Available online:
http://journal.imla.or.id/index.php/arabi
Arabi : Journal of Arabic Studies, 6 (1), 2021, 93-104
DOI: http://dx.doi.org/10.24865/ajas.v6i1.320

ITEM RESPONSE THEORY APPROACH: KALIBRASI BUTIR SOAL PENILAIAN AKHIR SEMESTER MATA PELAJARAN BAHASA ARAB

Rahmat Danni[1], Ajeng Wahyuni[2], Tauratiya[3]
[13] Institut Agama Islam Negeri Syaikh Abdurrahman Siddik, Indonesia
[2] Institut Agama Islam Negeri Ponorogo, Indonesia
Corresponding E-mail: rahmatdanni@iainsasbabel.ac.id

Abstract

This study describes the item details of the final semester questions in Arabic MAN 1 Pangkalpinang using the item response theory approach. The problem behind this research is that the development of Arabic final assessment items did not go through the correct stages. Therefore, this research is quantitative research. The subjects of this study were 176 students of class XI MAN 1 Pangkalpinang. The answer data is in the form of answers to questions in the final semester in Arabic which are 40 multiple-choice items with five answers. The results showed that the final results of the Arabic semester (1) proved valid, indicated by 40 items (100%) with loading factors; (2) proven to be reliable, indicated by the reliability coefficient of 0.884; (3) there are 33 items (82.5%) of the 40 items that have a good level of difficulty and distinguishing power so that they can be stored in the question bank and used in subsequent activities, while 7 items (17.5%) are item number 10, 26, 27, 29, 32, 34, and 35 do not meet the criteria for a good level of difficulty so they need to be revised or eliminated; and (4) suitable for use in students with low to moderate ability (θ) in the range -3.5 to +1.5 in logit. Future research is expected to be able to analyze Arabic language question items in the form of descriptive tests on a large scale or develop high-quality high-order thinking skills in Arabic.

Keywords: *Arabic, final assessment, item response theory, logistic parameter*

Abstrak

Penelitian ini bertujuan untuk mendeskripsikan karakteristik butir soal penilaian akhir semester bahasa Arab MAN 1 Pangkalpinang menggunakan pendekatan *item response theory*. Permasalahan yang melatarbelakangi penelitian ini adalah pengembangan butir soal penilaian akhir semester bahasa Arab yang tidak melalui tahapan yang benar. Penelitian ini merupakan penelitian deskriptif kuantitatif. Subjek penelitian berjumlah 176 peserta didik kelas XI MAN 1 Pangkalpinang. Data penelitian berupa jawaban soal penilaian akhir semester bahasa Arab berjumlah 40 butir pilihan ganda dengan lima alternatif jawaban. Hasil penelitian menunjukkan bahwa soal penilaian akhir semester bahasa Arab (1) terbukti valid ditunjukkan dengan 40 butir soal (100%) memiliki *loading factor* > 0,3; (2) terbukti reliabel ditunjukkan dengan koefisien reliabilitas > 0,7 yaitu 0,884; (3) terdapat 33 butir (82,5%) dari 40 butir soal memiliki tingkat kesulitan dan daya pembeda berkategori baik sehingga dapat disimpan dalam bank soal dan digunakan pada kegiatan penilaian selanjutnya, sedangkan 7 butir (17,5%) yaitu butir nomor 10, 26, 27, 29, 32, 34, dan 35 tidak memenuhi kriteria tingkat kesulitan yang baik sehingga perlu direvisi atau dieliminasi; dan (4) cocok digunakan pada peserta didik dengan kemampuan (θ) rendah hingga sedang dalam rentang -3,5 s.d. +1,5 dalam *logit*. Penelitian selanjutnya, diharapkan mampu melakukan analisis butir soal bahas Arab bentuk tes uraian dalam skala luas atau mengembangkan butir soal bahasa Arab berbasis *higher order thinking skills* (HOTS) yang berkualitas.

Kata Kunci: bahasa Arab, penilaian akhir semester, teori respons butir, parameter logistik

Pendahuluan

Pendidikan merupakan komponen penting dalam mencetak sumber daya manusia yang ahli di bidangnya dan mampu bersaing di era disruptif. Oleh sebab itu, lembaga pendidikan dituntut untuk meningkatkan kualitas pembelajaran agar melahirkan lulusan yang memiliki kecakapan abad 21. Trilling & Fadel (2009) mengungkapkan bahwa kecakapan abad 21 meliputi kemampuan berpikir kritis, pemecahan masalah, berkomunikasi dengan baik, dan mampu berkolaborasi. Pauw (2015) menambahkan kecakapan yang perlu dimiliki di antaranya menejmen diri, literasi informasi, dan kemampuan digital. Tuntutan yang tinggi, membuat guru harus mampu membangun sistem pembelajaran yang baik melalui desain pembelajaran yang mampu mengoptimalkan potensi diri peserta didik khususnya dalam menyelesaikan masalah yang kompleks (Prayogi & Estetika, 2019). Ditegaskan pula oleh Mardapi (2016) bahwa sistem pembelajaran yang dibangun dengan baik dapat menciptakan kualitas belajar yang baik.

Penilaian merupakan komponen yang selalu melekat pada proses belajar mengajar. Hasil penilaian bisa digunakan sebagai acuan guru untuk mengetahui keberhasilan dan meningkatkan kualitas pengajaran (Mardapi, 2016). Undang-undang nomor 14 tahun 2005 tentang Guru dan Dosen menegaskan bahwa guru tidak hanya sebatas mengajar, melainkan guru juga bertugas untuk melakukan penilaian dan evaluasi terhadap peserta didik. Oleh karena itu, selain dituntut mampu mendesain pembelajaran dengan baik, guru juga diharapkan mampu melakukan penilaian terhadap capaian peserta didik.

Stiggins & Chappuis (2012) mengungkapkan bahwa penilaian merupakan serangkaian kegiatan pengumpulan informasi terkait pencapaian peserta didik. Sedangkan Santoso, Kartianom, & Kassymova (2019) menyatakan bahwa penilaian adalah suatu kegiatan menafsirkan hasil pengukuran. Berdasarkan penjelasan tersebut, dapat disimpulkan bahwa penilaian merupakan penafsiran hasil pengukuran terkait capaian peserta didik. Selain itu, kegiatan penilaian dapat dilakukan setelah pengukuran. Hal ini menunjukkan bahwa penilaian dan pengukuran merupakan satu kesatuan (Suwandi, 2010). Penilaian tanpa melalui pengukuran yang benar akan menghasilkan informasi yang tidak akurat, sehingga capaian peserta didik dan keberhasilan pembelajaran tidak diketahui secara tepat. Oleh karena itu, dibutuhkan instrumen yang berkualitas dalam kegiatan pengukuran capaian belajar peserta didik.

