Afeksi: Jurnal Penelitian dan Evaluasi Pendidikan
Volume 3 Nomor 2 Tahun 2022
http://afeksi.id/jurna/index.php/afeksi
e-ISSN: 2745-9985

PENGEMBANGAN INSTRUMEN PENGUKUR KEMAMPUAN BERPIKIR TINGKAT TINGGI BAHASA ARAB MADRASAH ALIYAH

Rahmat Danni
Institut Agama Islam Negeri Syaikh Abdurrahman Siddik Bangka Belitung
Email: rahmatdanni@iainsasbabel.ac.id

Abstrak

Penelitian ini bertujuan untuk mengembangkan instrumen higher order thinking skills Bahasa Arab tingkat Madrasah Aliyah di Provinsi Kep. Bangka Belitung. Instrumen dikembangkan mengacu Model pengembangan Oriondo & Antonio. Sampel penelitian berjumlah 500 peserta didik kelas XI dari tiga Madrasah Aliyah yang ditentukan menggunakan Multistage random sampling. Hasil penelitian menunjukkan bahwa instruman higher order thinking skills terbukti valid dan reliabel sehingga layak digunakan untuk mengukur kemampuan berpikir tingkat tinggi peserta didik. Hal ini dibuktikan dari perolehan rerata indeks validitas isi instrumen melalui *expert judgement* yaitu sebesar 0,97. Hasi uji coba instrumen secara empirik menunjukkan *reliabilitas items* sebesar 0,93 dan *reliabilitas persons* sebesar 0,71, *Outfit mean square* (MNSQ) di rentang 0,87 – 1,59, *Outfit Z-standard* (ZTSD) berada di rentang -1,9 – 1,9, dan *Point measure correlation* di rentang 0,52 – 0,66 serta tingkat kesukaran butir tergolong sedang dengan rentang -1,9 sampai 2,46. Dengan demikian Instrumen HOTS bahasa Arab yang dikembangkan terbukti valid dan layak digunakan untuk mengukur HOTS bahasa Arab peserta didik.

Kata Kunci: Bahasa Arab, Kemampuan Berpikir Tingkat Tinggi, Madrasah Aliyah

PENDAHULUAN

Mewujudkan sumber daya manusia yang unggul tidak dapat dipisahkan dengan pendidikan yang berkualitas. Setiap manusia tentu mengalami proses pendidikan baik itu formal, nonformal maupun informal. Undang-undang nomor 20 tahun 2013 tentang sistem pendidikan nasional menjelaskan bahwa pendidikan formal adalah pendidikan yang berjenjang dan terstruktur yang terdiri atas pendidikan dasar, pendidikan menengah, dan pendidikan tinggi sedangkan pendidikan nonformal adalah pendidikan tambahan di luar pendidikan formal, kemudian pendidikan informal adalah jalur pendidikan keluarga dan lingkungan yang bentukya belajar secara mandiri.

Pada pendidikan formal, tugas guru begitu kompleks. Hal ini disebutkan dalam Undang-undang nomor 14 tahun 2005 bahwa guru adalah pendidik profesional dengan tugas utama mendidik, mengajar, membimbing, mengarahkan, melatih, menilai dan mengevaluasi peserta didik. Tugas guru yang kompleks harus diimbangi dengan keahlian guru yang kompleks pula. Guru diharapkan tidak hanya menguasi materi yang diajarkan, akan tetapi guru juga harus memiliki keahlian pedagogik seperti terampil dalam menyampaikan materi pelajaran, membuat soal ujian, serta menilai dan mengevaluasi pencapaian peserta didik.

Pencapaian peserta didik dalam kegiatan pembelajaran diketahui berdasarkan kompetensi-kompetensi yang telah ditentukan. Apabila peserta didik telah menguasi kompetensi yang diharapkan maka peserta didik dapat dinyatakan lulus atau naik pada jenjang berikutnya. Pencapaian peserta didik selama mengikuti kegiatan pembelajaran dapat diketahui melalui proses penilaian. Penilaian adalah usaha guru maupun peserta didik dalam mendapatkan informasi mengenai pencapaian pembelajaran (Rasyid & Mansyur, 2008). Senada dengan itu Hayat (2004) mendefinisikan penilaian sebagai proses pengumpulan

informasi yang dilakukan guru mengenai perkembangan dan pencapaian kegiatan pembelajaran melalui berbagai teknik secara tepat. Selain itu, hasil penilaian juga dapat menjadi gambaran bagi guru mengenai kualitas pembelajaran. Apabila sebagian peserta didik mendapatkan hasil yang kurang memuaskan maka guru harus mencari solusi guna meningkatkan kualitas pembalajaran.

