

Karakteristik Tes Ilmu Pengetahuan Alam

Prof. Dr. Suwanto M.Pd¹ dan Moh. Zain Bin Musa, Ph.D²

¹Universitas Veteran Bangun Nusantara Sukoharjo, e-mail: suwartowarto@yahoo.com

²Sekolah Bersepadu Musa Asiah, Cambodia

Received: Maret 09, 2022

Accepted: Maret 15, 2022

Online Published: Maret 26, 2022

Abstrak: Tujuan penelitian, memaparkan: karakteristik tes IPA berdasarkan teori tes klasik dan teori tes modern. Desain penelitian adalah kuantitatif dan deskriptif. Objek penelitian adalah tes prestasi IPA selama Covid-19, guru IPA, kepala sekolah, dan wakil kepala sekolah. Data diperoleh dari tanggapan siswa sejumlah 280 terhadap seluruh lembar jawaban siswa kelas VIII SMP MTA Gemolong Sragen sebagai populasi penelitian ini. Kunci jawaban soal IPA dan satu paket soal IPA (50 butir pilihan ganda) diperoleh dari guru IPA. Teknik penelitian dengan wawancara dan dokumenter. Analisis data dilakukan dengan menggunakan program Quest. Hasil penelitian: (1). Karakteristik tes IPA berdasarkan teori tes klasik: Validitas isi tidak terpenuhi, reliabilitas tes 0.960, kategori tingkat kesulitan butir dalam persentase adalah mudah: sedang: sulit = 10%:84%:6%, kategori daya beda butir dalam persentase adalah jelek: cukup: baik: sangat baik = 2%: 4%: 14%: 80%, jadi dominan sangat baik, sedangkan keberfungsian pengecoh dalam persentase adalah tidak efektif: efektif = 0,70%: 99,30%, jadi dominan pengecoh yang efektif. (2). Karakteristik tes IPA berdasarkan teori tes modern: kategori *Threshold* tes IPA dalam persentase adalah sangat sulit: sulit: sedang: mudah: sangat mudah = 0%:12%:78%:10%:0%, jadi *Threshold* tes IPA dominan sedang. Persentase kecocokan butir tes IPA dengan Rasch Model adalah 88%.

Kata-kata Kunci: Karakteristik Tes.

Characteristics of Science Test

Prof. Dr. Suwanto M.Pd¹ dan Moh. Zain Bin Musa, Ph.D²

¹Universitas Veteran Bangun Nusantara Sukoharjo, e-mail: suwartowarto@yahoo.com

²Sekolah Bersepadu Musa Asiah, Cambodia

Abstract: The research objective is to describe: the characteristics of science tests based on classical test theory and modern test theory. The research design is quantitative and descriptive. The objects of research are science achievement tests during Covid-19, science teachers, school principals, and deputy principals. The data was obtained from the responses of 280 students to all answer sheets of class VIII SMP MTA Gemolong Sragen as the population of this study. The answer keys for science questions and one package of science questions (50 multiple choice items) were obtained from the science teacher. Research techniques with interviews and documentaries. Data analysis was carried out using the Quest program. Research results: (1). Characteristics of science test based on classical test theory: Content validity is not met, test reliability is 0.960, item difficulty category in percentage is easy: moderate: difficult = 10%: 84%: 6%, item discriminatory category in percentage is poor: enough: good: very good = 2%: 4%: 14%: 80%, so the dominant distractor is very good, while the distractor function in percentage is not effective: effective = 0.70%: 99.30%, so the distractor dominant is effective. (2). The characteristics of the science test are based on modern test theory: the Threshold category for the science test in percentage is very difficult: difficult: medium: easy: very easy =

0%:12%:78%:10%:0%, so the dominant science test Threshold is moderate. The percentage of match between the IPA test items and the Rasch Model is 88%.

Keywords: *Test Characteristics.*

Pendahuluan

Setiap pendidikan pada waktu-waktu tertentu dalam suatu masa pendidikan selalu melakukan evaluasi (Sultana, 2018: 2). Artinya guru selalu melakukan penilaian terhadap hasil yang telah dicapai siswa pada waktu-waktu tertentu selama masa pendidikan. Penilaian hasil belajar siswa harus dilakukan secara berkesinambungan dengan kata lain guru harus terus menerus mengikuti hasil belajar yang telah dicapai oleh siswa dari waktu ke waktu. Guru sebagai pendidik adalah memberikan umpan balik kepada para siswanya tentang kemajuan mereka serta membantu memperbaiki perkembangan belajarnya.

Tes adalah salah satu alat ukur paling efektif yang digunakan guru untuk mengukur kuantitas dan kualitas pembelajarannya. Crocker dan Algina (1986) menggambarkan tes sebagai prosedur standar untuk mendapatkan sampel perilaku dari domain tertentu. Tes adalah instrumen yang dibuat dengan baik yang, secara total, mengukur hasil belajar yang realistis yang mewakili ciri-ciri perilaku yang diharapkan. Etsey (2004) menyarankan bahwa tujuan pembelajaran yang komprehensif mencakup perilaku yang dapat diamati, kondisi di mana perilaku yang dimaksudkan harus diwujudkan, dan tingkat kinerja yang dianggap cukup untuk menunjukkan penguasaan hasil belajar membantu dalam menilai pengetahuan dan konsep yang mengarah pada kognitif, afektif, dan perkembangan psikomotorik siswa.