Reynolds, Livingston, & Willson (2009) menyebutkan bahwa instrumen berkriteria baik adalah yang memenuhi syarat validitas dan reliabilitas. Instrumen dikatakan valid apabila dapat dibuktikan bahwa instrumen tersebut mengukur kemampuan peserta didik secara akurat sesuai dengan kompetensi yang diukur (Ramadhan, Mardapi, Prasetyo, & Utomo, 2019; Rindermann & Baumeister, 2015). Dengan demikian, alat ukut yang dipakai dalam mengukur kemampuan peserta didik harus dapat dibuktikan bahwa instrumen tersebut benar-benar mengukur kompetensi peserta didik. Sedangkan suatu instrumen memenuhi asumsi reliabilitas apabila skor amatan memiliki hubungan tinggi dengan skor sebenarnya (Allen & Yen, 1979). Uraian tersebut menunjukkan bahwa instrumen yang valid dan reliabel akan memberikan informasi secara akurat. Hal ini ditegaskan pula oleh Retnawati (2016) bahwa instrumen yang berkualitas akan selalu menghasilkan nilai informasi yang lebih tinggi dibandingkan kesalahan pengukuran.

Selain kriteria validitas dan reliabilitas, instrumen dengan bentuk tes perlu memenuhi kriteria lainnya, di antaranya meliputi tingkat kesulitan, daya beda, dan daya pengecoh/distraktor (Anita, A., Tyowati, S. & Zuldafrial, 2018; Iskandar & Rizal, 2018). Tingkat kesulitan menunjukkan peluang menjawab benar, apabila jumlah peserta tes yang mampu menjawab butir soal dengan benar lebih dominan maka soal tersebut cenderung mudah, begitu pula sebaliknya. Butir soal yang tergolong terlalu mudah atau terlalu sulit maka tidak dapat berfungsi secara maksimal dalam mengidentifikasi peserta didik yang berkemampuan tinggi dan rendah. Apabila daya pembeda suatu butir rendah, maka informasi yang diberikan oleh butir tersebut tidak akurat. Sedangkan daya pengecoh menginformasikan terkait keberfungsian opsi jawaban (Amiruddin, Mania, Ichiana, N., & Majid, A., 2020). Karakteristik butir soal dapat diketahui melalui analisis hasil pengujian butir soal

secara empirik. Terdapat dua pendekatan dalam menganalisis butir soal, yaitu yaitu teori tes klasik dan teori respons butir (Alfarisa & Purnama, 2019; Danni & Tauratiya, 2020; Himelfarb, 2019; Heri Retnawati, 2014).

Teori tes klasik merupakan pendekatan yang beranggapan bahwa skor amatan adalah hasil penjumlahan antara skor murni dan kesalahan pengukuran ($O_i = T_i + E_i$) (Himelfarb, 2019). Teori tes klasik juga menjadi pendekatan yang dominan digunakan (Petrillo, Cano, McLeod, & Coon, 2015; Rusch, Lowry, Mair, & Treiblmaier, 2017). Di Negara Indonesia, pendekatan teori tes klasik telah diaplikasikan pada sebagian besar disiplin ilmu hingga saat ini. Namun faktanya, pendekatan ini memiliki beberapa kelemahan yaitu hasil pengukuran memiliki ketergantungan pada karakteristik butir soal yang diujikan dan karakteristik butir soal bergantung pada kemampuan peserta tes (Culpepper, 2013; Mardapi, 2008; Heri Retnawati, 2014; Sumintono & Widhiarso, 2015; van der Linden & Hambleton, 2013). Sebagai upaya mengatasi kelemahan pendekatan teori tes klasik, berkembanglah pendekatan baru yaitu *item response theory*. Pendekatan ini hadir dengan mengusung tiga asumsi yaitu unidimensi, independensi lokal, dan invariansi parameter (DeMars, 2018). Asumsi tersebut merupakan jawaban atas kelemahan teori tes klasik, sehingga hasil pengukuran tidak lagi bergantung pada karakteristik butir soal yang diujikan karena pengukuran kemampuan didasari oleh peluang peserta tes dalam menjawab butir soal. Peserta tes yang berkemampuan tinggi akan berpeluang besar untuk dapat menjawab soal yang sukar, begitu pula sebaliknya peserta tes yang berkemampuan rendah berpeluang kecil dalam menjawab butir yang sukar.

Pendekatan teori respons butir memiliki tiga model parameter logistik (PL) (Susetyo, 2015). Model parameter yang memuat tingkat kesukaran (b_i) dikenal dengan sebuatan 1 parameter logistik (1PL), model parameter yang memuat tingkat kesukaran (b_i) dan daya pembeda (a_i) dikenal dengan sebutan 2PL, dan model parameter yang memuat tingkat kesulitan (b_i), daya beda (a_i), dan tebakan semu (c_i) dikenal dengan sebutan 3PL. Penentuan model parameter logistik didasari oleh kecocokan butir soal terhadap model (Naga, 1992). Berbanding terbalik dengan teori tes klasik, penggunaan pendekatan teori respons butir masih minim digunakan. Padahal teori respons butir merupakan jawaban atas kelemahan dari teori tes klasik. Informasi yang dihasilkan dari teori respons butir pun lebih akurat dibandingkan teori tes klasik (Santoso et al., 2019). Oleh karena itu, perlu adanya sosialisasi dan panduan dalam pengaplikasian teori respons butir khususnya pada lingkungan pendidikan di Indonesia.

