Tarbawy : Jurnal Pendidikan Islam
ISSN : 2407-4462 (Cetak), 2614-5812 (Elektronik)
Vol. 7, No. 1, 2020, Hal. 61-70
DOI: https://doi.org/ 10.32923/tarbawy.v7i1.1187

Analisis Kualitas Tes Buatan Guru Melalui Pendekatan *Item Response Theory* (IRT) Model Rasch

Dinar Pratama[1]
[1]IAIN Syaikh Abdurrahman Siddik Bangka Belitung

***Info Artikel*:**
Diterima 1 April 2020
Direvisi 18 April 2020
Dipublikasikan 27 April 2020

***Kata Kunci*:**
Tes
Teori Item Response
Model Rasch

ABSTRAK

Tujuan utama penelitian ini dilakukan adalah untuk menganalisis dan mendeskripsikan karakteristik tes buatan guru Akidah Akhlak melalui pendekatan Item Response Theory (IRT) model Rasch. Penelitian ini merupakan penelitian kuantitatif deskriptif. Subjek pada penelitian ini adalah pola respon siswa terhadap tes berjumlah 67 dengan lima alternatif jawaban. Perangkat tes diambil dari pelaksanaan Ujian Akhir Semester tahun pelajaran 2018/2019 melalui teknik dokumentasi. Analisis data dilakukan melalui pendekatan IRT model Rasch dengan bantuan *software* QUEST. Hasil analisis menunjukkan, dari 30 item terdapat 28 item fit dengan model Rasch. Ditinjau dari tingkat kesulitan, terdapat 25% item dengan kategori sangat sulit, 21.4% sulit, 7.14% sedang, 46.4% mudah, dan 0% sangat mudah. Rentang nilai tingkat kesukaran berkisar antara -2.94 sampai 4.18. Nilai *reliability of item estimate* sebesar 0.94 dengan kategori baik sekali dan nilai *reliability of case estimate* sebesar 0.38 dengan kategori lemah.

Keywords:
Test
Item Response Theory
Rasch Model

ABSTRACT

The main objective of this research is to analyze and describe the characteristics of the test made by the morality teacher through the Rasch Model Item Response Theory (IRT) approach. This research is a descriptive quantitative research. The subjects in this study were 67 students' response patterns to the test with five alternative answers. The test set is taken from the implementation of the Final Semester 2018/2019 academic year through documentation techniques. Data analysis was performed through the Rasch IRT approach with the help of QUEST software. The analysis showed that of the 30 items there were 28 items fit in the Rasch model. Judging from the level of difficulty, there are 25% of items with very difficult categories, 21.4% difficult, 7.14% moderate, 46.4% easy, and 0% very easy. The range of difficulty levels ranges from -2.94 to 4.18. The value of reliability of item estimate is 0.94 with excellent category and the value of reliability of case estimate is 0.38 with weak category.

Koresponden:
Dinar Pratama
E-mail: dinarpratama24@gmail.com

Pendahuluan

Idealnya, instrumen tes hasil belajar yang disusun oleh guru dapat memberikan banyak informasi mengenai kemampuan siswa. Akan tetapi, persoalan yang banyak muncul saat guru menyusun instrumen tes ini ternyata masih banyak ditemukan kelemahan dalam proses penyusunannya sehingga menghasilkan tes yang tidak memiliki daya ukur yang valid. Instrumen tes yang tidak memiliki daya ukur yang valid tidak akan dapat memberikan informasi apapun mengenai kemampuan peserta tes. Untuk mengevaluasi pencapaian proses pembelajaran yang umum dilakukan oleh guru adalah dengan menyusun instrumen tes hasil belajar.

Mungkin guru akan mendapatkan siswa yang memperoleh skor tertinggi, terendah, dan nilai rata-rata dari pengukuran hasil belajar siswa. Tentu informasi ini belum dapat memberikan banyak informasi mengenai kemampuan siswa. Mungkin siswa banyak mendapat skor di atas rata-rata, dan guru beranggapan bahwa pembelajarannya sudah tercapai dengan optimal. Padahal ada kemungkinan siswa mendapat skor tinggi karena soal yang disusun oleh guru memiliki tingkat kesulitan yang rendah, pilihan jawaban yang tidak homogen sehingga siswa dapat dengan mudah mengidentifikasi pilihan jawaban yang salah. Jadi, jika guru dalam menilai pencapaian hasil belajar siswa hanya dengan melihat nilai tertinggi, terendah, dan rata-rata justru dikhawatirkan informasi tersebut akan menyesatkan guru.

Hasil survey yang dilakukan Prasetya, (2012) mengenai kemampuan guru dalam menyusun tes hasil belajar di Kota Magelang menunjukkan hanya 64 persen guru memahami bagaimana menyusun tes hasil belajar. Hasil studi

lainnya yang dilakukan Nazaruddin, (2017) menunjukkan bahwa, kecenderungannya guru dalam menyusun tes hasil belajar biasanya menggunakan tes yang sudah ada. Guru tidak dari awal menyusun tes tersebut. Selain itu, tes yang disusun ternyata tidak sesuai dengan tujuan pembelajaran. Masih lemahnya kemampuan guru menyusun tes ini juga ditemukan pada guru-guru yang sudah memiliki sertifikat pendidik. Hasil temuan studi Kinerja Guru Profesional yang dilakukan Kartowagiran, (2011) terhadap 55 guru di Kabupaten Sleman yang sudah sertifikasi menunjukkan hasil yang belum baik. Dari 17 indikator yang diukur, hanya 7 indikator yang menunjukkan hasil baik. Penelitian Wiguna, (2018) menunjukkan bahwa, kemampuan guru Pendidikan Agama Islam (PAI) dalam menyusun tes hasil belajar tergolong rendah. Sebagian besar dari mereka kesulitan dalam melakukan analisis tes.