Tes lebih banyak digunakan untuk mengevaluasi hasil belajar siswa ditinjau dari ranah kognitif. Dalam penilaian pengajaran Ilmu Pengetahuan Alam (IPA), aspek kognitif sering dijadikan tolak ukur untuk mencapai hasil belajar bahasa IPA. Hal ini terlihat pada penilaian akhir pembelajaran bahasa IPA yang hanya menilai aspek kognitif saja karena butir tes yang digunakan hanya mengukur penguasaan pengetahuan terhadap materi yang diajarkan. Untuk mengevaluasi prestasi IPA siswa, guru biasanya memberikan siswa beberapa pertanyaan dalam sebuah tes. Guru dapat melaksanakannya setelah setiap bab materi selesai atau pada akhir semester. Tes tersebut disebut tes prestasi. Tes prestasi adalah tes yang menjadi fokus pengukuran adalah tujuan pembelajaran. Tes prestasi merupakan instrumen penugasan dalam pendidikan yang sangat penting sebagai sumber informasi untuk pengambilan keputusan. Ini adalah salah satu alat ukur yang paling banyak digunakan untuk menentukan hasil belajar siswa dalam proses belajar-mengajar atau program pendidikan. Hal ini penting dilakukan oleh guru, sekolah dan lembaga pendidikan untuk mengetahui sejauh mana siswa telah mencapai tujuan pembelajaran yang diharapkan. Oleh karena itu, peneliti menyimpulkan bahwa tes prestasi berupa tes yang disusun secara terencana untuk mengungkapkan kinerja maksimal yang telah diajarkan. Guru, sekolah, atau lembaga pendidikan lainnya dapat menggunakan hasil tes prestasi untuk mengambil keputusan atau memberikan umpan balik untuk meningkatkan proses belajar mengajar. Dalam kegiatan pendidikan formal di kelas, tes prestasi dapat berupa ulangan harian, ulangan formatif, ulangan sumatif bahkan ebtanas dan ujian masuk perguruan tinggi (Suwanto, 2013: 87).

Tes sumatif adalah kegiatan penilaian yang menghasilkan angka, yang kemudian digunakan untuk menentukan prestasi siswa. Tes ini dilakukan jika satuan pengalaman



belajar atau seluruh materi pelajaran telah tuntas. Tes sumatif digunakan untuk menentukan klasifikasi penghargaan pada akhir suatu mata kuliah atau program (Putri, 2017: 7 & Sugianto, 2017: 25). Di sisi lain, tes formatif digunakan untuk melacak bagaimana kemajuan siswa dalam studi mereka dan memberi mereka umpan balik yang dapat mereka gunakan untuk meningkatkan kinerja mereka sebagai guru dan siswa. Tes formatif membantu siswa lebih memahami kekuatan dan keterbatasan mereka dan bagaimana mereka dapat meningkatkan di bidang tersebut sementara juga memungkinkan guru untuk melihat di mana siswa mereka mengalami kesulitan dan mengambil tindakan cepat untuk membantu mereka.

Seorang guru sebagai pengembang tes harus mengetahui dasar-dasar penyusunan tes prestasi belajar yang baik untuk memperoleh hasil pengukuran yang valid dan reliabel. Pembelajaran, pengajaran, dan pengetahuan konten semuanya harus sinkron agar skor tes menjadi valid. Ketika ini terjadi, nilai tes adalah aktual. Hal ini didukung oleh Mulianah & Hidayat, 2013: 29; Suwarto, 2016: 3 dan Cheng, Yang & Du, 2019: 109, untuk memperoleh skor yang sebenarnya, diperlukan tes yang praktis untuk mengidentifikasi secara akurat. Sebuah tes yang baik harus terdiri dari item yang baik yang memenuhi kriteria tes dan menawarkan informasi aktual dengan kesalahan seminimal mungkin. Hasil tes yang berkualitas tinggi merupakan kunci yang dapat menjelaskan hasil belajar yang sebenarnya. Menurut Suwarto (2021: 358), suatu tes dikatakan tes yang baik, dan harus memenuhi ciri-ciri tes yang baik. Ini adalah; uji validitas, uji reliabilitas, tingkat kesulitan butir, daya beda butir dan terutama untuk tes pilihan ganda, memiliki distraktor yang efektif untuk setiap butir. Menganalisis butir tes diperlukan untuk menentukan tingkat validitas dan reliabilitas penilaian. Akibatnya, kualitas tes akan mempengaruhi hasil tes. Kualitas setiap butir mempengaruhi kualitas tes. Guru hendaknya fokus pada kualitas butir soal, sehingga guru perlu melakukan analisis butir soal karena dengan menganalisis butir soal, guru dapat mengidentifikasi kualitas setiap butir soal, mengetahui butir soal mana yang sesuai dengan kriteria, butir soal mana yang harus dihapus, dan butir mana yang harus direvisi.

Masa pandemi *Corona Virus Disease 2019 (Covid-19)* sejak Desember 2019 hingga saat ini, semua proses belajar mengajar, termasuk ujian, untuk sementara dilakukan di rumah. Hal itu perlu dilakukan untuk meminimalisir kontak fisik massal untuk memutus mata rantai penyebaran virus. Oleh karena itu, melalui pembelajaran jarak jauh dengan media handphone, PC (*personal Computer*), maupun laptop, dilakukan evaluasi dan tes. Sebuah media dinilai sangat efektif untuk mencegah penyebaran Covid-19 di lingkungan pendidikan. Guru memberikan tes yang dikirim melalui ponsel atau laptop siswa atau orang tua. Kemudian siswa mengerjakan tugas atau tes dari rumah. Di masa pandemi, tes prestasi IPA dilakukan oleh guru mata pelajaran secara mandiri karena keterbatasan jarak, sehingga tes yang baik yang dibuat oleh guru mata pelajaran perlu diselidiki. Berdasarkan wawancara yang dilakukan peneliti, tes prestasi IPA selama Covid-19 dilakukan oleh seorang guru IPA. Tes tidak diujicobakan; bahkan guru IPA membuat tes tanpa membuat kisi-kisi terlebih dahulu yang sesuai dengan silabus. Oleh karena itu, tujuan dari penelitian ini adalah untuk mengetahui validitas, reliabilitas, tingkat kesukaran butir, daya beda butir, dan keefektifan distraktor berdasarkan teori tes klasik, demikian juga akan dilihat bagaimana karakteristik tes berdasarkan teori tes modern. Penelitian ini diharapkan dapat memberikan masukan dan contoh bagi guru IPA, pendidik, pengembang tes, dan pihak lain



yang membuat tes prestasi. Selain itu, penelitian ini dilakukan untuk memberikan referensi bagi penelitian serupa di masa yang akan datang.

