MENEMUKAN KARAKTERISTIK BUTIR MENGGUNAKAN QUEST

Suparman

Dosen Program Studi Pendidikan Agama Islam

STAI Masjid Syuhada Yogyakarta

tsabita.bit@gmail.com

Abstrak

Penelitian ini bertujuan untuk menemukan karakteristik item evaluasi pada mata pelajaran Pendidikan Agama Islam. Data utama dalam penelitian ini adalah lembar jawaban peserta didik terhadap item soal evaluasi yang diselenggarakan oleh Dinas Pendidikan Nasional Kabupaten Sleman pada tanggal 14 Maret 2020. Data dikumpulkan pada tanggal 21 Maret 2020 dengan metode dokumentasi di Sekolah Dasar Islam Terpadu Bakti Insani Kabupaten Sleman. Data yang digunakan untuk analisa seluruh peserta didik kelas VI yang berjumlah 56 peserta didik dengan jumlah soal sebanyak 40 item. Analisa data dengan menggunakan Program QUEST untuk mendapatkan karakteristik dari item soal evaluasi. Hasil penelitian menunjukkan bahwa sebanyak 5 (lima) item *terjadi item has perfect score*, sebagian besar item memiliki tingkat kesukaran "Sedang". Sebanyak 48,5% item memiliki daya beda dengan kategori "sangat Baik". Sebanyak 88,5% item merupakan item yang sesuai untuk mengevaluasi kemampuan peserta didik pada mata pelajaran Pendidikan Agama Islam. Estimasi reliabilitas sebesar 0,75, dan jumlah distractor yang efektif sebanyak 50%.

Kata Kunci : Karakteristik Butir, Evaluasi, Quest.

Abstract

This study aims to find the characteristics of evaluation items in Islamic Religious Education subjects. The main data in this study are the students' answer sheets to the evaluation item items held by the Sleman District National Education Office on March 14, 2020. Data was collected on March 21, 2020 with the method of documentation in the Integrated Islamic Elementary School Bakti Insani, Sleman Regency. The data used for the analysis of all grade VI students, amounting to 56 students with a total of 40 items. Analyze data using the QUEST Program to get the characteristics of the evaluation items. The results showed that as many as 5 (five) items occurred item has perfect score, most items had a level of "Medium" difficulty. As many as 48.5% items have different power with the category of "very good". As many as 88.5% items are items that are infit for evaluating the ability of students in Islamic

Religious Education subjects. The estimated reliability is 0.75, and the number of effective distractors is 50%.

Keywords : Item Characteristic, Evaluation, Quest.

A. LATAR BELAKANG

Proses pendidikan secara tradisional digambarkan dalam bentuk In Put – Proses – Out Put. Dikatakan tradisional karena pada tahap proses digambarkan sebagai kotak hitam yang tidak dapat diamati. Hal ini memberikan pengertian destruktif, karena proses pendidikan menjadi suatu peristiwa yang absurd dan tidak dapat diterjemahkan secara tegas. Akibatnya sulit untuk mendeteksi penyebab terjadi hasil belajar yang di bawah standar. Terjadinya *overleaping* pada proses pengambilan kebijakan pada bidang pendidikan dapat terjadi karena penyebab yang tidak akurat sebagai dampak dari sulitnya mendeteksi pada tahap “proses”.

Penulis berpendapat bahwa proses pendidikan diawali dengan Seleksi Peserta didik – Pembelajaran – Hasil belajar. Berdasarkan tahapan ini, setiap proses akan dapat dievaluasi secara transparan. Apabila terjadi kekurangan pada setiap proses dapat diminimalisir, karena dapat dideteksi secara baik. Peserta didik dibantu oleh pendidik akan mendapatkan hasil belajar secara maksimal. Penyebab timbulnya masalah belajar akan dapat teridentifikasi sehingga dapat digunakan sebagai dasar untuk membuat kebijakan pada pendidikan secara *in line*.

Salah satu tahap pembelajaran adalah evaluasi. Pada tulisan ini akan dibahas tentang butir soal yang digunakan untuk evaluasi. Bahasan ini secara obyektif akan ditunjukkan butir soal yang sebaiknya dipakai dan butir soal yang sebaiknya tidak digunakan. Pada butir yang layak dipakai, sebaiknya dilakukan

dublikasi dan vareasi untuk menyiaapkan butir soal yang lebih berkualitas. Adapun butir yang tidak layak digunakan, sebaiknya didrop (angulir) agar tidak menurunkan kualitas butir valuasi.

B. KARAKTERISTIK BUTIR TES

Tes merupakan salah satu cara untuk menentukan tingkat kemampuan (*ability*) peserta didik. Menurut Arikunto tes merupakan alat atau prosedur yang digunakan untuk mengetahui atau mengukur sesuatu dalam suasana, dengan cara dan aturan-aturan yang sudah ditentukan¹³². Adapun Sudijono berpendapat bahwa tes adalah cara (yang dapat dipergunakan) atau prosedur (yang perlu ditempuh) dalam rangka pengukuran dan penilaian di bidang pendidikan, yang berbentuk pemberian tugas atau serangkaian tugas baik berupa pertanyaan-pertanyaan (yang harus dijawab), atau perintah-perintah (yang harus dikerjakan) oleh *testee*, sehingga (atas dasar data yang diperoleh dari hasil pengukuran tersebut) dapat dihasilkan nilai yang melambangkan tingkah laku atau prestasi *testee*; nilai mana dapat dibandingkan dengan nilai-nilai yang dicapai oleh *testee* lainnya atau dibandingkan dengan nilai standar tertentu¹³³. Berdasarkan pendapat di atas dapat dipahami bahwa tes merupakan Prosedur untuk menentukan kemampuan seseorang/peserta didik melalui perintah-perintah tertentu dimana yang dinilai (*testee*) menunjukkan tingkah laku yang menggambarkan tingkat kemampuannya.

