



PROPERTI PSIKOMETRIK *INTELLIGENZ* STRUKTUR TEST SUBTES KEMAMPUAN NUMERIK (*RECHENAUFGABEN* DAN *ZAHLEN REIHEN*)

Medianta Tarigan^{1✉}, Fadillah²

Departemen Psikologi, Universitas Pendidikan Indonesia

Info Artikel

Sejarah Artikel:
Disubmit 31 Agustus 2021
Direvisi 9 September 2021
Diterima 12 November 2021

Keywords:

Intelligenz Struktur Test (IST), Item Response Theory (IRT), Number Factor (N).

Abstrak

Salah satu tes intelegensi yang banyak digunakan untuk tujuan seleksi adalah Intelligenz Struktur Test (IST). Penelitian ini bertujuan untuk mengkaji ulang apakah IST masih berfungsi dalam menghasilkan pengukuran intelegensi yang berkualitas. Dalam penelitian ini, diuji properti psikometrik subtes IST yang mengukur kemampuan numerik, yaitu subtes Rechenaufgaben (RA) dan Zahlen Reihen (ZR). Sebanyak 3,662 partisipan terlibat sebagai sampel penelitian. Metode penelitian yang digunakan adalah analisis aitem dengan Teori Respon Butir 3PL, validasi *convergent-discriminant*, dan uji reliabilitas dengan *split-half*. Hasil analisis aitem berdasarkan estimasi parameter menunjukkan bahwa aitem-aitem pada subtes RA dan ZR termasuk kategori baik. Hasil validasi *convergent-discriminant* menunjukkan subtes RA berkorelasi positif dengan subtes ZR. Adapun uji reliabilitas menunjukkan bahwa *error* yang terjadi pada subtes RA dan ZR cenderung kecil, yaitu 21.1% dan 32.5% berturut-turut untuk subtes RA dan ZR. Hal ini berarti bahwa aitem-aitem pada subtes RA dan ZR mampu memberikan stimulasi kepada partisipan tes untuk mengukur kemampuan numeriknya. Akan tetapi, perlu dilakukan revisi terhadap 17.5% aitem soal yang terjangkit DIF.

Abstract

One of intelligence test that is widely used for selection purposes is the Intelligenz Structure Test (IST). This study aims to review whether IST is still functioning in producing quality intelligence measurements. In this study, the psychometric property of the IST sub-test was tested which measured numerical ability, namely the Rechenaufgaben (RA) and Zahlen Reihen (ZR). Total 3,662 participants were involved as the sample. The research method used is item analysis with item response theory 3PL (IRT-3PL), convergent-discriminant validation, and split-half reliability test. The results of the analysis of the item based on the parameter estimation show that the items in the RA and ZR subtest are in the good category. The results of the reliability showed that the error occurred on subtest RA and ZR tend to be small, ie 21.1% for the subtest RA and 32.5% for the subtest ZR, this means that the items on RA and ZR are still able to provide stimulation to the participants of the test to measure their numerical ability. However, it is necessary to revise 17.5% of the items affected by DIF.

© 2021 Universitas Negeri Semarang

✉ Alamat korespondensi:
Departemen Psikologi, Universitas Pendidikan Indonesia
medianta@upi.edu

PENDAHULUAN

Salah satu faktor yang menyebabkan manusia dapat berbuat sesuatu dengan cara tertentu adalah karena sejak lahir manusia memiliki kemampuan bawaan yang disebut inteligensi. Syah (2005) menyebutkan salah satu definisi intelegensi adalah kemampuan menyesuaikan diri dengan situasi baru secara cepat dan efektif. Sehingga ia menyimpulkan bahwa inteligensi dapat disinonimkan dengan kecerdasan. Tes inteligensi yang merupakan bagian dari tes psikologi memiliki syarat mutlak bahwa tes tersebut haruslah berkualitas secara psikometri. Kajian mengenai properti psikometri sebuah alat tes menjadi sangat esensial, karena kunci utama dari alat ukur dalam ilmu psikologi adalah soal presisi (validitas) dan konsistensi hasil pengukuran (reliabilitas) (Crocker & Algina, 1986).