Menurut Suryabrata (2005: 68) dan Fernandez (1984: 5), pengembangan tes prestasi spesifikitas memiliki wilayah, tes mata pelajaran, tujuan tes, materi yang akan dibuatkan tes, jenis tes, dan jumlah butir dalam tes tersebut. Kemudian, mereka merancang kisi-kisi tes yang meliputi tujuan khusus, nilai khusus, dan indikator. Dalam membangun tes, mereka mencetak butir tes sesuai dengan kisi-kisi yang telah dibuat, setelah itu, tes harus divalidasi secara kualitatif ataupun direview butir-butir tes, penilaian profesional, validasi kuantitatif, validasi teori, materi, konstruksi, validasi bahasa, validitas isi. Kemudian, mereka merevisi tes sesuai masukan para reviewer, setelah itu, semua butir tes yang sudah baik dirangkai menjadi sebuah tes. Setelah menyelesaikan suatu tes, maka harus melakukan uji coba kepada sekelompok siswa sesuai dengan mata pelajaran tes tertentu untuk menganalisis karakteristik tes. Untuk menganalisis, ada teori tes klasikal: tingkat kesukaran butir, daya beda butir, keberfungsian pengecoh, reliabilitas, dan validitas isi. Setelah itu, butir tes dipilih berdasarkan hasil analisis uji (teori uji klasik: diterima, revisi, dan ditolak atau teori tes modern: nilai ambang batas, diterima atau ditolak dan kesesuaian model Rasch atau model logistik satu parameter). Terakhir, butir-butir tes yang lolos disusun menjadi tes standar. Selanjutnya tes dicetak dan didistribusikan ke siswa atau di sekolah-sekolah.

Karakteristik tes yang baik akan difokuskan pada analisis butir soal kuantitatif. Richard & Sheila. (1999: 3) menjelaskan bahwa, analisis butir secara kuantitatif adalah studi butir berdasarkan data empiris dari tes yang diujikan. Ada dua macam dalam analisis butir secara kuantitatif, yaitu analisis berdasarkan teori tes klasik dan teori tes modern. Analisis butir soal berdasarkan teori tes klasik merupakan kajian soal melalui informasi yang diperoleh dari jawaban siswa guna meningkatkan kualitas soal dengan menggunakan teori tes klasik. Teknik ini memiliki beberapa keunggulan yaitu murah, mudah, dapat dilaksanakan dengan cepat, dan sederhana. Karakteristik tes adalah tingkat kesukaran soal, daya pembeda butir soal, pengecoh, validitas dan reliabilitas (Suwanto, 2021: 358). Penelitian yang pernah ada antara lain. Penelitian Huda dan Wahyuni (2019) dengan judul “Analisis Soal IPA *Try Out* USBN Tahun Ajaran 2018/2019 Dalam Kaitannya Dengan Tingkat Kognitif”. Mengetahui karakteristik soal *try-out* IPA USBN tahun ajaran 2018/2019 berdasarkan *Classical Test Theory* (CTT) dan mengetahui jumlah soal *try-out* IPA USBN 2018/2019 dalam kaitannya dengan tingkat kognitif. Jenis penelitian ini adalah penelitian deskriptif dengan pendekatan kuantitatif. Mereka menganalisis soal *try-out* IPA USBN menggunakan program ITEMAN. Mereka menggunakan dokumentasi untuk teknik pengumpulan data. Data yang diperoleh adalah Lembar Jawaban Komputer (LJK) sebanyak 5.022 siswa yang mengikuti *try-out* SD/MI USBN 2019 pada 21 Februari 2019 di Kabupaten Tulungagung.

Penelitian oleh Hamimi, Zamharirah & Rusydy (2020), penelitian untuk mengetahui tingkat kualitas soal yang terdiri dari validitas, tingkat kesukaran, reliabilitas, pengecoh, dan daya pembeda soal matematika di SMP N 1 Susoh semester gasal kelas VII tahun pelajaran 2017-2018. Soal-soal tes dibuat langsung oleh guru matematika, dan terlebih dahulu membuat kisi-kisi soal berdasarkan standar kompetensi dan kompetensi dasar untuk semester gasal tahun pelajaran 2017-2018. Bentuk soal tes yang diberikan kepada siswa berupa pilihan ganda dan essay. Namun, peneliti hanya menganalisis pilihan ganda. Tes itu dibuat oleh MGMP (Musyawarah Guru Mata Pelajaran) Matematika. Hasil penelitian menunjukkan bahwa pertanyaan yang digunakan relatif tidak valid, karena masih ada



beberapa pertanyaan yang memiliki validitas rendah dan sangat rendah. Solusinya, soal tersebut tidak digunakan dan direvisi. Selain itu, soal-soal yang diteliti juga memiliki tingkat reliabilitas yang rendah. Namun soal tersebut memiliki tingkat kesukaran yang relatif baik, dengan hasil tes menunjukkan sebagian besar soal memiliki tingkat kesukaran sedang. Hasil daya pembeda menunjukkan hasil yang baik.

Kesamaan penelitian Richard & Sheila dan Hamimi, Zamharirah & Rusydy adalah sama-sama analisis kuantitatif yang dilakukan dengan program Iteman. Program Iteman hanya mampu untuk menganalisis kualitatif berdasarkan teori tes klasik. Program Iteman tidak mampu untuk menganalisis kuantitatif berdasarkan teori tes modern. Adapun yang telah dilakukan oleh penulis adalah analisis kuantitatif berdasarkan teori tes klasik dan berdasarkan teori tes modern, dengan menggunakan program Quest.