¹³² Suharsimi Arikunto. *Dasar-Dasar Evaluasi Pendidikan (Edisi Revisi)*. Bumi Aksara : Jakarta. 2006. hlm. 53.

¹³³ Sudijono, A. *Pengantar Evaluasi Pendidikan*. PT RajaGrafindo Persada : Jakarta. 2011. hlm. 67.

Karakteristik butir tes dapat diketahui dari dua pendekatan, yaitu teori tes klasik dan teori respon butir. Allen & Yen menegaskan bahwa teori tes klasik, atau disebut juga teori tes skor murni klasik, didasarkan pada model aditif, yaitu skor amatan merupakan penjumlahan dari skor sebenarnya dan skor kesalahan pengukuran¹³⁴. Formulasi dasar dari teori klasik adalah $X = T + E$, dimana X adalah skor amatan yang diperoleh, T adalah skor murni, dan E adalah tingkat kesalahan yang terjadi selama penilaian berlangsung. Selanjutnya Hambleton, Swaminathan, & Rogers menjelaskan bahwa teori respon butir atau teori tes modern dikembangkan atas dasar pikiran: 1) hasil ujian seseorang dapat diprediksi dari kemampuan yang dimilikinya dan 2) hubungan antara hasil ujian dan kemampuan dinyatakan dalam sebuah fungsi yang disebut kurva karakteristik butir atau *Item Characteristic Curve*¹³⁵.

Terdapat dua asumsi yang mendasari teori respon butir menurut Hambleton, Swaminathan, & Rogers adalah 1) Asumsi pertama adalah unidimensional, yakni suatu asumsi yang menyatakan bahwa setiap butir hanya mengukur satu dimensi. Asumsi ini sulit untuk dipenuhi, akan tetapi setidaknya butir tersebut secara dominan dapat untuk mengukur satu kemampuan tertentu. Menurut Hattie (1985) dan Sinaga (1992) cara untuk menentukan unidimensi adalah dengan analisa faktor. 2) Asumsi kedua adalah kebebasan lokal (*local independence*), yaitu suatu anggapan bahwa respon peserta didik untuk menjawab butir tertentu tidak dipengaruhi oleh butir yang lain.

¹³⁴ Allen dan Yen. *Introduction to Measurement Theory*. Brooks/Cole Publishing Company Wadsworth Inc. : California.1979. hlm. 57.

¹³⁵ Hambleton, R.K., Swaminathan, H., & Rogers, H.J. *Fundamental of item response theory*. Sage Publication Inc : Newbury Park, CA .1991. hlm. 2-5.

Program QUEST merupakan salah satu software yang digunakan untuk analisa butir soal dengan menggunakan pendekatan teori respon butir. Adapun karakteristik yang dapat dihasilkan dengan menggunakan program QUEST adalah 1) estimasi tingkat kesukaran, 2) daya beda, 3) kecocokan butir, 4) estimasi reliabilitas, dan 5) efektifitas distraktor. Setiap karakteristik dapat dijelaskan sebagai berikut :

1. Tingkat kesukaran

Tingkat kesukaran merupakan karakteristik yang dapat digunakan untuk menentukan kualitas butir tersebut. Biasanya tingkat kesukaran diberi dengan simbol p . Pada dasarnya tingkat kesukaran menunjukkan berapakah peserta didik yang dapat menjawab benar dari semua peserta didik yang mengerjakan butir soal tersebut. Besarnya koefisien tingkat kesukaran berkisar antara 0,00 sampai dengan 1,00. Nilai tingkat kesukaran semakin mendekati 0,00 maka butir tersebut semakin sulit dan semakin mendekati 1,00 maka butir tersebut semakin mudah.

Allen dan Yen (1979 : 122) menyatakan bahwa secara umum indeks kesukaran suatu butir sebaiknya terletak pada interval (0,3 – 0,7)¹³⁶. Besar koefisien pada rentang tersebut dapat memberikan informasi tentang kemampuan peserta didik secara maksimal. Oleh karena itu butir soal tersebut layak digunakan sebagai butir evaluasi. Pada butir pilihan ganda untuk menentukan koefisien tingkat kesukaran dapat ditentukan dengan¹³⁷ :

$$p = \frac{n_i}{N}$$

¹³⁶ Allen & Yen. Op. cite. hlm. 122.

¹³⁷ Suparman. *Evaluasi Pendidikan*. STAIMS Press : Yogyakarta. 2017. hlm. 21.

Keterangan :

p : Koefisien indeks kesulitan

n_i : Jumlah peserta didik yang menjawab dengan benar pada butir no tertentu (i)

N : Jumlah peserta didik yang mengerjakan soal pada butir tertentu.

i : 1, 2, 3, ... dst.

Perlu dipertimbangkan besarnya koefisien tingkat kesukaran butir soal dengan tujuan tes. Penulis berpendapat apabila tes digunakan untuk tujuan evaluasi belajar maka butir soal tingkat kesukaran dengan proporsi 15% mudah, 70% sedang dan 15% sulit. Adapun tes yang digunakan untuk seleksi maka proporsinya 5% mudah, 5% sedang dan 90% sulit. Tes yang digunakan untuk tujuan diagnostik maka proporsi koefisien tingkat kesukarannya 80% mudah, 10% sedang dan 10% sulit. Interval mengenai tingkat kesukaran dapat dilihat pada tabel 1.