Sistem penilaian yang baik harus mampu (1) memberi informasi yang akurat, (2) memotivasi peserta didik dan guru, (3) meningkatkan kinerja lembaga, dan (4) meningkatkan kualitas pendidikan (Mardapi, 2012). Kontras dengan hal itu, selama ini sistem penilaian yang dipakai guru cenderung tidak menggambarkan sebagai sistem penilaian yang baik. Hasil penilaian lebih banyak digunakan untuk menilai peserta didik tanpa menjadi bahan penilaian diri guru dalam proses pembelajaran. Selain itu, proses penilaian yang dilakukan juga cenderung mengabaikan tahapan-tahapan yang benar karena guru lebih fokus kepada proses penyampaian materi ketimbang menilai dan mengavaluasi proses pembelajaran.

Sistem penilaian yang buruk berdampak pada hasil yang tidak akurat. Informasi yang tidak bisa merefleksikan keadaan peserta didik dapat menyebabkan pengambilan kebijakan yang keliru karena tidak sesuai dengan kebutuhan peserta didik. Guna mencegah hal tersebut, guru harus mampu membuat sistem peniliaian yang berkualitas. Ada empat hal penting dalam pembuatan penilaian kelas yang berkualitas, yaitu (1) *clear purposes,* artinya kegiatan penilaian harus memiliki tujuan yang jelas, (2) *clear targets*, target yang hendak diketahui harus jelas (3) *sound design*, instrumen yang digunakan harus tepat dan (4) *effective communication*, pelaporan hasil mudah untuk dipahami(Stiggins & Chappuis, 2012). Oleh karena itu, sistem penilaian harus dikonstruk melalui tahapan yang benar tanpa mengabaikan empat hal tersebut.

Selain itu, alat pengukur kemampuan dalam sistem penilaian yaitu berupa instrumen harus terdiri dari butir-butir soal yang baik (Suwarto, 2016). Reynolds, Livingston, & Willson (2009) mengatakan bahwa karakteristik instrumen yang baik meliputi validitas dan reliabilitas. Sedangkan Linn & Gronlund (1995) menyebutkan tiga kriteria instrumen berkategori baik yaitu validitas, reliabilitas, dan usabilitas. Validitas dan reliabilitas merupakan syarat penting pada suatu instrumen, akan tetapi masih saja banyak pendidik abai dengan keduanya. Burhan (Nurgiyantoro, 2001) mengatakan bahwa umumnya tes prestasi yang dirakit oleh guru tidak melalui tahap uji coba terlebih dahulu dikarenakan berbagai hal, seperti masalah waktu, tenaga, biaya dan kemampuan guru itu sendiri dalam menganalisisnya. Apabila bicara mengenai instrumen pengukur tes prestasi belajar maka tak bisa dipisahkan dari dua teori pengukuran yaitu *classical test theory* dan *item response theory* (Suryabrata, 2005).