Metode Penelitian

Desain penelitian adalah kuantitatif dan deskriptif. Penelitian kuantitatif karena peneliti menghitung karakteristik tes (tingkat kesukaran item, daya pembeda butir, keberfungsian pengecoh, reliabilitas tes, nilai ambang batas, Infit Meansquare, Outfit t, dan item fit) dengan menggunakan program Quest. Penelitian deskriptif karena peneliti menggambarkan karakteristik tes. Tempat penelitian di SMP MTA Gemolong Sragen. Objek penelitian adalah tes prestasi IPA selama Covid-19, guru IPA, kepala sekolah, dan wakil kepala sekolah. Ada satu set tes IPA yang terdiri dari 50 butir soal pilihan ganda. Data diperoleh dari tanggapan siswa terhadap seluruh lembar jawaban siswa kelas VIII sebagai populasi penelitian ini. Ada 280 lembar jawaban siswa untuk tahun 2021. Kunci jawaban soal IPA dan satu paket soal IPA diperoleh dari guru IPA. Peneliti telah melakukan wawancara tidak terstruktur berupa pertanyaan terbuka sebagai salah satu teknik pengumpulan data. Hal ini didasarkan pada metode penelitian yang digunakan oleh peneliti, yang sangat bergantung pada pemahaman peneliti dan data informasi yang diperoleh dari observasi dan wawancara. Peneliti meminta izin kepada pihak administrasi dan kepala sekolah untuk melakukan penelitian di SMP MTA Gemolong Sragen. Kedua, peneliti meminta guru IPA kelas VIII untuk mendapatkan informasi tentang kurikulum program sekolah, data siswa kelas VIII tahun 2021 secara keseluruhan, dan menanyakan bagaimana tes prestasi IPA dibuat selama masa pandemi Covid-19. Analisis data dilakukan dengan menggunakan program Quest.

Hasil Penelitian dan Pembahasan

Merancang tes prestasi IPA selama Covid-19, tidak ada tahapan. Itu dibuat oleh seorang guru IPA yang mengajar langsung untuk murid-muridnya. Berdasarkan wawancara peneliti dengan dia, dia langsung membuat tes tanpa membuat kisi-kisi. Jadi, dia membuatnya langsung soal menyesuaikan apa yang dia ajarkan di kelas selama periode tertentu. Dia hanya menyalin dan menempel dari tes sebelumnya yang dibuat MGMP, dirinya sendiri dan guru IPA lainnya. Selain itu, dia tidak mengujicobakan untuk menganalisis karakteristik tes seperti tingkat kesukaran butir, daya beda butir, dan keberfungsian pengecoh. Dan untuk keseluruhan tes juga tidak dianalisis seperti uji validitas dan uji reliabilitas.

Hasil Analisis berdasarkan Teori Tes Klasik.

Indeks kesulitan butir terendah adalah 0,225 pada butir 49 dan indeks kesulitan butir tertinggi adalah 0,739, yaitu butir 1 dan butir 27. Berdasarkan indeks kesulitan butir, dapat



disimpulkan bahwa butir tersulit dari tes prestasi IPA yang dibuat oleh guru IPA adalah butir 49 sedangkan butir tes yang paling mudah adalah butir 1 dan butir 27. Rangkuman tingkat kesukaran butir berdasarkan kategori dalam tes prestasi IPA disajikan dalam bentuk Tabel 1.

Tabel 1. Rangkuman Tingkat Kesulitan (p) butir dari Tes Prestasi IPA

Kategori	Nomor Butir	Total	Persentase
Mudah ($0,70 < p \leq 1,00$)	1,12,27,31,35	5	10
Sedang ($0,30 \leq p \leq 0,70$)	2,3,4,5,6,7,8,9,10, 11,13,14,15,16,17,18, 19,20,21,22,23,24,25,26,28,29,30,32,33,34,3 6,37,38,39,40,41,43,44,46,47,48,50	42	84
Sulit ($0,00 \leq p < 0,30$)	42,45,49	3	6
Total		50	100

Tes IPA memiliki 5 butir soal mudah dengan persentase 10%, 42 butir soal sedang dengan persentase 84%, dan 3 butir soal sulit dengan persentase 6%. Berdasarkan hasil tersebut, tingkat kesukaran butir soal lebih dominan pada butir-butir sedang, sehingga peneliti menyimpulkan bahwa soal tersebut tidak memiliki kesukaran butir yang proporsional, padahal tes yang ideal seharusnya terdiri dari 25% soal mudah, 50% soal sedang, dan 25% soal sulit (Kunandar, 2013 & Suwanto, 2021). Roid & Haladyna (1982) menyatakan bahwa tes yang tidak memiliki tingkat kesulitan butir proporsional tidak dapat mengungkapkan kompetensi siswa yang sebenarnya. Tes juga lebih dominan pada butir sedang. Brown (2004) menegaskan bahwa butir yang dibuat dengan baik tidak boleh terlalu mudah atau sulit, tesnya harus seimbang sehingga seorang guru IPA dapat memperoleh informasi tentang kompetensi IPA siswa. Sebaliknya, Haider et al. (2012) berpendapat bahwa kategori butir sedang dapat menunjukkan bahwa siswa memiliki pemahaman yang baik dalam menjawab tes karena lebih dari separuh siswa menjawab butir dengan benar. Tingkat kesukaran butir soal tes ini dapat dibandingkan dengan penelitian lain yang menguji tingkat kesukaran butir soal tes sumatif (Mulianah & Hidayat, 2013; Maharani & Putro, 2020; Saputra, Retnawati & Yusron, 2021), meskipun kondisi tesnya tidak sama. Penelitian-penelitian sebelumnya menemukan tingkat kesukaran soal yang memiliki butir-butir yang lebih sedang dibanding yang lain.