Tabel 1 : Interval Tingkat Kesukaran¹³⁸

No	Interval	Kategori
1	0,70 – 1,00	Mudah
2	0,30 – 0,70	Sedang
3	0,00 – 0,30	Sulit

2. Daya Beda

Pada dasarnya daya beda merupakan kemampuan butir soal untuk membedakan antara peserta didik yang berkemampuan tinggi dengan peserta didik yang berkemampuan rendah. Oleh karena itu soal yang baik adalah soal yang dapat membedakan antara peserta didik yang berkemampuan tinggi dengan peserta didik yang berkemampuan rendah. Menentukan indeks daya beda dapat menggunakan indeks korelasi biserial, indeks korelasi point biserial, dan indeks keselarasan. Apabila menggunakan

¹³⁸ Suharsimi Arikunto. *Op. cite.* hlm. 210

indeks korelasi point biserial maka menurut Allen & Yen rumusnya adalah¹³⁹ :

$$r_{ix} = \frac{\bar{X}_i - \bar{X}}{s_x} \sqrt{\frac{p_i}{1 - p_i}}$$

Keterangan :

- r_{ix} = Koefisien daya beda
- $\bar{X}_i$ = Rerata skor pada butir i
- $\bar{X}$ = Rerata skor
- s_x = Varians sekor
- p_i = Daya beda

Berdasarkan penjelasan dari Cecil dkk, indeks daya beda dapat dikategorikan sebagaimana terdapat pada tabel 2.

Tabel 2 : Indeks Daya Beda¹⁴⁰

No	Interval	Kategori
1	$\geq 0,40$	Sangat Baik
2	0,30 – 0,39	Baik
3	0,11 – 0,29	Sedang
4	0,00 – 0,10	Kurang

3. Kecocokan Butir

Kecocokan butir (item fit) merupakan informasi untuk menentukan apakah butir tersebut merupakan butir yang sesuai atau tidak sesuai untuk mengukur kemampuan peserta didik. Butir soal yang tidak fit dapat dieliminir supaya tidak mempengaruhi kualitas evaluasi. Adapun butir yang fit dapat digunakan dan dikembangkan untuk evaluasi lebih lanjut.

¹³⁹ Allen & Yen. *Op. cit.* hlm. 122

¹⁴⁰ Cecil R. R; Ronald B. L; and Victor W. *Measurement and Assessment in Education* (Second Edition). Pearson Education, Inc. : New Jersey. 2009. hlm. 152

Penetapan *fit item* secara keseluruhan dengan model dalam program QUEST ditentukan berdasarkan pada besarnya nilai rerata *INFIT Mean of Square (INFIT MNSQ)* dan besar simpangan baku atau nilai rerata *INFIT Mean of INFIT t*. Penetapan fit tiap butir pada program QUEST ditentukan berdasarkan besarnya nilai *INFIT MNSQ* atau *nilai INFIT t* item yang bersangkutan. Besarnya kuadrat tengah yang tidak tertimbang (*Unweighted Mean Square*) dalam program QUEST disingkat *OUTFIT MNSQ* maupun kuadrat tengah yang tertimbang (*Wighted Mean Square*) yang diharapkan adalah sebesar 1 (satu) dan varians sebesar 0 (nol)¹⁴¹. Sementara besarnya nilai harapan *Mean INFIT t* sama dengan 0 dengan varians sama dengan 1. Menurut Didik Setyawarno¹⁴² maupun Bambang Subali dan Pujiyati Suyata¹⁴³ item fit (butir yang fit) dapat ditentukan dengan mengacu pada ketentuan yang terdapat pada tabel 3.

Tabel 3 : Ketentuan Kecocokan butir¹⁴⁴

No	Nilai Infit MNSQ	Keterangan
1	> 1,33	Aitem tidak fit
2	0,77 – 1,33	Aitem fit
3	< 0,77	Aitem tidak fit

4. Reliabilitas

Reliabilitas merupakan keajegan alat ukur (butir evaluasi) untuk mengukur kemampuan peserta didik.

¹⁴¹ Didik Setyawarno. Upaya Peningkatan Kualitas Butir Soal Dengan Analisa Aplikasi QUEST. *Makalah PPM bagi guru IPA SMP di Sleman disampaikan pada hari Sabtu, 9 September 2017 di Lab. Komputer IPA. 2017. hlm. 8.*

¹⁴² Didik Setyawarno. *Ibid* hlm. 9

¹⁴³ Bambang Subali & Pujiyati Suyata. Panduan Menggunakan Program Quest. *Lembaga Penelitian UNY. 2011. hlm. 25*

¹⁴⁴ Didik Styawarno. Op. Cit.

Keajegan ini tentunya bermakna positif, yakni alat ukur tersebut menjamin secara konsisten hasil dari tes tersebut. Apabila ada dua atau lebih tes pada kelompok peserta didik yang sama, dengan aitem soal yang memiliki reliabilitas yang baik, maka hasil uji atau hasil tesnya akan relative sama. Hal ini sesuai dengan penjelasan dari Mehrens & Lehmann yang menyatakan bahwa reliabilitas merupakan derajat keajegan (*consistency*) di antara dua buah hasil pengukuran pada objek yang sama¹⁴⁵.