Tes intelegensi yang masih sering digunakan di Indonesia untuk tujuan seleksi adalah *Intelligenz Struktur Test* (IST). IST menjadi salah satu alat tes intelegensi yang masih sering digunakan karena IST dianggap memiliki subtes yang beragam. Selain itu IST dianggap mampu memberikan informasi yang cukup lengkap terkait kemampuan verbal, numerikal, figural dan daya ingat. Sebagai alat ukur inteligensi, IST menghasilkan skor mean dalam konteks inteligensi yang disebut dengan *Intelligence Quotient* (IQ) (Adinugroho, 2016). Akan tetapi, seiring berjalannya waktu banyak faktor yang dapat mempengaruhi kualitas dari suatu tes. Terlebih usia IST yang sudah sangat tua tetapi masih marak digunakan. Maka diperlukan pengkajian ulang apakah IST masih berfungsi dalam menghasilkan pengukuran inteligensi yang berkualitas sesuai tujuan awal IST disusun.

Salah satu faktor utama yang membentuk IST adalah faktor *Number* (N). Dimana faktor ini mengukur dimensi kecepatan dan akurasi dalam penyelesaian problem matematika. Dalam IST, faktor *Number* (N) ini dapat diukur melalui subtes *Rechenaufgaben* (RA) dan *Zahlen Reihen*

(ZR). Secara umum subtes *Rechenaufgaben* (RA) mengukur kemampuan memecahkan masalah praktis dengan berhitung. Sedangkan *Zahlen Reihen* (ZR) secara umum mengukur kemampuan berhitung yang didasarkan pada pendekatan analitis atas informasi aktual dalam bentuk angka. Sehingga ditemukan hubungan antara angka-angka tersebut. Subtes *Rechenaufgaben* (RA) dan *Zahlen Reihen* (ZR) menjadi aspek IST yang perlu dievaluasi. Hal ini disebabkan karena kedua subtes tersebut mengukur kemampuan numerik yang sangat penting dalam kehidupan individual. Kemampuan numerik berguna untuk mengukur kecerdasan individu karena kemampuan ini dapat mempengaruhi pola berfikir seseorang dalam memecahkan masalah. Bahkan setiap individu dapat menggunakan kemampuan numerik pada berbagai situasi dalam kehidupan sehari-hari. Misalnya ketika membuat keputusan, berurusan dengan informasi numerik, atau mencoba menilai relevansi angka. Kemampuan ini menjadi sangat penting dalam dunia kerja dalam konteks revolusi digital dimana kemampuan analisis data kuantitatif juga menjadi vital dalam pengambilan keputusan. Hal ini menjadi alasan kuat bahwa keterampilan berhitung dan memahami data secara kuantitatif diperlukan untuk menyelesaikan berbagai pekerjaan (Jonas, 2018).

Butir soal yang berkualitas bisa didapatkan dengan melakukan analisis butir soal (Nayla & Widayati, 2012). Ada dua jenis metode yang biasa dilakukan untuk melakukan analisis butir soal, yaitu teori tes klasik dan teori respon butir. Menurut Mardapi (2017), penggunaan teori tes klasik atau yang biasa disebut *Classical Test Theory* (CTT) dalam menganalisis butir memiliki beberapa kelemahan. Diantaranya: statistik butir tes berupa tingkat kesukaran dan daya beda butir soal sangat tergantung kepada karakteristik partisipan tes; estimasi kemampuan partisipan tergantung kepada butir soal yang diujikan;

tidak ada informasi tentang respon setiap partisipan tes terhadap tiap butir soal; dan estimasi keandalan alat tes (reliabilitas) sangat tergantung pada data skor hasil tes.

Teori respon butir muncul untuk mengatasi kelemahan teori tes klasik. Perilaku seseorang dapat dijelaskan oleh karakteristik orang yang bersangkutan sampai pada batas-batas tertentu. Karakteristik tersebut disebut *trait*, dan seseorang dapat memiliki lebih dari satu *trait*. Setiap *trait* merupakan unjuk kerja dari orang yang bersangkutan. Setiap *trait* merupakan dimensi kemampuan seseorang (Suryani, 2018). Dengan demikian, teori respon butir atau yang biasa disebut IRT ini dapat menghasilkan pengukuran yang ekuivalen pada partisipan dari berbagai kelompok eksperimen yang berbeda (Rahmawati, 2014). Sehingga, dapat dipastikan IRT ini adalah pendekatan yang baik untuk mengukur IST.

Beberapa penelitian mengenai evaluasi karakteristik psikometri IST telah dilakukan. Penelitian yang dilakukan oleh Rahmawati (2014) yang berjudul “Evaluasi Karakteristik Psikometri *Intelligenz Struktur Test* (IST) Menggunakan Pendekatan *Item Response Theory*”. Dengan responden partisipan tes IST di Pusat Penelitian dan Pengabdian Kepada Masyarakat (P3M) Universitas Sumatera Utara. Menunjukkan bahwa 53.125% dari aitem-aitem IST yang dianalisis memiliki karakteristik psikometri yang kurang baik menurut batasan psikometri. Adapun hasil analisis spesifik pada subtes RA menunjukkan bahwa dari aitem-aitem RA hanya 45% yang dikatakan baik dan pada subtes ZR hanya 55% yang dikatakan baik.