Classical test theory atau dikenal juga dengan teori tes klasik merupakan teori pengukuran dikalangan pakar psikologi dan pendidikan (Embretson & Reise, 2000). Tak heran jika teori ini masih banyak digunakan dalam pengukuran kemampuan peserta didik di dunia pendidikan. Allen & Yen (1979) mengungkapkan bahwa teori tes klasik adalah akumulasi dari *true score* dan *error* (X = T + E). Teori inilah yang mendasari penskoran kemampuan peserta didik di Indonesia. Apabila disuatu tes terdiri dari 10 butir soal pilihan ganda, kemudian seorang peserta didik mampu menjawab 7 butir soal dengan benar maka ia mendapatkan skor 7, sedangkan sisanya yaitu 3 adalah *error*. Namun seiring berkembangnya waktu, ditemukan beberapa kelemahan dari teori tes klasik, diantaranya adalah kemampuan peserta tes dan butir soal tidak independen (Fan, 1998; Hambleton & Swaminathan, 1985; Hambleton, Swaminathan, & Rogers, 1991; Lord, 1980). Sedangkan menurut Mardapi (2012) kelemahan teori tes klasik yaitu hasil pengukuran bergantung pada karakteristik butir soal, parameter butir soal bergantung pada kemampuan peserta tes, dan kesalahan pengukuran yang dapat diketahui hanya untuk kelompok individu. Oleh sebab itu, guna menutupi kelemahan itu terbentuklah teori baru yang dikenal dengan *item response theory* salah satunya adalah rasch model.

Akan tetapi meski teori tes klasik memiliki kelemahan dan *item response theory* dengan kebaharuannya tetap saja teori tes klasik lebih digandrungi oleh para praktisi pendidikan dalam mengukur kemampuan peserta didik di Indonesia. Bahkan *item response theory* masih menjadi suatu hal yang asing bagi pendidikan di Indonesia. Hal ini terbukti dengan masih minimnya penelitian yang menggunakan pendekatan *rasch model* khususnya di bidang pendidikan bahasa Arab.

Dani (2013) dalam penelitiannya menyatakan bahwa dari 40 butir soal tes ujian semester genap bahasa Arab di SMA Muhammadiyah 1 Purwokerto hanya ada 16 butir soal yang valid, 52,5% butir soal memiliki tingkat kesukaran (b_i) tergolongi mudah, 13 butir soal memiliki daya beda (a_i) tergolong sedang, 11 butir soal perlu direvisi, dan 16 butir soal gugur/ dihapus. Kelemahan teori tes klasik dalam hal ini akan tampak, karena pada teori tes klasik karakteristik butir soal inkonsisten. Karakteristik butir soal pada teori tes klasik akan selalu berubah-ubah mengikuti kemampuan peserta tes. Apabila disuatu ketika paket soal diujikan pada kelompok peserta yang memiliki kemampuan tinggi, maka tingkat kesukaran butir soal menjadi rendah dan jika diujikan lagi pada kelompok lain, maka karakteristik butir soal berubah kembali.

Berdasarkan beberapa permasalahan yang dihadapi pendidik dalam membuat instrumen pengukur kemampuan berpikir tingkat tinggi bahasa Arab yang berkarakteristik baik dan teruji secara empirik, maka perlu adanya pengembangan tes prestasi belajar bahasa Arab menggunakan pendekatan rasch model. Oleh karena itu, penelitian ini bertujuan untuk mengembangkan instrument HOTS Bahasa Arab yang berkualitas menggunakan Rasch model.

METODOLOGI PENELITIAN

Penelitian pengembangan ini mengadopsi model pengembangan instrumen Oriondo & Dallo-Antonio (1998). Tujuan penelitian ini adalah pengembangan instrumen HOTS bahasa Arab tingkat Madrasah Aliyah kelas XI. Tahapan pengembangan instrumen dibagi menjadi tiga tahapan yaitu tahap pertama perancangan instrumen, tahap kedua penilaian ahli, dan tahap ketiga uji coba tes. Sampel penelitian berjumlah 500 peserta didik dari tiga Madrasah Aliyah di Provinsi Kep. Bangka Belitung yang ditentukan menggunakan teknik Multistage random sampling. Instrumen HOTS bahasa Arab yang dikembangkan berjumlah 10 butir soal.