Pendapat lain disampaikan oleh Anastasi & Urbina yang menyatakan bahwa reliabilitas merujuk pada konsistensi skor yang dicapai oleh orang yang sama diuji dengan tes yang sama pada waktu yang berbeda¹⁴⁶. Berdasarkan pendapat tersebut, maka reliabilitas dapat diartikan sebagai suatu alat ukur yang memiliki implikasi bahwa subjek yang dikenai pengukuran memiliki hasil yang relatif sama pada saat dilakukan pengukuran pada waktu yang berbeda dengan alat ukur yang sama.

Program QUEST untuk mengestimasi reliabilitas dengan menggunakan formulasi koefisien alpha atau *cronbach alpha*. Kriteria koefisien reliabilitas dapat dilihat pada tabel 4. Adapun formulasi koefisien alpha adalah¹⁴⁷ :

$$\alpha = \frac{n}{n-1} \left(1 - \frac{\sum v_i}{v_t} \right)$$

Keterangan :

α = Koefisien reliabilitas

¹⁴⁵ Mehrens, W.A. & Lehmann, I.J. *Measurement and evaluation in education and psychology*. Hold, Rinehart and Wiston, Inc : New York. 1973. hlm. 102.

¹⁴⁶ Anastasi, A. & Urbina, S. *Psikologi Testing*. Edisi terjemah oleh Hariono, R & Imam, S. PT Prenhallindo : Jakarta. 1998. hlm. 63.

¹⁴⁷ Cronbach, L.J., Coefficient Alpha and The Internal Structure of Test, *Psychometrika*. 1951, Volume 16. hlm. 297-334.

- n = Banyaknya butir
 1 = Konstanta
 $\sum v_i$ = Jumlah varians butir ke i
 v_t = Varians total

Tabel 4 : Kriteria Koefisien Reliabilitas¹⁴⁸

No	Koefisien Reliabilitas	Klasifikasi
1	0,80 – 1,00	Sangat Tinggi
2	0,60 – 0,80	Tinggi
3	0,40 – 0,60	Sedang
4	0,20 – 0,40	Rendah
5	-1,00 – 0,20	Rendah Sekali

5. Efektivitas Distraktor

Distraktor merupakan pilihan (opsi) pada aitem evaluasi yang berbentuk pilihan ganda. Apabila terdapat opsi 5 maka komposisinya terdiri atas 1 (satu) jawaban kunci dan 4 (empat) distraktor atau pengecoh. Pengecoh yang baik adalah pengecoh yang dipilih oleh sebagian besar peserta didik yang memiliki kemampuan rendah, sedangkan jawaban kunci adalah pilihan yang dipilih oleh sebagian besar peserta didik yang memiliki kemampuan tinggi.

Anas Sudijono menyatakan bahwa cara menganalisa fungsi distraktor dapat dilakukan dengan melihat pola penyebaran jawaban pada aitem soal. Pola penyebaran jawaban menurut Sudijono adalah suatu pola yang dapat menggambarkan bagaimana peserta tes dapat menentukan pilihan jawaban terhadap kemungkinan-kemungkinan jawaban yang telah dipasangkan pada setiap butir/aitem¹⁴⁹.

¹⁴⁸ Guilford, J.P. *Fundamental Statistic in Psychology and Education*. 3rd Ed. McGraw-Hill Book Company, Inc. : New York. 1956. hlm. 145

¹⁴⁹ Anas Sudijono. *Pengantar Evaluasi Pendidikan*. Jakarta : Rajawali Pres. 2009. hlm. 411

Efektivitas distraktor, dapat dilihat dengan dua cara¹⁵⁰, yaitu (1) Melihat berapa banyaknya peserta didik yang memilih distractor. Apabila terdapat distraktor yang tidak dipilih oleh peserta didik, maka distraktor tersebut harus diganti dengan pilihan yang lebih tersamar. (2) Distraktor yang dipilih oleh peserta didik, memiliki koefisien point biserial yang lebih rendah dari kunci jawaban. Distraktor yang memiliki koefisien point biserial yang lebih rendah dari point biserial kunci jawaban, menunjukkan distraktor tersebut berfungsi dengan baik.

C. METODE PENELITIAN

Penelitian ini merupakan jenis penelitian diskriptif kuantitatif, data diambil dari perangkat soal latihan ujian nasional Sekolah Dasar putaran ke dua (2) tahun pelajaran 2019/2020 untuk mata pelajaran Pendidikan Agama Islam (PAI) Kabupaten Sleman, Daerah Istimewa Yogyakarta. Respon peserta didik dianalisa secara empiris berdasarkan pendekatan teori tes klasik dan teori respon butir untuk menguji kelayakan butir soal. Pengambilan data dilakukan pada tanggal 21 Maret 2020.

Subjek penelitian ini adalah peserta didik kelas VI Sekolah Dasar Islam Terpadu Bakti Insani, yang beralamatkan di Jln Letnan Sumanto, Srimulyo, Triharjo, Sleman, Yogyakarta. Jumlah peserta didik sebanyak 56 anak, dengan butir soal PAI sebanyak 40 item, ukuran sampel ini sudah memenuhi persyaratan karena menurut Bond & Fox, ukuran sampel khusus untuk model 1PL berupa Rasch Model antara 30 sampai

¹⁵⁰ Suparman. Panduan Membaca Out Put dari Quest. *Hand Out Materi Kuliah*. Tidak Diterbitkan. hlm. 8

300¹⁵¹. Data yang akan dianalisa adalah jawaban peserta didik terhadap 40 item soal pilihan ganda pada ujian mata pelajaran Pendidikan Agama Islam.