Salah satu sekolah yang menerapkan teori tes klasik dalam penilaian peserta didik adalah Madrasah Aliyah Negeri (MAN) 1 Pangkalpinang di Provinsi Bangka Belitung. Sekolah yang berdiri pada 2016 ini adalah satu-satunya sekolah keagamaan negeri setara Sekolah Menengah Atas di Kota Pangkalpinang. Sekolah yang berada di bawah naungan Kementerian Agama ini merupakan sekolah percontohan bagi sekolah lain khususnya tingkat madrasah Aliyah. Hasil survey prapenelitian dengan guru bahasa Arab MAN 1 Pangkalpinang, diketahui bahwa pengembangan butir soal mata pelajaran bahasa Arab baik pada tes formatif maupun sumatif tidak melalui tahapan uji coba butir soal. Hal tersebut menunjukkan bahwa soal untuk mengukur kemampuan bahasa Arab peserta didik belum dikalibrasi sehingga tidak diketahui kualitasnya namun telah digunakan untuk mengetahui kemampuan bahasa Arab peserta didik. Padahal pengembangan butir soal perlu dilakukan kalibrasi agar diketahui kualitasnya, sehingga butir soal yang digunakan hanya yang berkualitas baik (Retnawati & Hadi, 2014). Soal yang berkualitas baik akan memberikan informasi terkait peserta tes secara akurat, sedangkan butir yang tidak diketahui kualitasnya berpotensi menghasilkan informasi yang tidak sesuai dengan keadaan sebenarnya.

Berdasarkan uraian beberapa literatur dan fenomena yang terjadi di MAN 1 Pangkalpinang, maka diketahui bahwa sejauh ini guru MAN 1 Pangkalpinang masih pada tahap mampu membuat soal. Akan tetapi, belum mampu mengkalibrasi butir soal, sehingga perlu dilakukan pengkalibrasian butir soal agar proses pengembangan butir soal melalui tahapan yang benar. Oleh karena itu, penelitian ini bertujuan untuk mengetahui dan mendeskripsikan karakteristik butir soal bahasa Arab

di MAN 1 Pangkalpinang. Kualitas butir soal dideskripsikan melalui pembuktian validitas, estimasi reliabilitas, dan karakteristik butir soal yang dikalibrasi berdasarkan pendekatan teori respons butir.

Metode Penelitian

Penelitian ini merupakan penelitian deskriptif dengan pendekatan kuantitatif. Penelitian ini bertujuan untuk menganalisis dan mendeskripsikan kualitas butir soal penilaian akhir semester (PAS) bahasa Arab menggunakan pendekatan teori repons butir. Subjek penelitian ini adalah peserta didik kelas XI MAN 1 Pangkalpinang tahun ajaran 2019/2020 yang berjumlah 176 peserta didik. Instrumen pengumpul data berupa butir soal PAS mata pelajaran bahasa Arab berjumlah 40 butir soal berbentuk pilihan ganda dengan lima pilihan jawaban. Validitas instrumen dibuktikan berdasarkan validitas konstruk melalui *exploratory factor analysis* (EFA) dan reliabilitas soal diestimasi menggunakan *formula cronbach's alpha* melalui *software* SPSS v.23. Sedangkan pengkalibrasian butir soal PAS bahasa Arab dilakukan menggunakan pendekatan teori respons butir melalui aplikasi Bilog-MG. Karakteristik butir soal ditentukan berdasarkan kriteria indeks parameter butir, seperti parameter tingkat kesukaran butir (b_i) tergolong baik apabila indeks berada direntang -2 sampai +2, parameter daya pembeda (a_i) di rentang 0 s.d. +2, dan parameter tebakan semu (c_i) di rentang 0 s.d. 1/k (DeMars, 2018).

Hasil dan Pembahasan

Validitas dan Reliabilitas Instrumen PAS bahasa Arab

Pada pembuktian validitas konstruk menggunakan analisis faktor eksploratori, terdapat beberapa asumsi yang menjadi persyaratan, yaitu asumsi ketercukupan sampel yang diketahui melalui nilai *chi-square* pada uji Bartlett dan *Kaiser meyer olkin measure of sampling adequacy* (KMO-MSA), nilai eigen dan *scree plot* menunjukkan jumlah faktor dominan yang terbentuk pada instrumen, dan *component matrix* berisikan *loading factor*. Ketercukupan sampel pada analisis validitas konstruk instrumen PAS bahasa Arab kelas XI MAN 1 Pangkalpinang menggunakan *exploratory factor analysis* (EFA) melalui software SPSS v.22 ditunjukkan dengan hasil KMO-MSA dan uji Bartlett pada Tabel 1.

Tabel 1. KMO dan Uji Bartlett

Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		0,687
Bartlett's Test of Sphericity	Approx. Chi-Square	1426,784
	Df	780
	Sig.	0,000

Tabel 1 menunjukkan bahwa instrumen PAS bahasa Arab yang diujikan pada 176 peserta didik memperoleh nilai KMO 0,687 yang lebih besar dari 0,50 dan *p value* 0,000 pada uji barltett lebih kecil dari 0,05. Nilai KMO lebih dari 0,50 menunjukkan bahwa ukuran sampel telah tercukupi untuk analisis faktor (Nurmalita, 2018). Dengan demikian, dapat dikatakan bahwa ukuran sampel sebesar 176 peserta didik telah memenuhi syarat ketercukupan analisis factor dan pembuktian validitas konstruk instrumen PAS bahasa Arab menggunakan analisis faktor dapat dilanjutkan. Faktor yang terbentuk pada instrumen dapat diketahui melalui nilai eigen yang lebih besar dari 1 (Wagiran, 2013). Hasil analisis menunjukkan bahwa terdapat 16 komponen yang memiliki nilai eigen > 1, berikut 5 komponen teratas ditampilkan pada Tabel 2.

Tabel 2. Nilai Eigen Instrumen PAS bahasa Arab

Komponen	Total	% varian	Kumulatif %
1	5,767	14,419	14,419
2	1,976	4,940	19,359
3	1,695	4,237	23,596

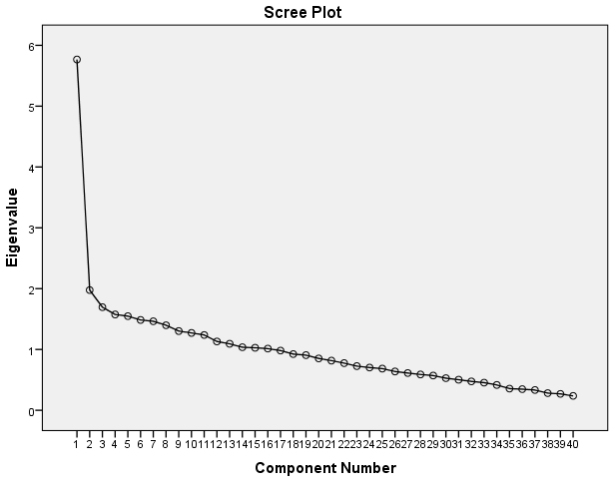
4	1,575	3,938	27,534
5	1,548	3,869	31,404