Tabel 2. Rangkuman Daya Beda Butir ($r_{Pt.Biser}$) Tes Prestasi IPA

Kategori	Nomor Butir	Total	Persentase
Jelek ($r_{Pt.Biser} \leq 0,19$)	6	1	2
Cukup ($0,20 < r_{Pt.Biser} < 0,29$)	42,49	2	4
Baik ($0,30 < r_{Pt.Biser} < 0,39$)	1,30,33,34,46,47,50	7	14
Sangat Baik ($0,40 \leq r_{Pt.Biser}$)	2,3,4,5,7,8,9,10,11,12,13,14,15,16,17,18,19 ,20,21,22,23,24,25,26,27,28,29,31,32,35,36 ,37,38,40,41,43,44,45,48	40	80
Total		50	100



Hal ini menunjukkan bahwa tes tersebut lebih banyak butir yang dibangun dengan baik daripada butir yang dibangun dengan jelek, namun tidak seimbang antara butir mudah, sedang, dan sulit. Sulitnya butir yang tidak seimbang diduga karena adanya pandemi Covid-19 yang mengharuskan siswa bekerja di rumah, sehingga siswa dapat minta pertolongan teman, atau bisa browsing internet. Itu semua dapat mempengaruhi indeks kesulitan butir. Daya beda butir terendah adalah 0,13 pada butir 6 dan daya beda butir tertinggi adalah 0,74 pada butir 28. Rangkuman daya beda butir berdasarkan kategori dalam tes prestasi belajar IPA yang dilakukan guru IPA disajikan pada Tabel 2. Daya beda butir soal tes ini baik. Berdasarkan program Quest menunjukkan bahwa 1 butir jelek dengan persentase 2%, 2 butir diterima dengan persentase 4%, 7 butir baik dengan persentase 14%, 40 butir sangat baik dengan persentase 80%. Hasil ini menunjukkan bahwa 1 butir yang jelek harus didrop dan 2 item yang dapat diterima harus direvisi (Dichoso & Joy, 2020). Hasil ini bagus karena 80% butir sangat bagus dan 14% butir baik (Dichoso & Joy, 2020). Artinya sebagian besar soal dapat digunakan untuk mengukur kompetensi IPA siswa yang sebenarnya. Butir-butir tersebut juga dapat membedakan siswa yang berprestasi tinggi, berprestasi sedang, dan berprestasi rendah. Hal ini sesuai dengan Suwarto (2021) semakin besar indeks diskriminasi butir menyiratkan bahwa butir tersebut semakin dapat membedakan antara siswa yang berprestasi rendah dan berprestasi tinggi. Hal ini untuk mendeteksi perbedaan individu di antara siswa. Hasil pengujian ini dapat dibandingkan dengan penelitian lain (Boopathiraj & Chellamani, 2013; Singh et al., 2014; Saputra et al., 2021), meskipun kondisi pengujiannya tidak serupa. Para peneliti tersebut menemukan daya beda butir yang baik. Sedangkan hasil yang berbeda ditemukan dari penelitian sebelumnya seperti (Sa'adah, 2017; Toksöz & Ertunç, 2017; Rehman, Aslam & Hassan, 2018; Manalu, 2019; Karim, Sudiro & Sakinah, 2021) yang melaporkan bahwa butir dengan daya beda jelek, maka butir tersebut tidak dapat membedakan antara siswa yang berprestasi tinggi dan siswa yang berprestasi rendah.

Distraktor adalah jawaban pilihan ganda yang pasti jawabannya salah. Fungsinya membuat siswa bingung atau salah perhitungan ketika memilih jawaban yang benar di antara alternatif-alternatif yang disediakan. Distraktor dikatakan efektif, jika dipilih lebih dari 5% dari responden, dalam bilangan desimal 0,050. Distraktor dikatakan tidak efektif, jika dipilih kurang dari 5% dari responden atau dalam bilangan desimal $< 0,050$. (Suwarto, 2021: 364). Berdasarkan pengecoh efektif dan pengecoh tidak efektif tes prestasi IPA yang dibuat oleh guru IPA adalah sebagai berikut. Persentase pengecoh yang tidak efektif dari tes IPA ada 0,70%. Persentase pengecoh efektif tes IPA ada 99,30%. Pengecoh tes ini memiliki 1 pengecoh tidak efektif (0,7%) dari 150 pengecoh yang harus direvisi dan 149 pengecoh efektif (99,30%) dari 150 pengecoh. Hasil prosentase distraktor efektif penelitian ini hampir sama dengan penelitian sebelumnya yaitu Maharani & Putro, 2020. Mereka menemukan 80% pengecoh efektif. Hasil ini dapat dijelaskan bahwa indeks pembeda butir soal dapat mempengaruhi indeks pengecoh. Sebagian besar butir tes IPA mampu membedakan antara yang berprestasi tinggi dan rendah yang dapat diasumsikan bahwa indeks daya beda butir yang tinggi dapat menyebabkan indeks pengecoh yang efektif (Kheyami, Jaradat, Al-Shibani, & Ali, 2018). Mereka juga mengatakan bahwa jumlah pengecoh yang ideal minimal 3 pengecoh perbutir. Hasil dari penelitian ini adalah pengecoh yang lebih efektif sehingga kualitas butir soal semakin baik. Hasil penelitian tes prestasi IPA, hampir semua butir memiliki pengecoh yang efektif. Diasumsikan guru IPA merancang tes itu sendiri, sehingga guru sudah mengetahui karakteristik siswa.