Selain itu, hasil penelitian Hamid, (2016) terhadap guru Madrasah juga menunjukkan masih rendahnya kemampuan guru menyusun dan menganalisis tes hasil belajar. Bukti lain lemahnya guru dalam menyusun tes hasil belajar ini juga dapat dilihat dari penelitian yang dilakukan oleh Setiadi, (2016) mengenai pelaksanaan penilaian pada kurikulum 2013. Penelitian yang dilakukan terhadap sekolah dasar dan menengah di 15 provinsi menunjukkan masih banyak guru yang belum mengerti dalam menyusun kisi-kisi soal, menganalisis instrumen, maupun menyusun laporan penilaian. Hasil penelitian ini juga memberikan informasi bahwa sedikit sekali guru yang merencanakan penyusunan tesnya dengan terlebih dahulu menyusun kisi-kisi tes.

Dalam hal analisis instrumen tes hasil belajar pada ujian semester atau akhir semester, umumnya para guru lebih banyak membuat tes dalam bentuk pilihan ganda (*multiple choice*) Berdasarkan kajian literatur beberapa hasil penelitian terhadap kemampuan guru dalam merencanakan penilaian hasil belajar masih menunjukkan hasil yang rendah. Salah satu komponen penting dalam melaksanakan evaluasi hasil belajar adalah analisis kualitas tes.

Sebagaimana yang telah disampaikan di awal bahwa, tes yang baik haruslah memenuhi beberapa kriteria. Dalam buku pedoman Puspendik Balitbang Depdiknas, (2007) sebagaimana yang dikutip Wardhani & Putra, (2016) kriteria tes yang baik setidaknya memenuhi tahapan analisis secara kuantitatif maupun kualitatif. Menurut Surapranata, (2009) dalam Pratama, (2019) analisis kualitatif dapat dilakukan dengan memeriksa aspek teknik penulisan, penggunaan bahasa, maupun kecocokan materi. Sedangkan analisis kuantitatif dapat dilihat dari karakteristik internal tes yang diperoleh dari hasil pengukuran empiris terhadap peserta tes. Selain itu, menurut Azwar, (2009) sebagaimana dikutip Pratama, (2019) analisis tes secara kualitatif dapat dilakukan dengan mengetahui seberapa akurat tes dapat menjangkau kawasan atau materi yang diukur. Untuk analisis kuantitatif baiknya tes divalidasi terlebih dahulu sebelum digunakan.

Dalam bidang pengukuran pendidikan terdapat dua pendekatan yang sering digunakan untuk melakukan analisis kualitas tes, yakni teori tes klasik (*Classical Test Theory*) dan teori respon butir (*Item Response Theory*) Akan tetapi, analisis kualitas tes menggunakan pendekatan teori klasik sudah banyak ditinggalkan karena memiliki banyak kelemahan. Menurut Mardapi, (2012) teori tes klasik setidaknya memiliki dua kelemahan yakni, 1) hasil dari pengukuran akan tergantung pada karakteristik tes yang dipakai, 2) parameter item bergantung pada kemampuan peserta tes, dan 3) *error measurement* atau kesalahan pengukuran hanya dapat mengetahui untuk kelompok, bukan individu. Terkait dengan lemahnya teori tes klasik, Wardhani & Putra, (2016) menyatakan bahwa teori tes klasik kurang mampu menggambarkan atau merefleksikan kemampuan peserta tes yang sebenarnya. Asumsi ini didasarkan pada kemampuan siswa yang hanya dilihat dari perolehan skor total dengan tidak memperhatikan korelasi antara kemamampuan peserta tes dengan karakteristik butir.

Berbeda dengan pendekatan *Item Response Theory* (IRT) yang memiliki asumsi bahwa peluang (*probability*) peserta tes menjawab benar setiap butir tergantung dengan kemampuan peserta tes. Sehingga, peserta tes yang memiliki kemampuan tinggi memiliki peluang menjawab benar lebih besar dibandingkan peserta tes dengan kemampuan rendah (Retnawati, 2014) Hambleton & Swaminathan (1985) dan Hambleton, Swaminathan, & Rogers (1991) dalam Retnawati, (2014) menyatakan ada tiga asumsi yang menjadi dasar IRT yaitu, unidimensi, independensi lokal, dan invariani parameter.

Menurut Hambleton et.al, (1991) dalam Sarea&Ruslan, (2019) unidimensi berarti setiap butir tes hanya mengukur satu kemampuan. Misalnya tes hasil belajar Akidah Akhlak. Ini berarti tes yang hanya dapat mengetahui kemampuan peserta tes pada mata pelajaran Akidah Akhlak saja bukan kemampuan lainnya. Akan tetapi, pada praktiknya asumsi ini sulit dilakukan dikarenakan beberapa faktor yang mempengaruhi seperti, faktor kognitif, kepribadian, lingkungan, atau kecemasan. Asumsi independensi lokal dalam hal ini dinyatakan bahwa tidak ada hubungan antara respon peserta tes dengan butir soal yang berbeda. Sedangkan invariansi parameter menyatakan bahwa karakteristik butir soal tidak tergantung pada sebaran parameter peserta tes dan parameter yang menjadi karakteristik peserta tes tidak tergantung pada karakteristik butir (Retnawati, 2014)