Muatan faktor yang diketahui melalui *component matrix* pada output SPSS v.22 menginformasikan bahwa semua butir soal memiliki besaran *loading factor* > 0,3 sehingga dapat disimpulkan bahwa 40 butir soal PAS bahasa Arab tergolong valid. *Loading factor* > 0,3 menunjukkan bahwa butir mengukur dimensi yang hendak diukur (Hair, Black, Babin, & Anderson, 2010; Nurosis, 1986, p. 12).Muatan faktor terendah dimiliki butir nomor 23 yaitu sebesar 0,351 dan tertinggi dengan muatan faktor sebesar 0,570 dimiliki oleh butir nomor 7. Diketahui pula berdasarkan muatan faktor tertinggi terdapat 30 butir soal (75%) mengukur komponen 1 dan 10 butir (25%) tersebar pada komponen lainnya. Hal ini menunjukkan bahwa terdapat 1 faktor dominan yang diukur oleh instrumen PAS bahasa Arab. Di samping itu, berdasarkan hasil estimasi reliabilitas instrumen PAS bahasa Arab kelas XI MAN 1 Pangkalpinang memperoleh koefisien reliabilitas sebesar 0,8442. Hasil tersebut menunjukkan bahwa instrument PAS bahasa Arab tergolong reliabel karena memiliki koefisien reliabilitas > 0,7. Instrumen memenuhi kriteria reliabilitas apabila memperoleh koefisien reliabilitas > 0,7 (Azwar, 2015).

Pengkalibrasian butir soal menggunakan pendekatan teori respons butir memerlukan uji asumsi, yaitu meliputi asumsi unidimensi, independensi lokal, dan invariansi parameter. Adapun hasil uji asumsi pada instrumen PAS bahasa Arab dipaparkan sebagai berikut.

Asumsi Unidimensi dan Independensi Lokal

Asumsi unidimensi dapat diketahui melalui dua cara, yaitu dengan melihat nilai eigen atau kecuraman pada *scree plot* (Heri Retnawati, 2016). Nilai eigen soal PAS bahasa Arab ditampilkan pada Tabel 2 dan *scree plot* pada Gambar 1.



Gambar 1. Nilai eigen soal PAS bahasa Arab

Tabel 2 menginformasikan bahwa nilai eigen pada komponen 1 adalah 5,767 dan komponen 2 sebesar 1,976. Apabila nilai eigen komponen 1 dibandingkan dengan komponen 2 terdapat selisih yang cukup jauh, sedangkan komponen 2 dengan komponen 3 (1,695) berjarak dekat hal ini menunjukkan bahwa soal PAS bahasa Arab mengukur 1 faktor dominan. Apabila nilai eigen pertama memiliki selisi nilai berkali kali dari komponen ke dua dan nilai eigen antar komponen selanjutnya hampir sama, maka asumsi unidimensi terpenuhi (Susetyo, 2015, p. 72).

Di samping itu, *scree plot* yang terbentuk berdasarkan nilai eigen memperkuat simpulan tersebut. Tampak pada Gambar 1 bahwa komponen 1 dengan komponen 2 membentuk kecuraman yang jauh, sedangkan komponen 2 ke 3 pendek dan landai. Banyaknya curaman pada *scree plot* menunjukkan banyaknya dimensi dominan dan bentuk landai tidak menunjukkan adanya dimensi (Heri Retnawati, 2016). Dengan demikian, berdasarkan nilai eigen dan diperkuat oleh *scree plot*

yang terbentuk maka dapat disimpulkan bahwa soal PAS bahasa Arab kelas XI MAN 1 Pangkalpinang bersifat unidimensi.

Selain unidimensi, terdapat pula asumsi yang harus terpenuhi pada pendekatan teori respons butir, yaitu independensi lokal. Naga (1992) menyebutkan bahwa independensi lokal merupakan asumsi yang mensyaratkan skor butir suatu tes tidak bergantung pada butir lainnya. Asumsi independensi lokal akan terpenuhi apabila suatu tes terbukti bersifat unidimensi (DeMars, 2018; Santoso et al., 2019). Dengan demikian, soal PAS bahasa Arab MAN 1 Pangkalpinang memenuhi asumsi independensi lokal karena telah terbukti bersifat unidimensi.

Kecocokan Butir Terhadap Model Logistik

Hambleton, R. K. Swaminathan & Rogers (1991) mengungkapkan bahwa pendekatan teori respons butir dapat digunakan apabila terdapat kecocokan antara model logistik dengan data tes. Pendekatan teori respons butir menawarkan tiga model logistik, yaitu 1 PL memuat parameter tingkat kesulitan (b_i), 2 PL memuat parameter tingkat kesulitan (b_i) dan daya pembeda (a_i), dan 3 PL memuat parameter tingkat kesulitan (b_i), daya beda (a_i), dan tebakan semu (c_i). Apabila parameter yang digunakan banyak, maka semakin rinci model logistik menginformasikan kemampuan peserta tes. Kecocokan butir soal terhadap model logistik dapat ditentukan dengan membandingkan *chi square* (χ^2) hitung dengan *chi square* (χ^2) tabel (Heri Retnawati, 2014). Selain itu, dapat pula diketahui dengan melihat nilai probability (*P value*), apabila *P value* lebih besar dibandingkan alpha (0,05) maka butir soal cocok terhadap model (*fit model*) (Danni, 2018). Ringkasan hasil uji kecocokan butir soal PAS bahasa Arab kelas XI MAN 1 Pangkalpinang dengan membandingkan *P value* $> 0{,}05$ ditampilkan pada Tabel 3.