Dokumen dikumpulkan dengan melakukan instrumen dokumen. Disebut instrumen karena sebanyak 56 peserta didik mengerjakan butir soal yang sama, sedangkan disebut dokumen karena lembar jawaban tersebut proses kegunaan utamanya telah selesai dan siap untuk disimpan sebagai arsip apabila suatu saat akan digunakan lagi. Secara teknis, pengambilan data dilakukan sepekan setelah ujian diselenggarakan.

Dokumen yang berupa lembar jawaban dari peserta didik yang telah diperoleh, kemudian ditabulasi dalam bentuk notepad yang berisi respon peserta didik pada setiap item soal PAI sebanyak 40 item. Langkah ini merupakan upaya untuk membuat file data. Setelah file data siap, langkah selanjutnya menyiapkan file program agar respon dari seluruh peserta didik dapat dianalisa menggunakan program QUEST. Setelah file data dan file program selesai disiapkan, selanjutnya dijalankan program QUEST untuk mendapatkan out put.

D. HASIL PENELITIAN

Berdasarkan out put dari program QUEST, dari 40 item soal yang dianalisa menunjukkan bahwa semua item soal dikerjakan oleh peserta didik (*0 items with zero scores*), artinya tidak ada item yang terlewat, kosong atau tidak diisi oleh peserta didik. Terdapat 5 item soal yang dapat dikerjakan oleh seluruh peserta didik (*5 items with perfect scores*) yakni item_1, item_3, item_6, item_20, dan item_29 . Program QUEST tidak

¹⁵¹ Bond, T.G. & Fox, Ch.M. *Applying the rasch model: Fundamental measurement in the human sciences. 2-nd ed.* Mahwah, Lawrence Erlbaum Associates, Publishers : New Jersey. 2007. hlm. 43.

menganalisa lima (5) item soal tersebut agar tidak mempengaruhi akurasi dari item yang lain.

Karakteristik item yang dihasilkan dari program QUEST adalah 1) estimasi tingkat kesukaran, 2) daya beda, 3) kecocokan butir, 4) estimasi reliabilitas, dan 5) efektivitas distraktor. Karakteristik tersebut merupakan besaran yang banyak didiskusikan dalam analisis item. Setiap karakteristik dilaporkan dalam bentuk tabel. Tabel 5 sampai dengan tabel 8 merupakan laporan dari hasil analisa¹⁵².

Berdasarkan Tabel 5 dapat diketahui dari 35 item yang dianalisa, terdapat 3 item soal dengan tingkat kesukaran kategori “MD”, yakni item nomor 36, 39 dan 40. Terdapat 29 item soal dengan tingkat kesukaran kategori “SD”, yakni item nomor 2, 4, 5, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 21, 22, 23, 24, 25, 26, 27, 30, 32, 33, 34, 35, 37, dan 38. Adapun sebanyak 3 item soal dengan tingkat kesukaran kategori “S”, yakni item no 7, 28, dan 31.

Item dengan tingkat kesukaran kategori “MD” mempunyai koefisien antara 0,92 sampai dengan 0,98. Item dengan tingkat kesukaran kategori “SD” mempunyai koefisien antara 0,30 sampai dengan 0,59. Item dengan tingkat kesukaran kategori “S” mempunyai koefisien antara 0,19 sampai dengan 0,28.

Berdasarkan Tabel 6 dapat diketahui dari 35 item yang dianalisa, terdapat 5 item soal dengan indeks daya beda kategori “K”, yakni item nomor 14, 18, 22, 27 dan 31. Terdapat 6 item soal dengan indeks daya beda kategori “S”, yakni item nomor 9, 13, 21, 32, 34 dan 35. Terdapat 7 item soal dengan indeks daya beda kategori “B”, yakni item nomor 4, 15, 24, 26, 28, 30, dan

¹⁵² Hasil analisa dengan menggunakan program quest. Tanggal 1 April 2020. Jam 14.12

33. Adapun sebanyak 17 item soal dengan indeks daya beda kategori “SB”, yakni item no 2, 5, 7, 8, 10, 11, 12, 16, 17, 19, 23, 25, 36, 37, 38, 39, dan 40.

Tabel 5 : Koefisien Tingkat Kesukaran Butir

No	Item	Koef	Ket
1	Item_2	0,30	SD
2	Item_4	0,50	SD
3	Item_5	0,48	SD
4	Item_7	0,28	S
5	Item_8	0,50	SD
6	Item_9	0,56	SD
7	Item_10	0,57	SD
8	Item_11	0,58	SD
9	Item_12	0,59	SD
10	Item_13	0,54	SD
11	Item_14	0,58	SD
12	Item_15	0,58	SD
13	Item_16	0,50	SD
14	Item_17	0,55	SD
15	Item_18	0,57	SD
16	Item_19	0,54	SD
17	Item_21	0,57	SD
18	Item_22	0,59	SD
19	Item_23	0,42	SD
20	Item_24	0,56	SD

No	Item	Koef	Ket
21	Item_25	0,43	SD
22	Item_26	0,58	SD
23	Item_27	0,36	SD
24	Item_28	0,19	S
25	Item_30	0,58	SD
26	Item_31	0,28	S
27	Item_32	0,56	SD
28	Item_33	0,45	SD
29	Item_34	0,57	SD
30	Item_35	0,56	SD
31	Item_36	0,98	MD
32	Item_37	0,51	SD
33	Item_38	0,55	SD
34	Item_39	0,93	MD
35	Item_40	0,92	MD

Keterangan : MD = Mudah SD = Sedang S = Sulit

Item dengan indeks daya beda kategori “K” mempunyai koefisien antara -0,10 sampai dengan 0,20. Item dengan indeks daya beda kategori “S” mempunyai koefisien antara 0,13 sampai dengan 0,24. Item dengan indeks daya beda kategori “B” mempunyai koefisien antara 0,30 sampai dengan 0,38. Adapun Item dengan indeks daya beda kategori “SB” mempunyai koefisien antara 0,40 sampai dengan 0,73.