Hambleton, R.K.,&Jones, R.W, (1993) dalam Andayani & Ramalis, (2019) menyebutkan beberapa kelebihan dari IRT yaitu; 1) skor menggambarkan kemampuan peserta tes dan tidak tergantung pada kesulitan tes, 2) dapat digunakan untuk menghubungkan item soal dengan kemampuan peserta tes, 3) tidak membutuhkan tes secara pararel untuk menentukan koefesien reliabilitas. Menurut Mardapi, (2012) salah satu model sederhana dan telah banyak digunakan ahli dalam mengembangkan suatu tes adalah model Rasch, dengan 1 parameter (1-P) Selain itu pemilihan model Rasch karena model ini setidaknya telah memenuhi prinsip-prinsip model pengukuran yaitu; 1) model ini mampu memberikan ukuran yang linier dengan interval yang sama, 2) mampu mengatasi persoalan data yang hilang, 3) dapat memberikan estimasi yang lebih tepat, 4) dapat mendeteksi ketidaktepatan sebuah model,

Tarbawy : Jurnal Pendidikan Islam
ISSN : 2407-4462 (Cetak), 2614-5812 (Elektronik)
Vol. 7, No. 1, 2020, Hal. 61-70
DOI: https://doi.org/ 10.32923/tarbawy.v7i1.1187

dan 5) memberikan instumen pengukuran yang independen dari parameter yang diteliti (Sumintono, B. & Widhiarso, W, 2014) dalam Purnomo, (2016)

Lebih lanjut, Mardapi, (2012) mengatakan bahwa, pengukuran pada model Rasch merupakan pembandingan langsung antara individu dengan butir. Dalam hal ini individu adalah kemampuan peserta tes dan butir adalah parameter tingkat kesulitan. Sehingga pada kondisi tertentu misalnya, kemampuan peserta tes naik, maka pelung menjawab benar butir tes menjadi besar. Oleh karena itu, peluang menjawab benar butir tes merujuk pada dua hal, yakni kemampuan peserta tes dan tingkat kesulitan butir. Terkait dengan konsep dasar model Rash hal yang sama juga dinyatakan Sumintono, (2014) bahwa, model Rash merupakan model pengukuran yang menentukan hubungan antara *ability* (kemampuan) peserta tes dengan tingkat kesulitan item tes. Lebih lanjut Sumintono, (2014) memberikan ilustrasi mengenai hubungan tersebut misalnya ada seorang peserta tes mampu menjawab 80% soal dengan benar tentu memiliki kemampuan yang lebih tinggi dari peserta tes yang hanya mampu menjawab 60% soal dengan benar.

Model Rash dalam penerapan analisis instrumen tes memiliki keunggulan dari model lainnya seperti, teori klasik. Menurut Sumintono&Widhiarso, (2014) dalam Ardiyanti, (2016) model Rash mampu melakukan prediksi terhadap data hilang berdasarkan pola respon individu. Hal inilah yang menjadikan model analisis Rash menjadi lebih akurat. Selain itu, model Rash juga dapat menghasilkan skor pengukuran eror standar untuk instrumen yang digunakan. Selain itu, model Rash juga melakukan kalibrasi pada tiga hal yakni, skala pengukuran, peserta tes (*person*), dan butir (*item*) Kalibrasi dimaksudkan untuk menjamin validitas dan reliabilitas hasil pengukuran sehingga tes dapat memberikan informasi yang komprehensif (Ardiyanti, 2016)

Menurut Mardapi, (2012) analisis instrumen tes menggunakan model Rash dapat dilakukan melalui langkah-langkah berikut; 1) menilai *item fit statistic*. Tahap ini merupakan tahap untuk menentukan item-item yang cocok dengan model Rasch. Jika ada item yang tidak cocok dapat disingkirkan. 2) menilai *person fit statistic*. Tahap ini menentukan peserta tes mana saja yang cocok dengan model Rasch. 3) Menentukan item dan peserta tes (*person*) mana yang cocok dengan model Rasch melalui analisis *goodness of fit*.

Penelitian mengenai analisis kualitas tes sudah banyak dilakukan, terutama pada bidang pendidikan. Dari hasil penelusuran, sebagian besar lebih dominan menggunakan pendekatan teori klasik. Sebagaimana telah diuraikan di atas bahwa, pendekatan teori klasik masih terdapat kelemahan. Sehingga pendekatan teori klasik sudah mulai ditinggalkan dan beralih pada pendekatan *item response theory* (IRT) Adapun hasil penelitian menggunakan model Rasch dilakukan oleh Alfarisa & Purnama, (2019) mengenai Analisis Butir Soal Ulangan Akhir Semester Mata Pelajaran Ekonomi SMA Menggunakan Rasch Model. Penelitian ini membuktikan bahwa, dari 40 butir terdapat 39 butir fit dengan model Rasch. Sainuddin, (2018) juga melakukan penelitian mengenai kualitas tes Matematika buatan guru ditinjau dari pendekatan IRT. Penelitian ini membukikan bahwa, kemampuan guru dalam menyusun tes tergolong kurang baik dengan persentase sebesar 48%. Untuk pembelajaran Pendidikan Agama Islam dan Budi Pekerti SD, penelitian yang dilakukan Sarea, (2018) melakukan estimasi pada tingkat kesulitan, daya beda, dan efektifitas pengecoh.

Dari beberapa penelusuran literatur di atas, penelitian mengenai analisis instrumen tes menggunakan model Rasch khusus pada mata pelajaran Akidah Akhlak siswa Madrasah Aliyah masih belum ditemukan. Penelitian kualitas tes buatan guru Madrasah penting dilakukan karena biasanya guru pada mata pelajaran Pendidikan Agama Islam (PAI) cenderung menganggap kurang penting dalam melakukan analisis soal. Hal ini dikarenakan mata pelajaran PAI tergolong mata pelajaran yang relatif mudah bagi siswa. Berdasarkan uraian di atas, maka penelitian ini bertujuan untuk: 1) mendeskripsikan karakteristik butir soal Ujian Akhir Semester Mata Pelajaran Akidah Akhlak kelas XI IPA Madrasah Aliyah Negeri Insan Cendekia Bangka Tengah tahun ajaran 2018/2019 yang meliputi validitas item, tingkat kesukaran item, *item fit*, dan reliabilitas.

