 0,92. Selain itu pada tingkat kesukaran soal memiliki variasi yaitu kategori soal yang sangat sulit terdapat sebanyak 9 soal, soal sulit sebanyak 16 soal, soal mudah sebanyak 28 soal, dan soal dengan kategori sangat mudah sebanyak 7 soal. Keseluruhan soal memiliki daya pembeda yang bagus karena memiliki nilai indeks dibawah 0,5 dan potensi bias (DIF) berdasarkan kelompok jenis kelamin atau gender terdapat 8 soal yang bias.

.

DAFTAR PUSTAKA

Amelia, Rizki Nor, and Kriswantoro Kriswantoro. 2017. "Implementation of Item Response Theory for Analysis of Test Items Quality and Students' Ability in Chemistry." *JKPK (Jurnal Kimia dan Pendidikan Kimia)* 2(1): 1.

Arikunto, S. 2021. "Dasar-Dasar Evaluasi Pendidikan Edisi 3 (R. Damayanti (Ed.)." : 1–400. https://books.google.com/books?hl=en&lr=&id=j5EmEAAAQBAJ&oi=fnd&pg=PA1&dq=pendidikan&ots=6uAPIgqLXM&sig=P6Zd6yrUVBrKlYSecTW8LvL-eJE (March 30, 2023).

Cordier, Reinie et al. 2018. "Using Rasch Analysis to Evaluate the Reliability and Validity of the Swallowing Quality of Life Questionnaire: An Item Response Theory Approach." *Dysphagia* 33(4): 441–56. https://link.springer.com/article/10.1007/s00455-017-9873-4 (April 2, 2023).

———. 2019. "Applying Item Response Theory (IRT) Modeling to an Observational Measure of Childhood Pragmatics: The Pragmatics Observational Measure-2." *Frontiers in Psychology* 10(FEB): 408.

Eaton, Philip, Keith Johnson, Barrett Frank, and Shannon Willoughby. 2019a. "Classical Test Theory and Item Response Theory Comparison of the Brief Electricity and Magnetism Assessment and the Conceptual Survey of Electricity and Magnetism." *Physical Review Physics Education Research* 15.

———. 2019b. "Classical Test Theory and Item Response Theory Comparison of the Brief Electricity and Magnetism Assessment and the Conceptual Survey of Electricity and

Magnetism." *Physical Review Physics Education Research* 15(1): 10102. https://doi.org/10.1103/PhysRevPhysEducRes.15.010102.

Fatimah, Laela Umi: Alfath, Khairuddin. 2019. "Analisis Kesukaran Soal Dan Fungsi Distraktor." *Jurnal Komunikasi dan Pendidikan Islam* 8(2): 37–64. https://journal.staimsyk.ac.id/index.php/almanar/article/view/115/104 (April 2, 2023).

Febriana, Rina. 2012. *Evaluasi Pembelajaran*. ed. Bunga Sari Fatmawati. Bumi Aksara Jl. Sawo Raya No. 18 Jakarta 13220.

Himelfarb, Igor. 2019. "REVIEW OF THE LITERATURE A Primer on Standardized Testing: History, Measurement, Classical Test Theory, Item Response Theory, and Equating." *J Chiropr Educ* 33(2): 151–63. www.journalchiroed.com (April 1, 2023).

Huda, Nuril, and Tutik Sri Wahyuni. 2020. "Penggunaan Aplikasi Item and Test Analysis (Iteman) Pada Soal Try Out UN IPA SMP Tahun 2019." *Jurnal Pembelajaran Sains* 4(1): 2527–9157. http://journal2.um.ac.id/index.php/jpsi/article/view/9738.

Lim, Hwanggyu, Edison M. Choe, and Kyung T. Han. 2022. "A Residual-Based Differential Item Functioning Detection Framework in Item Response Theory." *Journal of Educational Measurement* 59(1): 80–104.

Mardapi, D. 2016. *Pengukuran Penilaian Dan Evaluasi Pendidikan*. 2nd ed. Yogyakarta: Nuha Litera.

Marera, Alone. 2020. "A Study on The Quality of Final Exam Items Made by The Teacher at XII Grade Students of Senior High School in Sidrap Regency." *Journal of Biological Science & Education* 2(1): 32–41.

Martinková, Patrícia et al. 2017. "Checking Equity: Why Differential Item Functioning Analysis Should Be a Routine Part of Developing Conceptual Assessments." *CBE Life Sciences Education* 16(2).

Nima, Ali Al et al. 2020. "Validation of Subjective Well-Being Measures Using Item Response Theory." *Frontiers in Psychology* 10: 3036.

Nuryanti, Sri, Muhammad Masykuri, and E Susilowati. 2018. "Analisis Iteman Dan Model Rasch Pada Pengembangan Instrumen Kemampuan Berpikir Kritis Peserta Didik Sekolah Menengah Kejuruan." *Jurnal Inovasi Pendidikan IPA* 4(2): 224–33.

Salsabila, Fadiya, Juntika Nurihsan, and Yaya Sunarya. 2023. "Pengujian Validitas Dan Reliabilitas Instrumen Manajemen Diri Remaja: Rasch Model Analysis." *Jurnal Bimbingan dan Konseling Terapan* 7(1): 15–25. https://ojs.unpatti.ac.id/index.php/bkt/article/view/1741 (April 5, 2023).

Subali, Bambang, Kumaidi, and Nonoh Siti Aminah. 2020. "The Comparison of Item Test Characteristics Viewed from Classic and Modern Test Theory." *International Journal of Instruction* 14(1): 647–60.