Berdasarkan Tabel 7 dapat diketahui bahwa dari 35 item yang dianalisa terdapat 4 (empat) item soal atau 11,4% yang dinyatakan outfit (tidak cocok), yakni item 11, item 21, item 23, item 27, dan item 31. Adapun sebanyak 31 item atau 88,5 % dinyatakan infit (cocok). Adapun estimasi reliabilitas dari seluruh item tes tersebut sebesar 0,75. Hal ini menunjukkan bahwa seluruh item memiliki tingkat reliabilitas yang tinggi.

Tabel 6 : Koefisien Daya Beda

No	Item	Koef	Ket	No	Item	Koef	Ke t
1	Item_2	0,43	SB	21	Item_25	0,49	SB
2	Item_4	0,36	B	22	Item_26	0,35	B
3	Item_5	0,47	SB	23	Item_27	- 0,10	K
4	Item_7	0,60	SB	24	Item_28	0,32	B
5	Item_8	0,56	SB	25	Item_30	0,30	B
6	Item_9	0,13	S	26	Item_31	- 0,14	K
7	Item_10	0,55	SB	27	Item_32	0,24	S
8	Item_11	0,63	SB	28	Item_33	0,38	B
9	Item_12	0,41	SB	29	Item_34	0,24	S
10	Item_13	0,21	S	30	Item_35	0,24	S
11	Item_14	0,05	K	31	Item_36	0,48	SB
12	Item_15	0,33	B	32	Item_37	0,48	SB
13	Item_16	0,41	SB	33	Item_38	0,53	SB
14	Item_17	0,62	SB	34	Item_39	0,42	SB
15	Item_18	0,10	K	35	Item_40	0,43	SB
16	Item_19	0,73	SB				
17	Item_21	0,20	S				
18	Item_22	0,09	K				
19	Item_23	0,40	SB				
20	Item_24	0,31	B				

Keterangan : SB = Sangat Baik, B = Baik, S = Sedang,
K = Kurang

Tabel 7 : Kecocokan Butir

No	Item	Koef	Ket
1	Item_2	0,87	C
2	Item_4	1,04	C
3	Item_5	0,93	C
4	Item_7	0,73	C
5	Item_8	0,84	C
6	Item_9	1,14	C
7	Item_10	0,78	C
8	Item_11	0,69	TC
9	Item_12	0,88	C
10	Item_13	1,13	C
11	Item_14	1,13	C
12	Item_15	0,90	C
13	Item_16	0,98	C
14	Item_17	0,73	C
15	Item_18	1,12	C
16	Item_19	0,73	C
17	Item_21	0,98	C
18	Item_22	1,07	C
19	Item_23	1,03	TC
20	Item_24	0,99	C

No	Item	Koef	Ket
21	Item_25	0,92	C
22	Item_26	0,89	C
23	Item_27	1,52	TC
24	Item_28	0,92	C
25	Item_30	0,90	C
26	Item_31	1,68	TC
27	Item_32	1,11	C
28	Item_33	1,02	C
29	Item_34	1,08	C
30	Item_35	1,03	C
31	Item_36	0,80	C
32	Item_37	0,91	C
33	Item_38	0,84	C
34	Item_39	0,92	C
35	Item_40	0,88	C

Keterangan : C = Cocok

TC = Tidak Cocok

E. PEMBAHASAN

Pembahasan diawali dari butir soal dengan kasus *item has perfect score (IHPS)*. Berdasarkan analisa dapat diketahui bahwa terdapat 5 (lima) item soal yang dapat dijawab oleh semua peserta tes. Hal ini menunjukkan bahwa soal tersebut terlalu mudah untuk dikerjakan oleh peserta didik. Soal ini tidak dapat membedakan antara peserta didik yang memiliki kemampuan tinggi dengan peserta didik dengan kemampuan rendah. Soal dengan kasus *IHPS* hanya berfungsi untuk menambah skor nilai bagi peserta didik, akibat dari kasus *IHPS* ini adalah akan menimbulkan penilaian yang bias dimana

peserta didik dengan skor nilai yang tinggi tetapi memiliki kompetensi yang rendah.