Tahap perancangan instrumen berisikan perumusan materi mata pelajaran bahasa Arab kelas XI Madrasah Aliyah, penyusunan kisi-kisi, dan pembuatan butir soal. Tahap penilaian ahli bertujuan untuk mengetahui validitas isi instrumen HOTS bahasa Arab. Validitas isi instrumen pengukur kemampuan bahasa Arab diketahui melalui penilaian ahli menggunakan lembar telaah butir soal. Hasil peneilaian dianalisis menggunakan formula Aiken (Aiken, 1980, p. 956; Azwar, 2015). Indeks V Aiken direntang 0,4 – 0,8 tergolong valid dan diatas 0,8 tergolong sangat valid (Retnawati, 2016a). Tahap ketiga uji coba instrumen secara empirik untuk mengetahui karakteristik butir soal HOTS bahasa Arab berdasarkan rasch model. Reliabilitas didefinisikan sebagai suatu koefisien konsistensi hasil dari pengukuran (Hadi, 2013; Mardapi, 2016). Formula yang digunakan adalah cronbach's alpha (Cronbach, 1951). Apabila koefisien mendekati 1 itu menunjukkan bahwa butir memiliki reliabilitas yang layak untuk mengukur materi (Vakili & Jahangiri, 2018). Analisis butir soal HOTS bahasa Arab dilakukan menggunakan Rasch model melalui software jMetrik. Instrumen HOTS dapat diterima apabila memenuhi kriteria yaitu nilai Outfit mean square (MNSQ) berada di rentang 0,5<MNSQ<1,5, Outfit Z-standard (ZTSD) berada di rentang -2<ZSTD<+2, dan Point measure Correlation di rentang 0,3<PMC<0,85 (Boone et al., 2014; DeMars, 2018; Petrillo et al., 2015; Sumintono & Widhiarso, 2015). Dalam software Jmetrik

Outfit mean square (MNSQ) dikenal dengan WMS (Weighted mean square) dan ZTSD dikenal dengan Standardized WMS (Mayer, 2014).

HASIL DAN PEMBAHASAN

Perancangan tes

Pengembangan instrumen pengukur kemampuan berpikir tingkat tinggi peserta didik kelas XI diawali dengan perancangan instrumen. Perumusan materi Bahasa arab kelas XI Madrasah Aliyah. Instrumen pengukur HOTS Bahasa Arab kelas XI yang dikembangkan mengadopsi kurikulum 2013 mata pelajaran Bahasa Arab tingkat Madrasah Aliyah. Instrumen HOTS kelas XI disusun berdasarkan materi عدد ألف و مليون و مليار و بليون, حروف الجر و حروف العطف, الاسم النكرة و الاسم المعرفة. Kisi-kisi instrumen dikonstruk dengan memperhatikan Kompetensi Inti, Kompetensi Dasar, Materi Pelajaran, Indikator, dan Level Kognitif berdasarkan taksonomi bloom taksonomi bloom yang telah direvisi sebagai indikator, yaitu C4 (analysing), C5 (evaluating), dan C6 (creating) (Anderson & Krathwohl, 2001).

Penilaian ahli

Azwar (2015) mendefinisikan validitas isi sebagai ukuran sejauh mana butir tes dapat mengukur aspek atau ruang lingkup yang ingin diketahui. Pada penelitian ini, validitas isi diketahui berdasarkan *expert judgment*. Adapun hasil telah secara kualitatif disajikan pada Tabel 1. Validitas instrumen pengukur HOTS bahasa Arab dibuktikan menggunakan validitas isi melalui *expert judgement*. Penilaian diberikan oleh pakar di bidang Bahasa Arab. Hasil penilaian dianalisis menggunakan formula Aiken. Adapun hasil pembuktian validitas instrument HOTS Bahasa Arab menggunakan Formula Aiken disajikan pada Tabel 1.

Tabel 1. Validitas instrument HOTS bahasa Arab kelas XI

No Butir	Penilaian Validator	s1	∑s	Indeks Aiken
1	4	3	3	1,00
2	4	3	3	1,00
3	4	3	3	1,00
4	4	3	3	1,00
5	4	3	3	1,00
6	4	3	3	1,00
7	4	3	3	1,00
8	4	3	3	1,00
9	3	2	2	0,67
10	4	3	3	1,00
Rata-rata				0,97

Tabel 1 menunjukkan bahwa instrumen HOTS Bahasa Arab memiliki indeks V Aiken di rentang 0,67 – 1,00. Terdapat 90% butir dari 10 butir soal HOTS Bahasa Arab yang memperoleh V Aiken sebesar 1,00 dan 10% butir soal memperoleh V Aiken sebesar 0,67. Berdasarkan data pada Tabel 1, maka dapat dinyatakan bahwa 10 butir soal HOTS Bahasa Arab yang dikembangkan tergolong valid. Sebagaimana yang ditegaskan oleh Azwar (2015) bahwa butir soal tergolong valid apabila memiliki indeks >0,4. Dengan demikian instrument HOTS Bahasa Arab kelas XI terbukti dapat mengukur kemampuan HOTS Bahasa Arab peserta didik. Hasil penilaian ahli menguatkan bahwa instrumen HOTS bahasa Arab yang dikembangkan layak digunakan untuk mengukur kemampuan bahasa Arab peserta didik.