Reliabilitas tes prestasi IPA yang dibuat oleh guru IPA adalah 0,960. Hal ini menunjukkan bahwa butir-butir tes sangat reliabel. Sebuah tes dengan tingkat reliabilitas yang tinggi tergolong tes yang baik (Sa'adah, 2017). Selain itu, tes yang baik dapat digunakan untuk pengujian waktu berikutnya. Hasil penelitian ini juga menunjukkan bahwa sejauh mana pengukuran tes prestasi IPA tetap konsisten setelah diulang pada subjek dan dalam kondisi yang sama (Rudyatmi & Rusllowati, 2017). Uji reliabel ini hampir sama dengan penelitian sebelumnya (Anggreyani, 2009; Mulianah & Hidayat, 2013; Pascual, 2016; Sugianto, 2017; Manalu, 2019; Saputra et. al., 2021) meskipun kondisi pengujian tidak sama. Mereka menemukan tes yang andal. Estimasi reliabilitas tes dapat dipercaya karena jauh di atas batas koefisien reliabilitas 0,700. Beberapa faktor yang mempengaruhi estimasi reliabilitas, antara lain homogenitas kelompok, alokasi waktu, dan lama tes. Selain itu, faktor lain yang mempengaruhi estimasi reliabilitasnya adalah jumlah butir soal yang tergolong sukar (Crocker dan Algina, 1986).

Analisis berdasarkan teori tes klasik di atas memiliki kelemahan yaitu karakteristik butir soal bergantung pada kelompok peserta tes yang dikenai butir soal. Pada teori tes klasik statistik soal seperti indeks kesukaran soal tergantung pada kelompok peserta tes, jika tes dikerjakan oleh siswa yang pandai maka soalnya terasa mudah (tingkat kesukaran soal menjadi besar) dan sebaliknya jika tes dikerjakan oleh siswa yang tidak pandai maka soal menjadi sulit (tingkat kesukarannya menjadi kecil). Oleh karena itu, karakteristik soal tidak konsisten atau berubah-ubah tergantung pada kemampuan siswa yang mengikuti ujian. Analisis berdasarkan teori tes klasik memiliki kelemahan karena karakteristik tes tergantung pada kelompok tinggi dan kelompok rendah. Jika tes dikerjakan oleh yang berprestasi tinggi, maka tesnya mudah, sehingga indeks kesukaran butir soal menjadi lebih tinggi. Sebaliknya, jika tes dikerjakan oleh siswa yang berprestasi rendah, maka tes tersebut sulit, sehingga indeks kesukaran butir soal menjadi lebih rendah. Dengan demikian, hal ini menunjukkan bahwa analisis tes berdasarkan teori tes klasik, karakteristik tes tidak konsisten atau berubah tergantung pada prestasi siswa (Hambleton, Swaminathan & Rogers, 1991: 8). Oleh karena itu, peneliti melanjutkan menganalisis karakteristik berdasarkan teori tes modern untuk menganalisis karakteristik tes.

Hasil Analisis Berdasarkan Teori Tes Modern.

Tabel 3. Rangkuman Kategori *Threshold (b)* Tes IPA

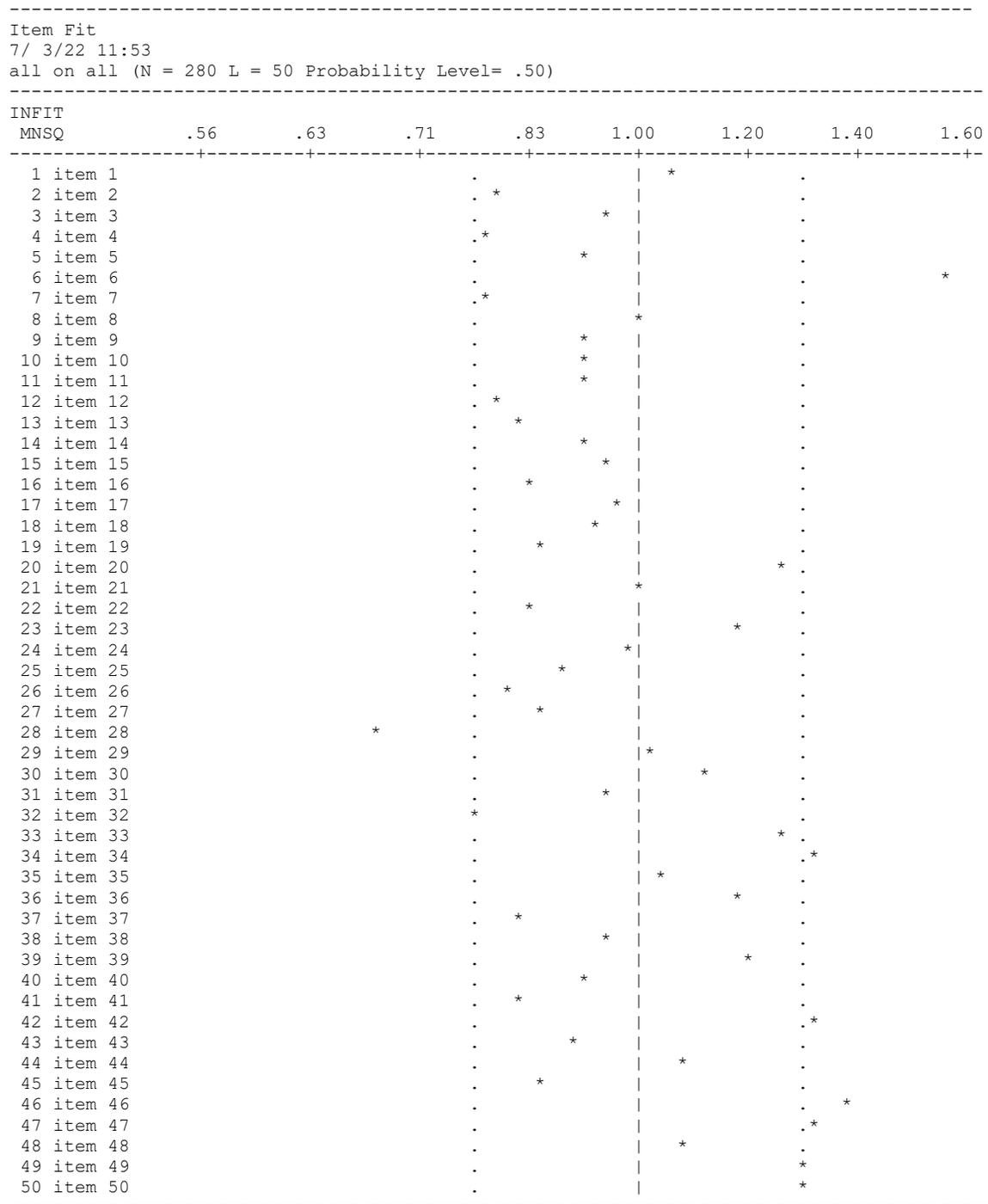
Analisis karakteristik tes berdasarkan teori tes modern menggunakan logistik satu