Tabel 3. Ringkasan Hasil Uji Kecocokan Butir Soal Terhadap Model Logistik

Model Logistik	Butir Cocok	Butir Tidak Cocok
1PL	24	16
2PL	33	7
3PL	28	12

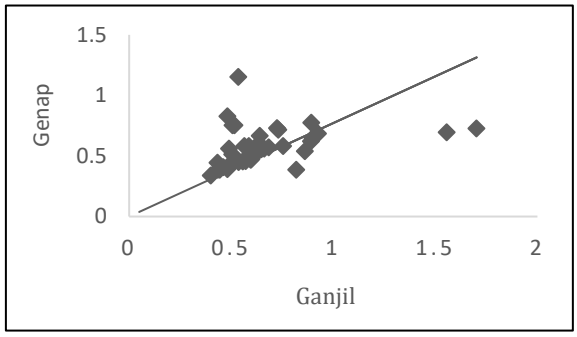
Hasil uji kecocokan butir sebagaiman ditampilkan pada Tabel 3 mengeinformasikan bahwa soal PAS bahasa Arab kelas XI MAN 1 Pangkalpinang cocok terhadap model logistik 2 PL. Hal ini dibuktikan dengan adanya 33 (82,5%) butir soal yang cocok dengan model 2 PL. Dengan demikian, kalibrasi butir soal yang digunakan adalah model 2PL sehingga memuat parameter tingkat kesukaran (b_i) dan daya pembeda (a_i).

Asumsi Invariansi Parameter

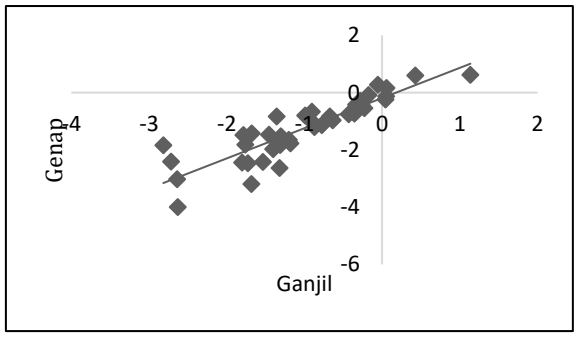
Invariansi parameter merupakan asumsi yang mengharuskan parameter suatu butir tidak tergantung pada peserta tes dan begitu juga sebaliknya parameter siswa tidak memiliki ketergantungan dengan parameter butir (Duskri, Kumaidi, & Suryanto, 2014). Oleh karena itu, invariansi parameter dapat diketahui berdasarkan invariansi parameter butir dan kemampuan peserta tes (Alfarisa & Purnama, 2019; Heri Retnawati, 2016; van der Linden & Hambleton, 2013). Invariansi parameter butir dapat diketahui melalui pembagian peserta tes ke dalam dua kelompok ganjil dan genap atau laki-laki dan perempuan kemudian diestimasi secara tepisah lalu dibentuk *scree plot* antar keduanya. Apabila sebaran parameter mendekati garis diagonal (y=x) maka dapat dikatakan bahwa parameter butir telah memenuhi asumsi invariansi parameter (Heri Retnawati, 2014).

Selain itu, pembuktian invariansi parameter dapat juga dilakukan dengan mengkorelasikan antar kelompok, apabila antar kelompok memiliki hubungan yang kuat maka dapat disimpulkan bahwa parameter bersifat invarian (Susetyo, 2015). Pembuktian asumsi invariansi parameter pada penelitian ini menggunakan *scree plot* sebaran parameter butir.

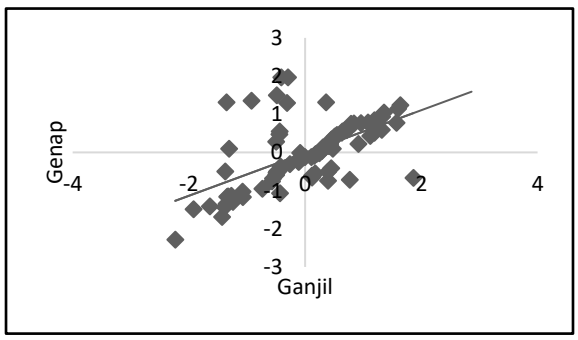
Pengujian invariansi parameter disesuaikan dengan kecocokan butir terhadap model logistik. Dikarenakan butir soal bahasa Arab kelas XI MAN 1 Pangkalpinang cocok terhadap model 2PL maka pengujian invariansi parameter dilakukan berdasarkan parameter tingkat kesukaran (b_i), daya pembeda (a_i), dan kemampuan peserta (θ). Hasil uji invariansi parameter butir dan kemampuan peserta pada PAS bahasa Arab melalui pembagian peserta tes ganjil dan genap digambarkan pada Gambar 2, Gambar 3, dan Gambar 4.



Gambar 2. Sebaran Parameter Daya Pembeda (a_i)



Gambar 3. Sebaran Parameter Tingkat Kesukaran (b_i)



Gambar 4. Sebaran Kemampuan Peserta Tes (θ)

Hasil uji invariansi parameter butir dan kemampuan peserta tes pada soal PAS bahasa Arab kelas XI MAN 1 Pangkalpinang melalui pembagian peserta tes ganjil dan genap menunjukkan bahwa sebaran parameter tingkat kesukaran (b_i), daya pembeda (a_i), dan kemampuan peserta (θ) baik kelompok ganjil maupun genap mendekati garis diagonal. Hasil tersebut menunjukkan bahwa parameter tingkat kesukaran, daya pembeda, dan kemampuan peserta bersifat invarian. Hal ini ditegaskan oleh Heri Retnawati (2014) bahwa apabila sebaran parameter mendekati garis diagonal (y=x) maka parameter bersifat invarian.

Karakteristik Butir Soal

Hasil kalibrasi butir soal PAS bahasa Arab kelas XI MAN 1 Pangkalpinang menggunakan pendekatan teori respons butir dengan model 2PL menginformasikan parameter tingkat kesulitan (b_i) dan daya beda (a_i). Hasil kalibrasi menunjukkan butir soal dengan indeks kesukaran terendah

adalah butir nomor 29 dengan indeks sebesar -3,164 dalam *logit*, sedangkan butir soal tertinggi adalah nomor 31 dengan *logit* 0,868. Terdapat 33 butir (82,5%) memiliki indeks kesukaran berkategori baik dan 7 butir (17,5%) berkategori kurang baik karena memiliki tingkat kesukaran berkategori sangat mudah sehingga perlu direvisi/diperbaiki, yaitu butir nomor 10, 26, 27, 29, 32, 34, dan 35. Tingkat kesukaran butir yang baik di rentang -2 sampai dengan +2 dalam skala *logit* (DeMars, 2018; Hambleton & Swaminathan, 1985; Mardapi, 1998; Sumintono & Widhiarso, 2015). Sedangkan butir yang memiliki indeks kesukaran diluar rentang tersebut dapat diperbaiki atau dieliminasi (Heri Retnawati, 2014). Ditegaskan pula oleh Susetyo (2015) bahwa butir dengan rentang mendekati *logit* +2 maka butir cenderung lebih sulit, sebaliknya apabila butir mendekati -2 relatif mudah, dan direntang $-1{,}0 < b < +1{,}0$ butir tergolong sedang. Secara kesluruhan tingkat kesukaran butir soal PAS bahasa Arab tergolong mudah dengan rata-rata indeks kesukaran sebesar -1,16 dalam logit dengan rincian 18 butir (45%) tergolong sedang, 15 butir (37,5%) tergolong mudah, dan 7 butir (17,5%) tergolong sangat mudah.