Tabel 8 : Efektivitas Distraktor

No	Item	Opsì (%)				Ket
		A	B	C	D	
1	Item_2	13,30	28,30	8,30	50,00*	Efektif
2	Item_4	83,30*	8,30	3,30	5,00	Efektif
3	Item_5	1,70	3,30	80,00*	15,00	Efektif
4	Item_7	23,30	6,70	23,30	46,70*	Efektif
5	Item_8	83,30*	5,00	5,00	6,70	Efektif
6	Item_9	1,70	93,30*	1,70	3,30	Efektif
7	Item_10	1,70	0,00	95,00*	3,30	Opsi B diganti
8	Item_11	0,00	1,70	1,70	96,70*	Opsi A diganti
9	Item_12	0,00	98,30*	0,00	1,70	Opsi A dan C diganti
10	Item_13	90,00*	0,00	6,70	3,30	Opsi B diganti
11	Item_14	0,00	3,30	0,00	96,70*	Opsi A dan C diganti
12	Item_15	1,70	0,00	96,70*	1,70	Opsi B diganti
13	Item_16	1,70	83,30*	3,30	11,70	Efektif
14	Item_17	3,30	91,70*	1,70	3,30	Efektif
15	Item_18	1,70	1,70	96,60*	0,00	Opsi D diganti
16	Item_19	3,30	90,00*	5,00	1,70	Efektif
17	Item_21	1,70	3,30	95,00*	0,00	Opsi D diganti
18	Item_22	1,70	0,00	0,00	98,30*	Opsi B dan C diganti
19	Item_23	23,30	0,00	6,70	70,00*	Opsi B diganti
20	Item_24	0,00	93,30*	3,30	3,30	Opsi A diganti
21	Item_25	13,3	8,3	6,7	71,70*	Efektif
22	Item_26	1,70	0,00	96,70*	1,70	Opsi B diganti
23	Item_27	0,00	33,30	60,00*	6,70	Opsi A diganti
24	Item_28	21,70	31,70*	15,00	31,70	Efektif
25	Item_30	1,70	1,70	0,00	96,70*	Opsi A diganti
26	Item_31	46,70*	6,70	33,30	13,30	Efektif
27	Item_32	6,70	93,30*	0,00	0,00	Opsi C dan D diganti
28	Item_33	15,00	5,00	5,00	75,00*	Efektif
29	Item_34	5,00	0,00	95,00*	0,00	Opsi B dan D diganti
30	Item_35	0,00	0,00	93,30	6,70	Opsi A dan B diganti
31	Item_36	98,30*	1,70	0,00	0,00	Opsi C dan D diganti
32	Item_37	13,30	85,00*	1,70	0,00	Opsi D diganti
33	Item_38	5,00	91,70*	3,30	0,00	Opsi D diganti
34	Item_39	93,30*	1,70	0,00	5,00	Opsi C diganti
35	Item_40	91,70*	0,00	5,00	3,30	Opsi B diganti

Keterangan : tanda * merupakan kunci jawaban

Pada karakteristik tingkat kesukaran item untuk tes materi Pendidikan Agama Islam, item soal memiliki koefisien tingkat kesukaran antara 0,30 sampai dengan 0,70. Hal ini sesuai dengan pendapat dari Allen dan Yen, yang menyatakan bahwa secara umum indeks kesukaran suatu butir sebaiknya terletak pada interval antara $(0,3 - 0,7)$ ¹⁵³.

Berdasarkan indeks kesukarannya item soal ini sudah sesuai dengan standar evaluasi pendidikan, namun perlu kurangi proporsi soal dengan kategori “Sedang” dan ditambah proporsi soal dengan kategori “Sulit”. Pada butir soal evaluasi ini masih dominan butir soal dengan kategori “Sedang”.

Karakteristik selanjutnya adalah daya beda. Berdasarkan analisa dapat dilihat bahwa koefisien daya beda menyebar secara proporsional. Daya beda dengan kategori “SB” memiliki jumlah yang lebih banyak. Hal ini menunjukkan bahwa sebagian besar item soal evaluasi tersebut dapat membedakan dengan sangat baik antara peserta didik yang berkemampuan tinggi dengan peserta didik yang berkemampuan rendah.

Soal evaluasi ini akan lebih baik apabila item 27 dan item 31 diganti dengan soal yang lebih baik, karena menghasilkan daya beda dengan tanda negative. Dua item ini membingungkan, dimana terdapat peserta didik yang berkemampuan tinggi menjawab item secara salah, sedangkan peserta didik berkemampuan rendah menjawab item dengan benar. Selanjutnya jumlah item yang memiliki daya beda dengan kategori “K” lebih baik jika dikurangi atau ditiadakan.

Karakteristik berikutnya adalah kecocokan butir. Berdasarkan analisa dapat diketahui bahwa terdapat 4 (empat) item yang terdeteksi tidak cocok (*outfit*). Ketidak cocokan ini

¹⁵³ Allen & Yen. *Op. cite*. hlm. 122.

karena koefisien *infit* kurang dari 0,77 sebagaimana yang terjadi pada item 11. Adapun item yang tidak fit karena koefisiennya lebih besar Dari 1,33 adalah item 23, 27, dan 31. Khusus pada item 27 dan 31 ini terkonfirmasi bahwa dua butir tersebut selain tidak cocok juga tidak memenuhi persyaratan daya beda.

Item yang cocok yakni item yang memiliki koefisien *infit* berkisar antara 0,77 sampai dengan 1,33. Interval *infit* ini sesuai dengan pendapat Didik Setyawarno¹⁵⁴ dan Bambang Subali dengan Pujiyati Suyoto¹⁵⁵. Proporsinya sebesar 88,5 %, hal ini menunjukkan bahwa sebagian besar item evaluasi merupakan item yang cocok untuk digunakan sebagai cara untuk memberikan evaluasi kepada peserta didik. Adapun sebagai upaya untuk mempertahankan kualitas evaluasi maka item yang tidak *fit* sebaiknya dikurangi atau ditiadakan.

Selanjutnya item evaluasi ini memiliki karakteristik reliabilitas yang tinggi. Berdasarkan *out put* dari Program QUEST menunjukkan bahwa koefisien reliabilitas item evaluasi ini sebesar 0,75. Menurut Guilford¹⁵⁶ koefisien tersebut merupakan koefisien yang tinggi. Oleh karena itu, item evaluasi ini merupakan item yang memiliki keajegan dan dapat digunakan sebagai item evaluasi.