Uji coba instrumen

Hasil uji coba instrumen HOTS bahasa Arab secara empirik dilakukan pada 500 peserta didik kelas XI dari tiga madrasah Aliyah di Provinsi Kep. Bangka Belitung dianalisis menggunakan rasch model. Hasil analisis menunjukkan reliabilitas items 0.93 dan reliabilitas persons sebesar 0.71. Ini membuktikan bahwa items instrumen HOTS bahasa Arab memiliki konsistensi insternal yang tinggi (Tran et al., 2018). Sedangkan pada reliabilitas person masih tergolong cukup. Reliabilitas yang dapat diterima berada pada range diatas 0.67, sedangkan dibawah 0.67 tergolong rendah (Mohamad et al., 2015). Hasil analisis butir soal berdasarkan hasil uji coba secara empirik ditampilkan pada Tabel 2.

Tabel 2. Hasil uji coba instrumen HOTS Bahasa Arab

Butir	Outfit MNSQ	Outfit ZTSD	Pt. Measure Correlation	Ket.
5	1,59	1,9	0,56	Fit
7	0,97	-0,2	0,60	Fit
1	1,20	1,6	0,58	Fit
9	0,87	-1,3	0,64	Fit
3	0,81	-1,9	0,66	Fit
6	0,98	-0,2	0,62	Fit
4	0,93	-0,7	0,63	Fit
10	1,02	0,2	0,63	Fit
8	1,04	0,3	0,60	Fit
2	0,87	-0,2	0,52	Fit

Tabel 2 menunjukkan bahwa instrumen HOTS bahasa Arab yang dikembangkan memiliki Outfit mean square (MNSQ) di rentang 0,87 – 1,59 dan Outfit Z-standard (ZTSD) berada di rentang -1,9 – 1,9, dan Point measure correlation di rentang 0,52 – 0,66. Kecocokan butir dengan rasch model ditentukan berdasarkan kriteria *Outfit mean square (MNSQ)* di rentang 0,5<MNSQ<1,5, *Outfit Z-standard (ZTSD)* di rentang -2<ZSTD<+2, dan *Point measure Correlation* di rentang 0,4<PMC<0,85 (Abdellatif, 2023; Mayer, 2014; Sumintono & Widhiarso, 2015). Berdasarkan kriteria tersebut dapat diketahui bahwa bahwa semua butir soal pengukur HOTS Bahasa Arab baik kelas XI semua memenuhi kriteria dan fit terhadap rasch model.

Kriteria berikutnya adalah tingkat kesukaran butir. Bond & Fox (2015) mengungkapkan bahwa kriteria tingkat kesukaran ideal beradaa pada rentang -2 sampai +2. Hasil analisis tingkat kesukaran butir soal HOTS bahasa Arab kelas XI ditampilkan pada Tabel 3.

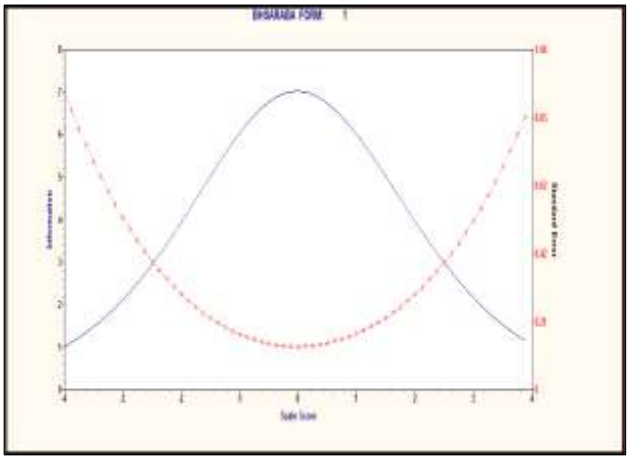
Tabel 3. Indeks tingkat kesukaran butir soal HOTS Bahasa Arab