Kategori	Nomor butir	Total (%)
Sangat sulit ($b > 2$)	-	0 (0%)
Sulit ($1 < b \leq 2$)	6,21,42,45,48,49	6 (12%)
Sedang ($-1 \leq b \leq 1$)	2,3,4,5,7,8,9,10,11,13,14,15,16,17,18,19,20,22,23,24,25, 26,28,29,30, 32,33,34,36,37,38,39,40,41,43,44,46,47,50	39 (78%)
Mudah ($-1 > b > -2$)	1,12,27,31,35	5 (10%)
Sangat mudah ($b < -2$)	-	0 (0%)
Total		50 (100%)

parameter (1PL) karena program Quest hanya dapat menganalisis satu parameter model



logistik (Adams & Khoo, 1996: 25). Berdasarkan Tabel 3, persentase Threshold tes prestasi IPA = sangat sulit: sulit: sedang: mudah: sangat mudah = 0%:12%:78%:10%:0%.



Gambar 1. Item fit map untuk Tes Prestasi IPA

Berdasarkan Gambar 1, dapat dilihat bahwa 6 butir soal tes prestasi IPA tidak fit karena statistik asteris out of fit yang berada di antara dua garis vertikal putus-putus, yaitu: butir 6,



28, 34, 42, 46, dan 47, sedangkan 44 butir lainnya fit (Adams & Khoo, 1996: 25). Persentase kecocokan butir tes IPA dengan Rasch Model = $44/50 \times 100\% = 88\%$.

Tabel 4. Analisis Butir Tes Berprestasi IPA yang Diterima dan Ditolak

Kategori (Kriteria)	Item Tes	Jumlah (%)
Diterima (<i>Outfit</i> $t \leq 2,00$)	1,2,3,4,5,7,8,9,10,11,12,13,14,16,17,18,19,21,22,23, ,25,26,27,28,29, 30,31,32,33,35,36,37,38,39,40,41,43,44,45,48	40 (80%)
Ditolak (<i>Outfit</i> $t > 2,00$)	6,15,20,24,34,42,46,47,49,50	10 (20%)
Total		50 (100%)

Berdasarkan Tabel 4, persentase soal tes IPA yang lolos = diterima: ditolak = 80%:20%. Sedangkan kalau dilihat *Item Fit Map* tes IPA dapat dilihat Gambar 1.

Simpulan dan Saran

Karakteristik tes IPA berdasarkan teori tes klasik: Validitas isi tidak terpenuhi, reliabilitas tes 0,960, kategori tingkat kesulitan butir dalam persentase adalah mudah: sedang: sulit = 10%:84%:6%, kategori daya beda butir dalam persentase adalah jelek: cukup: baik: sangat baik = 2%: 4%: 14%: 80%, jadi domina sangat baik, sedangkan keberfungsian distraktor pengkecoh dalam persentase adalah tidak efektif: efektif = 0,70%: 99,30%, jadi dominan pengkecoh yang efektif. Karakteristik tes IPA berdasarkan teori tes modern: kategori Threshold tes IPA dalam persentase adalah sangat sulit: sulit: sedang: mudah: sangat mudah = 0%:12%:78%:10%:0%, jadi *Threshold* tes IPA dominan sedang. Persentase kecocokan butir tes IPA dengan Rasch Model adalah 88%.

Saran yang dapat diberikan tingkat kesulitan butir seharusnya dibuat 25 persen mudah, 50 persen sedang, 25 sulit. Dengan demikian, kemampuan siswa yang rendah, sedang, dan tinggi semuanya dapat diukur. Validitas isi seharusnya terpenuhi, sehingga tes IPA memiliki butir-butir yang dapat mengukur apa yang seharusnya diukur (semua aspek yang harus diukur terwakili dalam butir-butir tes IPA). Butir-butir yang menyusun tes IPA seharusnya berkesuaian dengan Rasch Model.

Daftar Rujukan

- Adams, R. J. & Khoo, S. T. (1996). *Quest the Interactive Test Analysis System*. Australia: The Australian Council for Educational Research Ltd.
- Angg्रेyani, A. (2009). *Penerapan Teori Uji Klasik dan Teori Respon Butir dalam Mengevaluasi Butir Soal*. Departemen Statistika Fakultas Matematika Dan Ilmu Pengetahuan Alam Institut Pertanian Bogor.
- Boopathiraj, C. & Chellamani, K. (2013). Analysis of Test Items on Difficulty Level and Discrimination Index in The Test pfor Research in Education. *International Journal of Social Science & Interdisciplinary Research* 2(2), 189-193. Retrieved from: indianresearchjournals.com.