Metode

Penelitian ini merupakan penelitian kuantitatif deskriptif dengan tujuan untuk mendapatkan gambaran mengenai kualitas tes melalui karakteristik tes Model Rash. Subjek pada penelitian ini sejumlah 67 pola respon siswa terhadap instrumen tes Akidah Akhlak buatan guru dengan lima alternatif jawaban. Perangkat tes buatan guru ini diambil dari hasil pelaksanaan Ujian Akhir Semester tahun pelajaran 2018/2019 melalui teknik dokumentasi. Analisis data kuantitatif dilakukan melalui pendekatan IRT model Rasch dengan bantuan program QUEST.

Hasil dan Pembahasan

Insrumen tes hasil Belajar Akidah Akhlak memiliki jumlah butir sebanyak 30 butir dengan lima pilihan jawaban. Analisis pola jawaban responden dianalisis menggunakan model Rasch melalui *software* QUEST. Kualitas soal dalam model Rasch dapat diketahui dengan melakukan estimasi pada parameter seperti, validitas, reliabilitas, daya beda, tingkat kesukaran, dan item fit dengan model Rasch.

1. Estimasi Validitas Item

Untuk menguji validitas menggunakan program QUEST sebagaimana yang diungkapkan Setyawarno, (2017) dapat dibandingkan melalui kriteria di bawah ini:

Tabel 1. Kriteria Nilai INFIT MNSQ

Nilai INFIT MNSQ	Keterangan
>1.33	Tidak cocok dengan model
0.77-1.33	Cocok dengan model
<0.77	Tidak cocok dengan model

Hasil analisis Nilai INFIT MNSQ pada program QUEST dapat dilihat pada gambar di bawah ini:

Gambar 1. Rekapitulasi Validitas Item

Item Estimates (Thresholds) In input Order
all on all (N = 66 L = 30 Probability Level= .50)

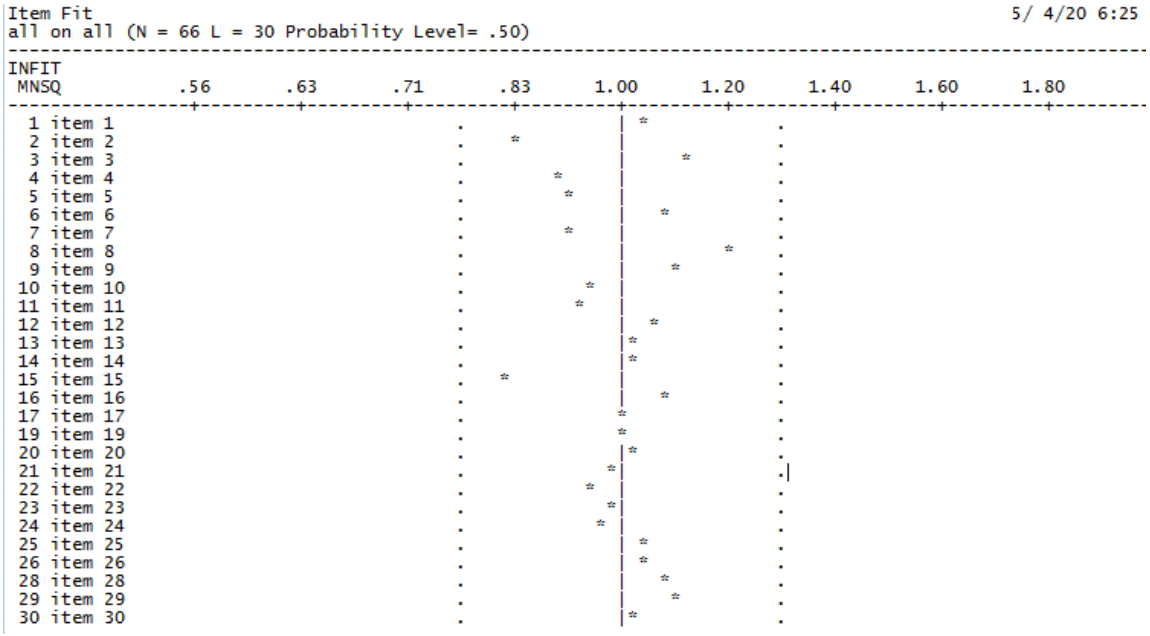
	ITEM NAME	SCORE	MAXSCR	THRSH 1	INFT MNSQ	OUTFT MNSQ	INFT t	OUTFT t
1	item 1	48	66	.28 .29	1.04	.98	.4	.0
2	item 2	52	64	-.22 .33	.83	.67	-.8	-1.2
3	item 3	4	65	4.18 .52	1.11	2.07	.4	1.5
4	item 4	21	65	2.13 .28	.89	.87	-1.0	-.6
5	item 5	61	66	-1.29 .47	.91	.67	-.1	-.5
6	item 6	10	66	3.17 .35	1.09	1.08	.5	.3
7	item 7	33	66	1.34 .26	.90	.87	-1.2	-.8
8	item 8	10	66	3.17 .35	1.21	1.65	.9	1.7
9	item 9	10	66	3.17 .35	1.09	1.11	.5	.4
10	item 10	22	66	2.09 .28	.94	.89	-.6	-.5
11	item 11	44	66	.59 .28	.93	.85	-.6	-.7
12	item 12	26	66	1.80 .27	1.06	1.10	.7	.6
13	item 13	43	66	.66 .27	1.01	1.02	.1	.2
14	item 14	62	66	-1.53 .52	1.02	1.19	.2	.5
15	item 15	59	66	-.92 .41	.82	.47	-.6	-1.4
16	item 16	61	66	-1.29 .47	1.07	1.24	.3	.6
17	item 17	61	66	-1.29 .47	1.00	.77	.1	-.3
18	item 18	0	0	Item has perfect score				