Selain karakteristik tingkat kesukaran, hasil kalibrasi juga menginformasikan karakteristik daya pembeda butir soal. Hasil kalibrasi menunjukkan daya pembeda terendah diperoleh oleh butir soal nomor 22 dengan indeks 0,374 dalam *logit*, sedangkan indeks tertinggi diperoleh butir nomor 7 dengan *logit* 1,107. Karakteristik daya pembeda soal PAS bahasa Arab kelas XI MAN 1 Pangkalpinang tergolong baik dengan indeks rata-rata sebesar 0,578 dan indeks daya pembeda 40 butir soal (100%) berada di rentang 0 sampai dengan +2 dalam logit. Sebagaimana yang dinyatakan DeMars (2018) bahwa daya pembeda yang baik memiliki rentang 0 s.d. +2. Dengan demikian, butir soal PAS bahasa Arab kelas XI MAN 1 Pangkalpinang dapat dikatakan mampu membedakan peserta didik berkemampuan tinggi dan berkemampuan rendah. Ringkasan hasil kalibrasi butir soal PAS bahasa Arab melalui Bilog MG dan materi pokok serta level kognitif butir berdasarkan kisi-kisi soal ditampilkan pada Tabel 4.

Tabel 4. Karakteristik Butir Soal Bahasa Arab Kelas XI MAN 1 Pangkalpinang

Butir	Materi Pokok	Level Kognitif	Tingkat Kesukaran	Daya Pembeda	Keterangan
1	تطور المدرسة اللإسلمية	C1	0,699	-0,279	Baik
2	تطور المدرسة اللإسلمية	C1	0,459	-1,594	Baik
3	تطور المدرسة اللإسلمية	C1	0,699	0,494	Baik
4	تطور المدرسة اللإسلمية	C1	0,528	-0,771	Baik
5	تطور المدرسة اللإسلمية	C1	0,492	-1,767	Baik
6	تطور المدرسة اللإسلمية	C1	0,461	-1,677	Baik
7	تطور المدرسة اللإسلمية	C1	1,107	-1,407	Baik
8	تطور المدرسة اللإسلمية	C1	0,800	-1,404	Baik
9	تطور المدرسة اللإسلمية	C1	0,484	-0,620	Baik
10	تطور المدرسة اللإسلمية	C1	0,444	-3,010*	Kurang Baik
11	تطور المدرسة اللإسلمية	C1	0,815	-0,940	Baik
12	تطور المدرسة اللإسلمية	C1	0,530	-1,033	Baik
13	تطور المدرسة اللإسلمية	C1	0,741	-1,782	Baik
14	تطور المدرسة اللإسلمية	C1	0,606	-0,389	Baik
15	تطور المدرسة اللإسلمية	C1	0,558	-1,129	Baik
16	تطور المدرسة اللإسلمية	C1	0,533	-1,462	Baik
17	تطور المدرسة اللإسلمية	C1	0,577	-0,815	Baik
18	تطور المدرسة اللإسلمية	C1	0,421	-1,802	Baik
19	تطور المدرسة اللإسلمية	C1	0,699	-1,561	Baik
20	تطور المدرسة اللإسلمية	C1	0,484	-1,489	Baik
21	مراحل التعلم	C1	0,508	-0,377	Baik
22	مراحل التعلم	C2	0,374	0,113	Baik
23	مراحل التعلم	C2	0,494	-0,353	Baik
24	مراحل التعلم	C2	0,535	-0,036	Baik
25	مراحل التعلم	C2	0,626	-0,771	Baik

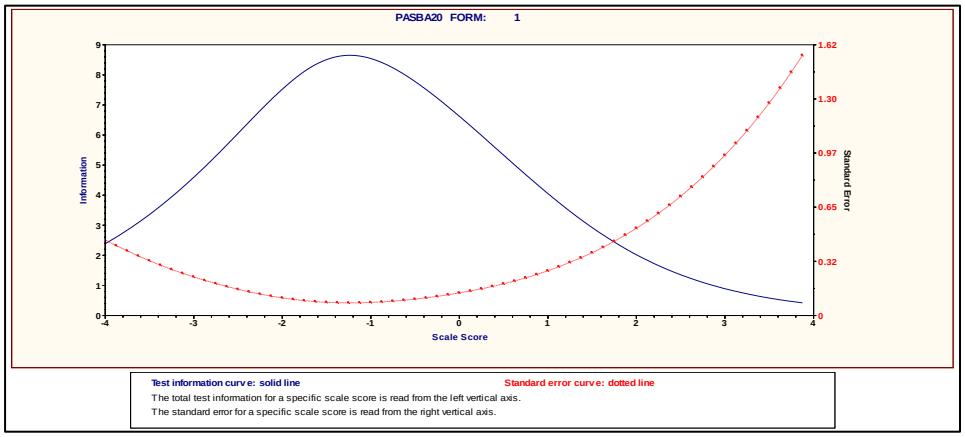
Butir	Materi Pokok	Level Kognitif	Tingkat Kesukaran	Daya Pembeda	Keterangan
26	مراحل التعلم	C2	0,538	-2,334*	Kurang Baik
27	مراحل التعلم	C2	0,568	-2,418*	Kurang Baik
28	مراحل التعلم	C2	0,612	-0,088	Baik
29	مراحل التعلم	C2	0,607	-3,164*	Kurang Baik
30	مراحل التعلم	C2	0,816	-1,767	Baik
31	مراحل التعلم	C2	0,575	0,868	Baik
32	مراحل التعلم	C2	0,490	-2,162*	Kurang Baik
33	مراحل التعلم	C2	0,396	-1,945	Baik
34	مراحل التعلم	C2	0,677	-2,084*	Kurang Baik
35	مراحل التعلم	C2	0,443	-2,862*	Kurang Baik
36	مراحل التعلم	C2	0,720	0,114	Baik
37	مراحل التعلم	C2	0,481	-0,965	Baik
38	مراحل التعلم	C2	0,449	-0,141	Baik
39	مراحل التعلم	C1	0,608	-0,583	Baik
40	مراحل التعلم	C1	0,475	-1,011	Baik