- Brown, H. D. (2004). *Language Assessment: Principle and Classroom Practice*. United States of America: Pearson Education.
- Cheng, Y., Yang, Y. & Du, D. (2019). A class of asymptotically optimal group testing strategies to identify good items . *Discrete Applied Mathematics Journal*. 260, 109–116.
- Crocker, L. & Algina, J. (1986). *Introduction to Classical and Modern Test Theory*. New York: CBS College.
- Dichoso, A. A. & Joy M. R. J. (2020). Test Item Analyzer using Point-Biserial Correlation and P-Values. *International Journal Of Scientific & Technology Research*, 9(4). 2122-2126
- Etsey, Y. K. (2004). “Assessing performance in schools: Issues and practice,” *Ife Psychologia*, vol. 13, no. 1, 123-135
- Fernandez. H. J. X. (1984). *Testing and Measurement*. Jakarta: National Education Planning Evaluation and Curriculum Development.
- Hambleton, R.K., Swaminathan, H., & Rogers, H.J. (1991). *Fundamentals of Item Response Theory*. London: Sage Publication.
- Hamimi, L., Zamharirah, R. & Rusydy. (2020). Analisis Butir Soal Ujian Matematika Kelas VII Semester Ganjil Tahun Pelajaran 2017/2018. *Mathema Journal*. 2(1). 57-66.
- Huda, N. & Wahyuni, T. S. (2019). Analisis Butir Soal IPA Try Out USBN Tahun Ajaran 2018/2019 dalam Kaitannya dengan Level Kognitif. *Jurnal Pendidikan dan Pembelajaran Dasar*. 12(1), 29-39.
- Kheyami, D., Jaradat, A., Al-Shibani, T., & Ali, F. A. (2018). Item Analysis of Multiple-Choice Questions at the Department of Paediatrics, Arabian Gulf University, Manama, Bahrain. *Sultan Qaboos Univesity Medical Journal*, 18(1), 68-74.
- Kunandar, K. (2013). *Penilaian Autentik: Penilaian Hasil Belajar Peserta Didik Kurikulum 2013*. RajaGrafindo Persada.
- Maharani, A. V. & Putro, N. H. P. S. (2020). Item Analysis of English Final Semester Test. *Indonesian Journal of EFL and Linguistics*. 5(2), 2020. 492-504.
- Manalu, D. (2019). An Analysis of Students Reading Final Examination by Using Item Analysis Program on Eleventh Grade of SMA Negeri 8 Medan. *Journal of English Teaching & Applied Linguistics*, 1(1), 13-19.
- Mulianah, S. & Hidayat, W. (2013). Pengembangan Tes Berbasis Komputer. *Kuriositas*, 2(6), 27- 43.
- Pascual, G. R. (2016). Analysis of The English Achievement Test for ESL Learners in Northern Philippines. *International Journal of Advanced Research in Management and Social Sciences*, 5(12),1-5. retrieved from www.garph.co.uk.
- Putri, B. D. T. (2017). the Validity Analysis of English Summative Test of Junior High School. *Journal of Languages and Language Teaching*. 5(1), 6-11.
- Rehman, A., Aslam, A. & Hassan, S. H. (2018). Item Analysis of Multiple-Choice Questions. *Pakistan Oral & Dental Journal*, 38(2), 291-293.
- Richard & Sheila. (1999). *Item Analysis for Criterion-Referenced Tests*. New York: Research Foundation of SUNY/Center for Development of Human Services.
- Roid, G. H. & Haladyna, T. M. (1982). *A Technology for Test-Item Writing*. London: Academic Press, Inc.
- Rudyatmi, Ely & Rusllowati, A. (2017). *Evaluasi Pembelajaran*. Semarang: Faculty of Mathematics and Science Unnes.



- Sa'adah, N. (2017). The Analysis of English Mid-Term Test Items based on the Criteria of a Good Test at the First Semester of the Eighth Grade Students Of Mts. Mathalibul Huda Mlonggo In The Academic Year Of 2016/2017. *Journal Edulingua*, 4(1),45-58.
- Saputra, A. N. S., Retnawati, H. & Yusron, E. (2021). Analysis Difficulties and Characteristics of Item Test of on Biology National Standard School Examination. *Advances in Social Science, Education and Humanities Research*, 542, 8-14.
- Singh, J. P., Kariwal P., Gupta S.B., & Shrotriya V.P. (2014). Improving Multiple Choice Questions (MCQs) through item analysis: An assessment of the assessment tool. *International Journal of Sciences & Applied Research*, 1(2), 53-57. Retrived from: www.ijisar.in.
- Sugianto, A. (2017). Validity and Reliability of English Summative Test for Senior High School. *Indonesian EFL Journal: Journal of ELT, Linguistics, and Literature*, 3(2), 22-38. P-ISSN: 2460-0938. E-ISSN: 2460-2604.
- Sultana, N. (2018). Test Review of the English Public Examination at the Secondary Level in Bangladesh. *Language Testing in Asia*, 8(16), 1-9.
- Suryabrata, S. (2005). *Pengembangan Alat Ukur Psikologis*. Yogyakarta: C.V Andi Offset.
- Suwarto. (2013). Difficulty, Difference, and Reliability Level of New Student Selection Test for Veteran Bangun Nusantara Sukoharjo University. *National Seminar on Science Education*. 652- 658.
- Suwarto. (2013). *Pengembangan Tes Diagnostik Dalam Pembelajaran*. Yogyakarta: Pustaka Pelajar.
- Suwarto. (2016). The Biology Test Characteristic of 7th Grade by The Period of The Odd Term. *Jurnal Penelitian Humaniora*. 17(1), 1-8.
- Suwarto. (2021). The Characteristics of Indonesia a Second- Semester Final Test for Eight-grade Students. *Turkish Online Journal of Qualitative Inquiry*. 12(9), 356-370.
- Toksöz, S. & Ertunç., A. (2017). Item Analysis of a Multiple-Choice Exam. *Advances in Language and Literary Studies*, 8(6), 141-146.