Butir	Tingkat Kesukaran (b)
5	2,46
7	1,18
1	0,76
9	0,56
3	0,30
6	0,12
4	0,11

Butir	Tingkat Kesukaran (b)
10	-0,64
8	-1,44
2	-1,90

Pada Tabel 3 tampak bahwa tingkat kesukaran butir soal HOTS Bahasa Arab kelas XI tertinggi dimiliki butir soal nomor 5 dengan indeks 2,46 dan terendah nomor 2 dengan indeks -1,90. Berdasarkan hasil tersebut maka dapat disimpulkan bahwa terdapat 10% atau 1 butir soal yang tidak memenuhi sebagai butir soal dengan indeks ideal, yaitu butir nomor 5. Butir soal yang tidak memenuhi indeks kesukaran ideal yaitu direntang 2 sampai -2 dinyatakan gugur karena tidak dapat berfungsi dengan baik untuk mengukur HOTS Bahasa Arab. Butir soal nomor 5 memiliki indeks kesukaran 2,46 menunjukkan bahwa butir tersebut tergolong sangat sulit sehingga peserta didik sebagian besar tidak mampu menjawab dengan benar.

Selain itu, diperoleh pula fungsi infomasi paket soal. Retnawati (2014) mengatakan bahwa fungsi informasi adalah cara untuk menjelaskan sumbangan suatu butir soal dalam memberikan informasi mengenai kemampuan (θ) peserta tes yang hendak diukur. Informasi yang diperoleh dari suatu perangkat tes akan selalu berbanding terbalik dengan kesalahan pengukuran/ *standard error of measurement* (SEM), sehingga semakin besar fungsi informasi yang diperoleh maka semakin kecil kesalahan pengukuran. Fungsi informasi tes disajikan pada Gambar 1.



Gambar 1. Fungsi informasi instrument pengukur higher order thinking skills Bahasa Arab

Gambar 1 berupa kurva yang menunjukkan *test information function* dan *standar error of measurement* (SEM). Pertemuan kurva fungsi informasi dan SEM menginformasikan terkait kehandalan instrument untuk mengukur kemampuan (Gao & Liu, 2024). Berdasarkan Gambar 1 dapat diketahui bahwa instrumen HOTS bahasa Arab yang dikembangkan dapat memberikan informasi secara optimal apabila diberikan pada peserta didik dengan kemampuan/ability rentang -2,5 sampai +2,6. Instrumen ini memberikan informasi tertinggi sebesar 7 dengan Standar Error of Measurement sebesar 0.3 pada peserta didik dengan kemampuan 0. Dengan demikian, instrumen HOTS bahasa Arab dapat memberikan informasi maksimal bila diujikan pada peserta didik dengan kemampuan sedang.

KESIMPULAN

Berdasarkan hasil penelitian yang telah dipaparkan maka dapat disimpulkan bahwa instrumen pengukur kemampuan berpikir tingkat tinggi bahasa Arab untuk peserta didik kelas XI Madrasah Aliyah terbukti valid dan reliabel secara empiris serta telah memenuhi kriteria butir soal yang baik. 9 dari 10 butir soal pengukur kemampuan berpikir tingkat tinggi bahasa arab telah memenuhi kriteria rasch model. Validitas isi instrumen berdasarkan penilaian ahli memiliki indeks aiken direntang 0,67 sampai 1,0 dengan rata-rata 0,97. Pengestimasian reliabilitas menunjukkan *reliabilitas items* sebesar 0,93 dan *reliabilitas persons* sebesar 0,71. Berdasarkan Rasch model kriteria tingkat kesukaran butir tergolong sedang dengan rentang -1,9 sampai 2,46, *Outfit mean square* (MNSQ) di rentang 0,87 – 1,59 dan *Outfit Z-standard* (ZTSD) berada di rentang -1,9 – 1,9, dan *Point measure correlation* di rentang 0,52 – 0,66. Dengan demikian Instrumen HOTS bahasa Arab yang dikembangkan terbukti valid dan layak digunakan untuk mengukur HOTS bahasa Arab peserta didik.