Tarbawy : Jurnal Pendidikan Islam
ISSN : 2407-4462 (Cetak), 2614-5812 (Elektronik)
Vol. 7, No. 1, 2020, Hal. 61-70
DOI: https://doi.org/ 10.32923/tarbawy.v7i1.1187

19	item 19	60	66	-1.09 .44	1.00	1.08	.1	.3
20	item 20	64	66	-2.24 .72	1.01	.88	.2	.2
21	item 21	65	66	-2.94 1.01	.98	.48	.3	-.1
22	item 22	60	66	-1.09 .44	.94	.85	-.1	-.2
23	item 23	19	64	2.28 .29	.98	.98	-.2	.0
24	item 24	64	66	-2.24 .72	.96	.46	.2	-.5
25	item 25	64	66	-2.24 .72	1.04	1.17	.3	.5
26	item 26	64	66	-2.24 .72	1.03	1.01	.3	.3
27	item 27	0	0	Item has zero score				
28	item 28	62	66	-1.53 .52	1.08	1.62	.3	1.1
29	item 29	47	64	.25 .30	1.10	1.11	.8	.5
30	item 30	65	66	-2.94 1.01	1.01	.86	.3	.3
Mean				.00	1.00	1.00	.1	.1
SD				2.09	.09	.35	.5	.7

Gambar 1 di atas memberikan informasi mengenai validitas item dimana semua item fit atau cocok dengan model Rasch dengan rentang nilai INFIT MNSQ antara 0.82 – 1.21. Pada hasil analisis tersebut didapat item 13 bernilai 0 karena semua peserta tes dapat menjawab benar. Dalam pemodelan Rasch, item yang dijawab benar dan salah semua oleh peserta tes tidak dihitung. Sedangkan untuk Item 27 tidak ada satu pun peserta tes yang dapat menjawab benar sehingga tidak dihitung. Untuk mengetahui apakah item cocok dengan Model Rasch juga dapat dilihat melalui peta item fit melalui gambar di bawah ini.

Gambar 2. *Fit Map* Model Rasch

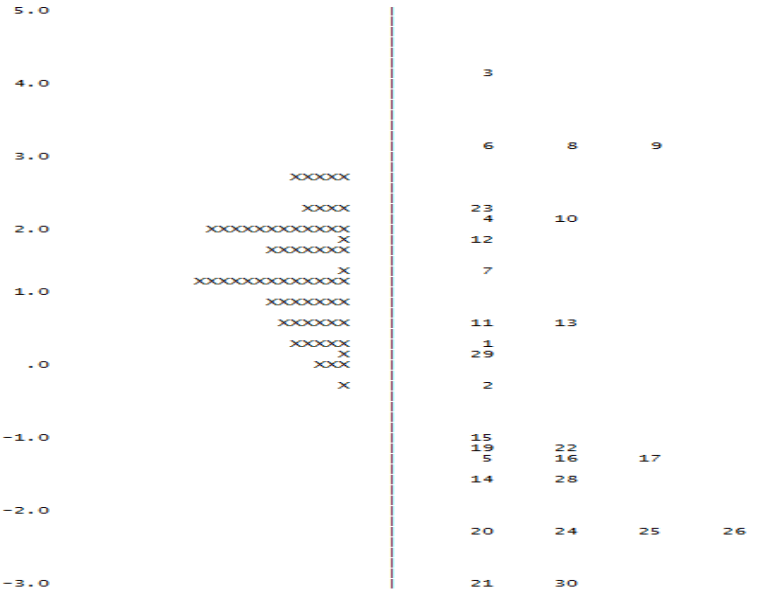


Jika dilihat dari *fit map* model di atas diketahui bahwa seluruh item berada pada rentang nilai INFIT MNSQ 0.77 – 1.30. Tanda titik-titik sebelah kiri menunjukkan nilai 0.77 sedangkan tanda titik-titik sebelah kanan menunjukkan nilai 1.30.

2. Estimasi Tingkat Kesukaran

Untuk mengetahui tingkat kesukaran item melalui program QUEST dapat diketahui dengan melihat hasil analisis *item estimate* (*Threshold*) Adapun kriteria untuk menentukan tingkat kesulitan item berkisar pada nilai -2.0 – 2.0. Jika rentang atau sebaran item atau peserta tes <-2.0, maka item termasuk kategori mudah. Sedangkan jika rentang atau sebaran item atau peserta tes >2.0, maka item termasuk kategori sulit. Agar lebih detail melihat sebaran tingkat kesulitan item dapat dilihat pada gambar di bawah ini.

Gambar 3. Sebaran Tingkat Kesukaran Item



Jika memperhatikan :m yang paling sulit. Jika dibandingkan dengan kemampuan peserta tes sekalipun, kemungkinan peserta tes menjawab benar item nomor 3 sangat kecil atau boleh dikatakan tidak mungkin. Selain itu, untuk item nomor 30 merupakan item paling mudah dan sesuai dengan kemampuan peserta tes. Tingkat kesukaran item melalui program QUEST juga dapat dilihat dari nilai (*threshold*) *item estimate* dengan kriteria yaitu;

b>2 sangat sulit
1<b≤2 sulit

Tarbawy : Jurnal Pendidikan Islam
ISSN : 2407-4462 (Cetak), 2614-5812 (Elektronik)
Vol. 7, No. 1, 2020, Hal. 61-70
DOI: https://doi.org/ 10.32923/tarbawy.v7i1.1187

-1≤b≤1 sedang
-1 ≤b≥-2 mudah
b<2 sangat mudah.