Karakteristik terakhir adalah efektivitas distraktor. Pada karakteristik ini, msdih bsnysk terdapat opsi yang perlu diperbaiki agar item evaluasi memiliki tingkat kualitas yang baik. Berdasarkan analisa terdapat 18 item yang opsi distraktornya tidak dipilih oleh semua peserta didik. Bahkan ada yang dua (2) opsi disyaktor tidak dipilih oleh semua peserta didik dalam satu item soal evaluasi.

¹⁵⁴ Didik Setyawarno. *Op Cite*.

¹⁵⁵ Bambang Subali & Pujiyati Suyoto. *Op. Cite*.

¹⁵⁶ Guilford, J.P. *Op cite*.

Berdasarkan analisa terdapat 17 item yang memiliki distraktor secara efektif, artinya pada evaluasi ini terdapat hampir 50% yang distraktornya berfungsi secara efektif dan 50% lainnya tidak berfungsi secara efektif. Jumlah yang distraktornya tidak efektif masih terlalu banyak karena mendekati 50%. Pengembangan opsi khususnya untuk distraktor harus dikembangkan agar kualitas evaluasi menjadi semakin baik.

Indikator dari distraktor yang tidak efektif adalah opsi distraktor tersebut tidak dipilih oleh semua peserta didik bahwa distraktor tersebut. Semua peserta didik mengetahui bahwa opsi tersebut adalah opsi yang salah. Salah satu upaya untuk meningkatkan efektifitas distraktor adalah dengan memilih opsi yang lebih tersamar, sehingga peserta didik yang berkemampuan rendah “terpaksa” memilih. Diksi “terpaksa” penulis pilih karena menunjukkan peserta didik tersebut belum memahami materi pelajaran atau bahan evaluasi, sehingga peserta didik tersebut mudah terkecoh dengan opsi yang sebenarnya tidak benar.

F. KESIMPULAN

Berdasarkan out put yang dihasilkan dari program QUEST, analisa dan pembahasan yang telah dilakukan, maka dapat diambil beberapak kesimpulan tentang karakteristik item soal latihan ujian berstandar Nasional Pendidikan Agama Islam untuk kelas VI, yaitu :

1. Item evaluasi sebanyak 40 butir, akan tetapi 5 item mengalami *Item Has Perfect Score (IHPS)* sehingga item tersebut tidak disertakan dalam analisa.

2. Sebagian besar aitem memiliki tingkat kesukaran kategori “Sedang” dengan koefisien antara 0,30 sampai dengan 0,59.
3. Terdapat 48,5 % atau 17 dari 35 item soal yang dianalisa memiliki indeks daya beda dengan kategori “Sangat Baik”.
4. Terdapat 31 item atau 88,5 % merupakan item yang sesuai (*infit*) untuk mengevaluasi kemampuan peserta didik pada mata pelajaran Pendidikan Agama Islam.
5. Tingkat reliabilitas soal evaluasi ini sebesar 0,75 yang berarti bahwa item soal ini memiliki tingkat konsistensi yang baik.
6. Sebanyak 50% distraktor memiliki fungsi secara efektif, akan tetapi untuk item yang distraktornya tidak efektif, perlu diperbaiki dengan mengganti opsi yang lebih tersamar.

DAFTAR PUSTAKA

- Allen dan Yen. *Introduction to Measurement Theory*. California : Brooks/Cole Publishing Company Wadsworth Inc. 1979.
- Anas Sudijono. *Pengantar Evaluasi Pendidikan*. Jakarta : Rajawali Pres. 2009.
- Anastasi, A. & Urbina, S. *Psikologi Testing*. Edisi terjemah oleh Hariono, R & Imam, S. Jakarta : PT Prenhallindo. 1998.
- Bambang Subali & Pujiyati Suyata. *Panduan Menggunakan Program Quest. Lembaga Penelitian UNY*. 2011.
- Bond, T.G. & Fox, Ch.M. *Applying the rasch model: Fundamental measurement in the human sciences. 2-nd ed.* New Jersey : Mahwah, Lawrence Erlbaum Associates Publishers.2007.
- Cecil R. R; Ronald B. L; and Victor W. New Jersy : *Measurement and Assessment in Education (Second Edition)*. Pearson Education, Inc. 2009.
- Cronbach, L.J., Coeficient Alpha and The Internal Structure of Test, *Psychometrika*. 1951, Volume 16. hlm. 297-334.
- Didik Setyawarno. Upaya Peningkatan Kualitas Butir Soal Dengan Analisa Aplikasi QUEST. *Makalah PPM bagi guru IPA SMP di Sleman disampaikan pada hari Sabtu, 9 September 2017 di Lab. Komputer IPA. 2017. hlm. 9*
- Guilford, J.P. *Fundamental Statistic in Psychology and Education*. 3rd Ed. New York : McGraw-Hill Book Company, Inc. 1956.
- Hambleton, R.K., Swaminathan, H., & Rogers, H.J. *Fundamental of item response theory*. Newbury Park, CA : Sage Publication Inc.1991
- Mehrens, W.A. & Lehmann, I.J. *Measurement and evaluation in education and psychology*. New York : Hold, Rinehart and Wiston,Inc. 1973.
- Sudijono, A. *Pengantar Evaluasi Pendidikan*. Jakarta : PT RajaGrafindo Persada. 2011.
- Suharsimi Arikunto. *Dasar-Dasar Evaluasi Pendidikan (Edisi Revisi)*. Jakarta : Bumi Aksara. 2006.
- Suparman. *Evaluasi Pendidikan*. Yogyakarta : STAIMS Press. 2017.
- _____. *Panduan Membaca Out Put dari Quest. Hand Out Materi Kuliah*. Tidak Diterbitkan.