DAFTAR PUSTAKA

Aiken, L. R. (1985). Three Coefficients for Analyzing the Reliability and Validity of Ratings. *Educational and Psychological Measurement*, *45*(1), 131–142.

Allen, M. J., & Yen, W. M. (1979). *Introduction to Measurement Theory*. Monterey: Books Cole Publishing Company.

Azwar, S. (2015). *Reliabilitas dan Validitas* (4th ed.). Yogyakarta: Pustaka Belajar.

Baker, F. B. (2001). *The Basic of Item Response Theory* (2nd ed.). USA: ERIC Clearinghouse on Assessment and Evaluation.

Dani, H. (2013). *Pelaksanaan evaluasi analisis butir soal bahasa Arab di SMA Muhammadiyah 1 Purwokerto*. Universitas Islam Negeri Sunan Kalijaga, Yogyakarta.

DeMars, C. (2010). *Item response theory*. New York: Oxford University Press.

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologist*. NJ: Lawrence Erlbaum Associates Inc.

Fan, X. (1998). Item Response Theory and Classical Test Theory: An Empirical Comparison of Their Item/Response Person Statistics. *Educational and Psychological Measurement*, *58*(3), 357–381.

Hair, J. F., Black, W. C., Babin, B. J., & Anderson, R. E. (2010). *Multivariate data analysis* (7th ed.). Ottawa: Pearson. Retrieved from https://www.pearson.com/us/higher-education/program/Hair-Multivariate-Data-Analysis-7th-Edition/PGM263675.html

Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory*. Boston: Kluwer Inc.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamental of item response theory*. Newbury Park: Sage Publication Inc.

Hayat, B. (2004). Penilaian kelas (classroom assessment) dalam penerapan standard kompetensi. *Jurnal Pendidikan Penabur*, *III*, 108–112.

Linn, & Gronlund. (1995). *Measurement and assessment in teaching* (7th ed.). Englewood Cliffs, NJ: Prentice-Hall.

Lord, F. (1980). *Application of Item Response Theory to Practical Testing Problems. Hillsdale*. New Jersey: Lawrence Erlbaum Associates Publishers.

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

Naga, D. S. (1992). *Pengantar teori sekor pad pengukuran pendidikan*. Jakarta: Gunadarma.

Nurgiyantoro, B. (2001). *Penilaian dalam pengajaran bahasa dan sastra*. Yogyakarta: BPFE UGM.

Oriondo, L. L., & Dallo-Antonio, E. M. (1998). *Evaluation educaational outcomes*. Manila: Rex Printing Compagny inc.

Rasyid, H., & Mansyur. (2008). *Penilaian hasil belajar*. Banndung: Wacana Prima.

Retnawati, H. (2014). *Teori respons butir dan penerapannya*. Yogyakarta: Nuha Medika.

Retnawati, H. (2016). *Validitas, reliabilitas dan karakteristik butir*. Yogyakarta: Nuha Medika.

Reynolds, C. R., Livingston, R. B., & Willson, V. (2009). *Measurement and assessment in education* (2nd ed.). Upper Saddle River, NJ: Pearson.

Stiggins, R., & Chappuis, J. (2012). *Introduction to student invoved assessment for learning* (6th ed.). Boston: Addison Wesley.

Suwarto. (2016). Karakteristik tes biologi kelas 7 semester gasal. *Proceeding Biology Education Conference*, *13*, 159–163.

Wagiran. (2015). *Metodologi Penelitian Pendidikan (Teori dan Implementasi)*. Yogyakarta: Deepublish.

Widyastuti, A. (2015). *Analisis butir soal ulangan tengah semester mata pelajaran bahasa Arab kelas VII semester ganjil SMP.I Muhammadiyah 2 Depok*. Universitas Islam Negeri Sunan Kalijaga, Yogyakrata.

Wright. (2008). *Educational assessment*. California: Sage Publications.