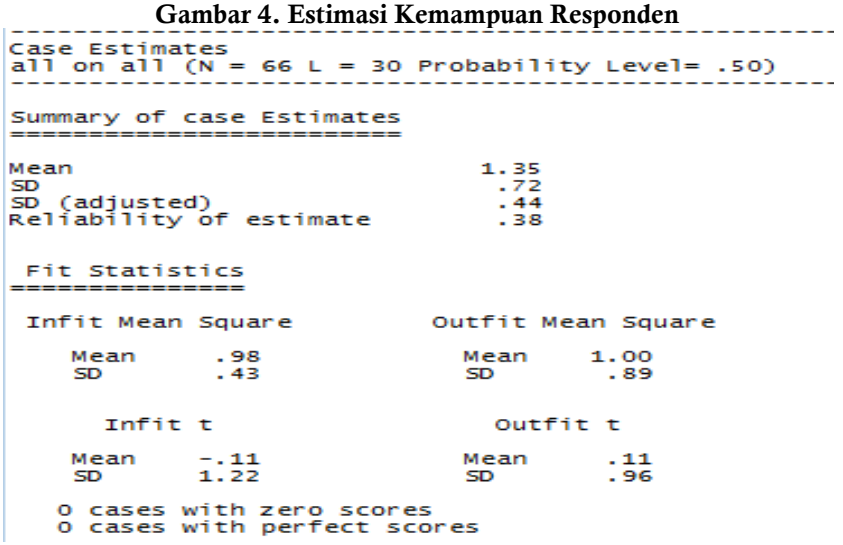
Berikut ini disajikan rekapitulasi tingkat kesukaran masing-masing item.

Tabel 2. Rekapitulasi Tingkat Kesukaran Soal Model Rasch

Item	Nilai *Threshold*	Interpretasi	Item	Nilai *Threshold*	Interpretasi
1	0.28	Sulit	16	-1.29	Mudah
2	-0.22	Sedang	17	-1.29	Mudah
3	4.18	Sangat Sulit	18	0.00	-
4	2.13	Sangat Sulit	19	-1.09	Mudah
5	-1.29	Mudah	20	-2.24	Mudah
6	3.17	Sangat Sulit	21	-2.94	Mudah
7	1.34	Sulit	22	-1.09	Mudah
8	3.17	Sangat Sulit	23	2.28	Sangat Sulit
9	3.17	Sangat Sulit	24	-2.24	Mudah
10	2.09	Sangat Sulit	25	-2.24	Mudah
11	0.59	Sulit	26	-2.24	Mudah
12	1.80	Sulit	27	0.00	-
13	0.66	Sulit	28	-1.53	Mudah
14	-1.53	Mudah	29	0.25	Sulit
15	-0.92	Sedang	30	-2.94	Mudah

Tingkat kesukaran berdasarkan tabel 2 di atas dapat digambarkan bahwa, item soal dengan kategori sangat sulit sebanyak 7 item atau sebesar 25%. Item dengan kategori sulit sebanyak 6 item atau 21.4%, kategori item sedang sebanyak 2 item atau sebesar 7.14%, kategori mudah sebanyak 13 item atau sebesar 46.4%, dan 0% untuk kategori item soal sangat mudah. Secara umum kemampuan peserta tes di bawah tingkat kesulitan item. Hal ini dibuktikan dengan sedikitnya peserta tes yang mampu menjawab benar item dengan sangat sulit atau sulit. Untuk mengetahui kemampuan peserta tes melalui program QUEST dapat dilihat pada *Summary of Case Estimate* pada *reliability of estimate* dengan kriteria, jika nilai Estimate > 1.00 kategori kemampuan tinggi, -1.00 – 1.00 kemampuan sedang, dan < -1.00 kemampuan rendah.

Gambar 4. Estimasi Kemampuan Responden



```
Case Estimates
all on all (N = 66 L = 30 Probability Level= .50)

Summary of case Estimates
=========================

Mean                              1.35
SD                                 .72
SD (adjusted)                      .44
Reliability of estimate            .38

 Fit Statistics
===============

 Infit Mean Square          Outfit Mean Square

    Mean      .98              Mean     1.00
    SD        .43              SD        .89

      Infit t                    Outfit t

    Mean     -.11              Mean      .11
    SD       1.22              SD        .96

   0 cases with zero scores
   0 cases with perfect scores
```

Gambar 4 di atas memberikan keterangan bahwa peserta tes memiliki kemampuan sedang, dengan nilai *reliability estimate* sebesar 0.38 atau dengan rentang -1.00 – 1.00.

3. Estimasi Item Lolos (*Fit*)

Untuk mengetahui item mana saja yang gugur atau lolos didasarkan pada nilai OUTFIT t pada program QUEST. Jika nilai OUTFIT t ≤ 2.00 maka item lolos dan jika nilai OUTFIT t ≥ 2.00 item gugur.

Tabel 3. Rekapitulasi Item *Fit*

Item	Nilai OUTFIT t	Keterangan	Item	Nilai OUTFIT t	Keterangan
1	0.0	Lolos	16	0.6	Lolos
2	-1.2	Lolos	17	-0.3	Lolos
3	1.5	Lolos	18	-	-
4	-0.6	Lolos	19	0.3	Lolos
5	-0.5	Lolos	20	0.2	Lolos
6	0.3	Lolos	21	-0.1	Lolos
7	-0.8	Lolos	22	-0.2	Lolos
8	1.7	Lolos	23	0.0	Lolos
9	0.4	Lolos	24	-0.5	Lolos
10	-0.5	Lolos	25	0.5	Lolos
11	-0.7	Lolos	26	0.3	Lolos
12	0.6	Lolos	27	-	-
13	0.2	Lolos	28	1.1	Lolos
14	0.5	Lolos	29	0.5	Lolos
15	-1.4	Lolos	30	0.3	Lolos

Berdasarkan tabel 3 di atas, diketahui bahwa semua item lolos sehingga dapat disimpulkan bahwa semua item dapat digunakan. Walaupun demikian, item dengan tingkat tingkat kesulitan paling sulit tinggi dan paling mudah sebaiknya tidak diikutsertakan dalam tes. Karena berdasarkan kemampuan peserta tes, sangat sedikit peserta tes mampu menjawab benar item paling sulit. Walaupun demikian, perlu diperhatikan item soal yang sangat sulit. Pada analisis ini terdapat 46.4% item soal sulit. Baiknya proporsi item sulit ini dikurangi untuk mengimbangi peserta tes. Apalagi berdasarkan hasil analisis, peserta tes masuk pada kategori kemampuan sedang.