* $b < -2$

Tabel 4 juga menginformasikan soal PAS bahasa Arab kelas XI MAN 1 Pangkalpinang masih mengukur level kognitif tingkat rendah. Hal tersebut ditunjukkan dari level kognitif C1 (mengetahui) dan C2 (memahami) yang mana level tersebut adalah bagian dari kemampuan berpikir tingkat rendah *(Lower order thinking skills)* dalam taksonomi Bloom (Anderson & Krathwohl, 2001; Fan & Yan, 2020). Padahal dalam pembelajaran bahasa Arab model berpikir tingkat tinggi seperti berpikir kreatif, kritis, dan inovatif perlu ditekankan (Wahab, 2015). Pengembangan butir soal berbasis *higher order thinking skills (HOTS)* memang bukanlah suatu hal yang mudah (Ramadhan et al., 2019). Oleh karena itu, pendidik bahasa Arab maupun *stakeholder* diharapkan dapat menerapkan pembelajaran dan penilaian berbasis *HOTS* khususnya di MAN 1 Pangkalpinang. Fenomena ini menegaskan bahwa pembelajaran bahasa Arab masih memiliki permasalahan dan tantangan yang harus dihadapi (Albantani & Madkur, 2019).

Fungsi Informasi

Fungsi informasi pada pendekatan teori respons butir berguna untuk menunjukkan kontribusi butir dalam mengestimasi kemampuan peserta tes (Nurcahyo, 2017). Retnawati (2016) menyatakan bahwa instrumen yang berkualitas akan selalu menghasilkan nilai informasi yang lebih tinggi dibandingkan kesalahan pengukuran. Oleh karena itu, fungsi informasi akan selalu berbanding terbalik dengan kesalahan pengukuran *(standard error of measurement)*. Semakin tinggi informasi yang diberikan suatu tes, maka semakin kecil kesalahan pengukurannya. Hubungan fungsi informasi dengan *standard error of measurement (SEM)* sering digambarkan dalam bentuk kurva. Adapun fungsi informasi soal bahasa Arab kelas XI MAN 1 Pangkalpinang ditampilkan pada Gambar 5.



Gambar 5. Fungsi informasi soal PAS bahasa Arab

Gambar 5 menunjukkan bahwa soal PAS bahasa Arab kelas XI MAN 1 Pangkalpinang akan memberikan informasi maksimum apabila diujikan pada peserta didik dengan kemampuan (θ) -3,5 sampai dengan +1,5 dalam *logit*. Sedangkan fungsi informasi tertinggi sebesar 8,647 dengan *SEM* 0,35 pada kemampuan (θ) peserta tes -1,25 dalam *logit*.

Simpulan

Berdasarkan hasil penelitian tersebut, dapat disimpulkan bahwa soal PAS bahasa Arab kelas XI MAN 1 Pangkalpinang terbukti valid ditunjukkan dengan 40 butir soal (100%) memiliki *loading factor* > 0,3 dan terbukti reliabel ditunjukkan dengan koefisien reliabilitas > 0,7 yaitu 0,884. Selain itu, tingkat kesukaran soal PAS bahasa Arab tergolong mudah dengan rata-rata indeks kesukaran sebesar -1,16 dalam *logit*, dengan rincian 33 butir soal (82,5%) yang memiliki tingkat kesukaran berkategori baik sehingga dapat disimpan dalam bank soal dan digunakan pada penilaian selanjutnya, sedangkan 7 butir soal (17,5%) yaitu butir nomor 10, 26, 27, 29, 32, 34, dan 35 tidak memenuhi kriteria tingkat kesukaran yang baik sehingga perlu direvisi atau dieliminasi. Daya pembeda 40 butir soal (100%) bahasa Arab tergolong baik karena berada pada rentang 0 s.d. +2 dengan rata-rata indeks daya pembeda sebesar 0,578 dalam *logit*. Soal PAS bahasa Arab akan memberikan informasi maksimum apabila diujikan pada peserta didik dengan kemampuan (θ) rendah hingga sedang yaitu dalam rentang -3,5 sampai dengan +1,5 dalam *logit*. Penelitian ini terbatas pada analisis butir soal pilihan ganda, sehingga kualitas tes uraian belum tergambarkan. Penelitian selanjutnya diharapkan mampu melakukan analisis butir soal bahas Arab bentuk tes uraian dalam skala luas atau mengembangkan butir soal bahasa Arab berbasis *higher order thinking skills* yang berkualitas.[]

Daftar Rujukan

Abdul Wahab, Muhbib. 2015. "Pembelajaran Bahasa Arab di Era Posmetode", *Arabiyat : Jurnal Pendidikan Bahasa Arab dan Kebahasaaraban*, Vol. 2, No. 1.

Albantani, A. M., & Madkur, A. 2019. "Teaching Arabic in the era of Industrial Revolution 4.0 in Indonesia: Challenges and opportunities", *ASEAN Journal of Community Engagement*, Vol. 3, No. 2.

Alfarisa, F., & Purnama, D. N. 2019. "Analisis Butir Soal Ulangan Akhir Semester Mata Pelajaran Ekonomi SMA Menggunakan RASCH Model", *Jurnal Pendidikan Ekonomi Undiksha*, Vol. 11, No. 2.

Allen, M. J., & Yen, W. M. 1979. *Introduction to measurement theory*. Monterey: Books Cole Publishing Company.

Amiruddin, K., Mania, S., Ichiana, N., N., & Majid, A., F. 2020. "Analisis Butir Soal Ujian Akhir Sekolah (UAS) Mata Pelajaran Matematika", *Alauddin Journal of Mathematics Education*, Vol. 2, No. 2.

Anderson, L. W., & Krathwohl, D. R. 2001. *A Taxonomy for Learning Teaching and Assessing: A Revision of Bloom's Taxonomy of Education Objectives*. New York: Addison Wesley Longman, Inc.

Anita, A., Tyowati, S., & Zuldafrial, Z. 2018. "Analisis Kualitas Butir Soal Fisika Kelas X Sekolah Menengah Atas", *Edukasi: Jurnal Pendidikan*, Vol. 16, No. 1.