4. Estimasi Reliabilitas

Nilai reliabilitas model Rasch menggunakan program QUEST dilihat pada *reliability of item estimate* dan *reliability of case estimate*. Pada nilai *reliability of item estimate* sebesar 0.94. Dalam pemodelan Rasch, reliabilitas ini disebut sebagai reliabilitas sampel. Kriteria nilai reliabilitas model Rasch sebagaimana pendapat Perdana, (2018) sebagai berikut; <0.67 lemah, 0.67-0.80 cukup, 0.81 – 0.90 baik, 0.91 – 0.94 baik sekali, >0.94 sempurna. Nilai *reliability of item estimate* sebesar 0.94 ini berhubungan dengan banyaknya item yang cocok dengan model. Nilai 0.94 termasuk reliabilitas dengan kategori baik sekali sehingga berpengaruh pada item yang fit dengan model. Semakin tinggi reliabilitas, maka semakin banyak pula item fit dengan model. Sedangkan nilai *reliability of case estimate* atau reliabilitas peserta tes sebesar 0.38 tergolong lemah. Nilai ini menunjukkan bahwa adanya inkonsistensi sebagaimana yang diungkapkan Ardiyanti, (2017) pada jawaban peserta tes. Inkonsistensi jawaban peserta tes ini juga dapat berarti peserta tes asal-asalan dalam menjawab soal sehingga mempengaruhi nilai reliabilitas person/subjek menjadi rendah.

Jika nilai *reliability of case estimate* dengan kategori baik maka, jawaban peserta tes menunjukkan konsistensinya. Rendahnya nilai *reliability of case estimate* ini dipengaruhi oleh jumlah peserta tes. Dalam hal ini, jawaban peserta tes kurang dari 100 yakni sebanyak 67. Temuan ini sejalan dengan penelitian yang dilakukan oleh Kurniawan, (2016) yang membuktikan bahwa, tinggi rendahnya nilai *reliability of case estimate* ditentukan oleh jumlah peserta tes. Semakin banyak jumlah peserta tes, maka nilai reliabilitasnya makin tinggi.

Penelitian yang dilakukan Purba, (2018) mengenai analisis instrumen tes prestasi menggunakan model Rasch membuktikan bahwa, jumlah peserta tes < 100 berpengaruh terhadap nilai reliabilitas peserta tes. Dalam hal ini, jumlah peserta tes pada penelitian Purba, (2018) sebanyak 428 siswa dengan nilai reliabilitas sebesar 0.92. Selain itu, berdasarkan temuan penelitian tersebut dapat dipahami juga bahwa jumlah item soal tidak mempengaruhi nilai reliabilitas peserta tes. Hal ini sejalan dengan penelitian yang dilakukan Istiyono, Mardapi, & Suparno, (2014) dimana jumlah item yang dianalisis menggunakan model Rasch sebanyak 26 item dan jumlah peserta tes sebanyak 1001 siswa. Penelitian Hakiki, Fitri, & Agung, (2018) juga menunjukkan hasil yang sama bahwa, jumlah item tidak berpengaruh terhadap nilai reliabilitas tes. Jumlah item pada penelitian ini sebanyak 20 item dengan jumlah responden sebanyak 293.

Tarbawy : Jurnal Pendidikan Islam
ISSN : 2407-4462 (Cetak), 2614-5812 (Elektronik)
Vol. 7, No. 1, 2020, Hal. 61-70
DOI: https://doi.org/ 10.32923/tarbawy.v7i1.1187

Kesimpulan

Berdasarkan hasil analisis Tes Akidah Akhlak Buatan Guru dapat digambarkan beberapa karakteristik tes dan peserta tes sebagai berikut; 1) estimasi validitas item fit atau cocok dengan model Rasch untuk 28 item dengan rentang nilai INFIT MNSQ antara 0.82 – 1.21. 2) 2 item tidak dianalisis karena pola jawaban benar dan salah untuk seluruh peserta tes. Estimasi tingkat kesukaran item soal dengan kategori sangat sulit sebanyak 7 item atau sebesar 25%. Item dengan kategori sulit sebanyak 6 item atau 21.4%, kategori item sedang sebanyak 2 item atau sebesar 7.14%, kategori mudah sebanyak 13 item atau sebesar 46.4%, dan 0% untuk kategori item soal sangat mudah. Secara umum kemampuan peserta tes di bawah tingkat kesulitan item. 3) Semua item pada tes dapat digunakan berdasarkan hasil estimasi nilai OUTFIT t ≤ 2.00. 4) Nilai *reliability of item estimate* sebesar 0.94 dengan kategori baik sekali dan nilai *reliability of case estimate* sebesar 0.38 dengan kategori lemah. Berdasarkan hasil analisis menggunakan model Rasch maka, secara umum istrumen tes Akidah Akhlak buatan guru ini dapat digunakan. Akan tetapi, kurang tepat jika hasil pengukurannya digunakan untuk pengambilan keputusan berdasarkan kemampuan siswa.

Referensi

Alfarisa, F., & Purnama, D. N. (2019). Analisis Butir Soal Ulangan Akhir Semester Mata Pelajaran Ekonomi SMA Menggunakan RASCH Model. *Jurnal Pendidikan Ekonomi Undiksha*, *11*(2), 366–374.

Andayani, A., & Ramalis, T. R. (2019). Kajian implementasi teori respon butir dalam menganalisis instrumen tes materi fisika. *Seminar Nasional Fisika*, *1*(1), 37–42.

Ardiyanti, D. (2016). Aplikasi Model Rasch pada Pengembangan Skala Efikasi Diri dalam Pengambilan Keputusan Karir Siswa. *Jurnal Psikologi*, *43*(3), 248–263.

Ardiyanti, D. (2017). Aplikasi Model Rasch pada Pengembangan Skala Efikasi Diri dalam Pengambilan Keputusan Karir Siswa. *Jurnal Psikologi*, *43*(3), 248–263.

Hakiki, A. W., Fitri, A. R., & Agung, I. M. (2018). Analisis Properti Psikometri Subtes Merkaufgaben (ME) dengan Rasch Model. *Jurnal Psikologi*, *14*(1), 40–49.

Hamid, A. (2016). Implementasi Kompetensi Guru Dalam Evaluasi Pembelajaran Pada Madrasah Aliyah Al-Balad Kamande. *J-Alif: Jurnal Penelitian Hukum Ekonomi Syariah Dan Budaya Islam*, *1*(1), 28–42.

Istiyono, E., Mardapi, D., & Suparno, S. (2014). Pengembangan tes kemampuan berpikir tingkat tinggi fisika (pysthots) peserta didik SMA. *Jurnal Penelitian Dan Evaluasi Pendidikan*, *18*(1), 1–12.

Kartowagiran, B. (2011). Kinerja guru profesional (Guru pasca sertifikasi). *Jurnal Cakrawala Pendidikan*, *3*(3).

Kurniawan, R. (2016). Apakah mahasiswa psikologi islam bahagia? Gambaran psychological well-being dengan pendekatan pemodelan rasch. *Al-Qalb: Jurnal Psikologi Islam*, *7*(1), 65–73.

Mardapi, D. (2012). *Pengukuran Penilaian dan Evaluasi Pendidikan*. Yogyakarta: Yuha Medika.

Nazaruddin, N. (2017). Kemampuan Guru dalam Menyusun Tes Hasil Belajar melalui Workshop di SD Negeri Lamteubee. *Serambi Akademica*, *5*(1), 32–42.

Perdana, S. A. (2018). Analisis Kualitas Instrumen Pengukuran Pemahaman Konsep Persamaan Kuadrat Melalui Teori Tes Klasik Dan Rasch Model. *Jurnal Kiprah*, *6*(1), 41–48.

Prasetya, T. I. (2012). Meningkatkan Keterampilan Menyusun Instrumen Hasil Belajar Berbasis Modul Interaktif Bagi Guru-Guru IPA SMP N kota Magelang. *Journal of Educational Research and Evaluation*, *1*(2).

Pratama, D. (2019). Analysis Of Clasical Test Theory (Ctt) Approch On Academic Ability Test Instrument. *JISAE (Journal of Indonesian Student Assesment and Evaluation)*, *5*(2), 43–54.

Purba, S. E. D. (2018). Analisis model Rasch instrumen tes prestasi pada mata pelajaran dasar dan pengukuran listrik. *Wiyata Dharma: Jurnal Penelitian Dan Evaluasi Pendidikan*, *6*(2), 142–147.

Purnomo, S. (2016). *Pengembangan soal matematika model PISA konten space and shape untuk mengetahui level kemampuan berpikir tingkat tinggi berdasarkan analisis model rasch*.

Retnawati, H. (2014). Teori respons butir dan penerapannya: Untuk peneliti, praktisi pengukuran dan pengujian, mahasiswa pascasarjana. *Yogyakarta: Nuha Medika*.

Sainuddin, S. (2018). Analisis Karakteristik Butir Tes Matematika Pada Tes Buatan Mgmp Matematika Kota Makassar Berdasarkan Teori Moderen (Teori Respon Butir). *Proximal: Jurnal Penelitian Matematika Dan Pendidikan Matematika*, *1*(1).

Sarea, M. S. (2018). Karakteristik Soal Ujian Akhir Semester Pendidikan Agama Islam Dan Budi Pekerti Tingkat Sekolah Dasar. *An-Nahdhah*, *11*(2), 303–318.

Sarea, M. S., & Ruslan, R. (2019). Karakteristik Butir Soal: Classical Test Theory Vs Item Response Theory? *Didaktika: Jurnal Kependidikan*, *13*(1), 1–16.

Setiadi, H. (2016). Pelaksanaan penilaian pada Kurikulum 2013. *Jurnal Penelitian Dan Evaluasi Pendidikan*, *20*(2), 166–178.

Setyawarno, D. (2017). *Upaya peningkatan kualitas butir soal dengan analisis aplikasi quest*. Yogyakarta: Fakultas Matematika dan Ilmu Pengetahuan Alam Universitas Negeri Yogyakarta.

Sumintono, B. (2014). *Model Rasch untuk penelitian sosial kuantitatif*.

Wardhani, D. F., & Putra, A. P. (2016). Pengembangan Instrumen Tes Standar Kognitif pada Mata Pelajaran IPA Kelas 7 SMP Di Kabupaten Banjar. *Proceeding Biology Education Conference: Biology, Science, Enviromental, and Learning*, *13*(1), 75–82.

Wiguna, S. (2018). *Kemampuan Guru PAI Dalam Merancang Tes (Analisis Aplikasi Anates Ganda Di Sekolah SMA Negeri 1 Hinai.* Universitas Islam Negeri Sumatera Utara.