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

Culpepper, S. A. 2013. "The reliability and precision of total scores and IRT estimates as a function of polytomous IRT parameters and latent trait distribution", *Applied Psychological Measuremen*, Vol. 37, No. 3.

Danni, R. 2018. "Pengembangan Tes Prestasi Belajar Bahasa Arab Menggunakan Pendekatan Item

Response Theory", *The 1st National Conference of Genuine Psychology; Understanding the Meaning of Being Human from the Perspective of Health Psychology*, Fakultas Psikologi UIN Raden Fatah Palembang.

Danni, R., & Tauratiya, T. 2020. "Analisis Kemampuan Berpikir Kritis Mahasiswa Program Studi Hukum Keluarga Islam IAIN Syaikh Abdurrahman Siddik Bangka Belitung", *Tarbawy : Jurnal Pendidikan Islam*, Vol. 7, No. 1.

DeMars, C. E. 2018. "Classical Test Theory and Item Response Theory", *The Wiley Handbook of Psychometric Testing*.

Duskri, M., Kumaidi, K., & Suryanto, S. 2014. "Pengembangan Tes Diagnostik Kesulitan Belajar Matematika di SD", *Jurnal Penelitian dan Evaluasi Pendidikan*, Vol. 18, No. 1.

Fan, T., & Yan, X. 2020. "Diagnosing English reading ability in Chinese senior high schools", *Studies in Educational Evaluation*, Vol. 67.

Hair, J. F. J., Black, W. C., Babin, B. J., & Anderson, R. E. 2010. *Multivariate data analysis 7th edition*. New York: Pearson Prentice Hall.

Hambleton, R. K. Swaminathan, H., & Rogers, H. J. 1991. *Fundamentals of Item Response Theory*. CA: Sage Publication Inc.

Hambleton, R. K., & Swaminathan, H. 1985. *Item Response Theory: Principles and Application*. Boston: Kluwer Inc.

Himelfarb, I. 2019. "A primer on standardized testing: History, measurement, classical test theory, item response theory, and equating", *Journal of Chiropractic Education*, Vol. 33.

Iskandar, A., & Rizal, M. 2018. "Analisis Kualitas Soal di Perguruan Tinggi Berbasis Aplikasi TAP", *Jurnal Penelitian Dan Evaluasi Pendidikan*, Vol. 22, No. 1.

Mardapi, D. 1998. "Analisis Butir dengan Teori Tes Klasik dan Teori Respons Butir", *Jurnal Kependidikan*, Vol. 28.

Mardapi, D. 2008. *Teknik Penyusunan Instrumen Tes dan Nontes*. Yogyakarta: Mitra Cendikia Offset.

Mardapi, D. 2016. *Pengukuran penilaian dan evaluasi pendidikan* (2nd ed.). Yogyakarta: Nuha Litera.

Miller, Linn, & Grounlund. 2009. *Measurement and assessment in teaching*. Englewood Cliffs: Prentice-Hall.

Naga, D. S. 1992. *Pengantar teori sekor pada pengukuran pendidikan*. Jakarta: Gunadarma.

Nurcahyo, F. A. 2017. "Aplikasi IRT dalam Analisis Aitem Tes Kognitif", *Buletin Psikologi*, Vol. 24, No. 2.

Nurmalita, C. 2018. "Karakteristik Perangkat Ujian Kompetensi Teori Teknik Perkayuan Tahun 2014 SMK di Jawa Timur Menggunakan Model IRT", *Ilmu Pendidikan: Jurnal Kajian Teori Dan Praktik Kependidikan*, Vol. 3, No. 2.

Nurosis, J. M. 1986. *SPSS/PC+for the imbbc/xt/ at*. Chicago: SPSS Inc.

Pauw, I. 2015. "Educating for the future: The position of school geography", *International Research in Geographical and Environmental Education*, Vol. 24, No. 4.

Petrillo, J., Cano, S. J., McLeod, L. D., & Coon, C. D. 2015. "Using classical test theory, item response theory, and rasch measurement theory to evaluate patient-reported outcome measures: A comparison of worked examples", *Value in Health*, Vol. 18, No. 1.

Prayogi, R. D., & Estetika, R. 2019. "Kecakapan Abad 21: Kompetensi Digital Pendidik Masa Depan", *Jurnal Manajemen Pendidikan*, Vol. 14, No. 2.

Ramadhan, S., Mardapi, D., Prasetyo, Z. K., & Utomo, H. B. 2019. "The development of an instrument to measure the higher order thinking skill in Physics", *European Journal of Educational Research*, Vol. 8, No. 3.

Retnawati, H. 2014. *Teori Respons Butir dan Penerapannya*. Yogyakarta: Nuha Medika.

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

Retnawati, H., & Hadi, S. 2014. "Sistem bank soal daerah terkalibrasi untuk menyongsong era desentralisasi", *Jurnal Ilmu Pendidikan*, Vol. 20, No. 2.

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

Rindermann, H., & Baumeister, A. E. E. 2015. "Validating the interpretations of PISA and TIMSS tasks: A rating study", *International Journal of Testing*, Vol. 15, No. 1.

Rusch, T., Lowry, P. B., Mair, P., & Treiblmaier, H. 2017. "Breaking free from the limitations of classical test theory: Developing and measuring information systems scales using item response theory", *Information and Management*, Vol. 54, No. 2.

Santoso, A., Kartianom, K., & Kassymova, G. K. 2019. "Kualitas butir bank soal statistika (Studi kasus: Instrumen ujian akhir mata kuliah statistika Universitas Terbuka)", *Jurnal Riset Pendidikan Matematika*, Vol. 6, No. 2.

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

Sumintono, B., & Widhiarso, W. 2015. *Aplikasi Pemodelan Rasch pada Assessment Pendidikan*. Bandung: Trim Komunikata.

Susetyo, B. 2015. *Prosedur Penyusunan dan Analasis Tes*. Bandung: PT Refika Aditama.

Suwandi, S. 2010. *Model Assesmen dalam Pembelajaran* (2nd ed.). Surakarta: Yuma Pustaka.

Trilling, B., & Fadel, C. 2009. *21st Century skills.* San Fransisco: John Wiley & Sons, Inc.

van der Linden, W. J., & Hambleton, R. K. 2013. *Handbook of modern item response theory*. New York: Springer Science & Business Media.

Wagiran. 2013. *Metodologi Penelitian Pendidikan (Teori dan Implementasi)*. Yogyakarta: Deepublish.