Selanjutnya penelitian yang dilakukan oleh Sari (2011) yang meneliti mengenai “Karakteristik Psikometri Subtes *Rechenaufgaben* (RA) Pada *Intelligenz Struktur Test* (IST)”. Dilakukan dengan menggunakan pendekatan teori tes klasik ditemukan bahwa dari total 20 aitem pada subtes RA, 16 aitem yang dianggap baik

berdasarkan indeks diskriminasinya. Adapun RA memiliki koefisien reliabilitas sebesar 0,851 sehingga dianggap kurang reliabel dalam mengukur fungsi ukurnya. Selain itu, berdasarkan analisis validitas konstraknya ditemukan bahwa subtes RA tidak lagi berfungsi sebagaimana tujuan awal subtes ini disusun. Sehingga dalam penelitian ini menyarankan perlunya peninjauan ulang terhadap aitem-aitem subtes RA. Hal ini perlu dilakukan sebelum digunakan sebagai tes inteligensi. Khususnya pada individu dengan latar belakang pendidikan SMA, D3, dan S1.

Penelitian ini melanjutkan penelitian sebelumnya mengenai evaluasi IST subtes verbal. Penelitian tersebut menunjukkan bahwa IST berbahasa Indonesia khususnya aspek verbal masih relevan digunakan di Indonesia. Sebagaimana yang telah dilaporkan bahwa dari ketiga subtes yang mewakili kemampuan verbal, yaitu Subtes *Satzergaenzung*, *Wortauswahl*, dan *Analogien*, diperoleh 71,67% aitem memiliki kualitas yang baik (valid) berdasarkan analisis aitem *Item Response Theory* (IRT) serta estimasi daya beda aitem yang dapat diterima (Tarigan & Fadillah, 2021). Meskipun dalam penelitian ini menambahkan catatan bahwa sejumlah aitem, yaitu 25% keseluruhan aitem memerlukan perbaikan. Hal ini disebabkan karena terindikasi ada bias jenis kelamin berdasarkan analisis DIF. Oleh karena itu, perlu adanya analisis bagi subtes yang mewakili aspek lainnya.

Berdasarkan pemaparan di atas, maka penelitian ini bertujuan untuk melakukan evaluasi dan pengkajian ulang apakah IST masih berfungsi dalam menghasilkan pengukuran inteligensi yang berkualitas, khususnya pada subtes RA dan ZR dengan menganalisis karakteristik psikometri, meliputi analisis indeks kesukaran aitem, indeks daya beda aitem, peluang tebakan, dan analisis *Differential Item Function* (DIF).

METODE PENELITIAN

Penelitian ini merupakan penelitian kuantitatif yang bertujuan untuk menguji validasi pengukuran inteligensi dengan menggunakan *Item Response Theory* (IRT). Dimana ditambahkan analisis reliabilitas dengan menggunakan *Split Half*. Metode pengumpulan data yang digunakan dalam penelitian ini dengan menggunakan metode dokumentasi, yaitu data penelitian ini diperoleh dari respon jawaban partisipan psikotes. Data yang digunakan berupa data dikotomi (benar dan salah) sehingga model dalam penelitian ini disebut model logistik dikotomus. Model logistik dikotomus dinamai sesuai dengan banyaknya parameter yang dilibatkan dalam model (Hambleton, Swaminathan, & Rogers, 1991). Adapun model IRT yang digunakan adalah model IRT 3 PL.

Partisipan penelitian ini adalah individu lulusan SMA hingga Strata 1 di Kota Bandung yang berjumlah sebanyak 3,662 partisipan. Tes IST RA dan ZR dilakukan dalam metode administrasi tes IST yang terstruktur sehingga seluruh standar, ketentuan, dan batasan waktu yang disediakan di setiap subtes. Hal ini mengikuti panduan

standar IST yang dilakukan selama periode 2016 sampai 2019.

Instrumen penelitian ini adalah *Intelligenz Struktur Test* (IST) subtes kelima *Rechenaufgaben* (RA) dan keenam *Zahlen Reihen* (ZR). Setiap subtes terdiri dari 20 aitem soal. Proses skoring dalam IST adalah memberikan nilai 1 (satu) untuk jawaban benar dan nilai 0 (nol) untuk jawaban salah pada masing-masing subtes. Dengan menghitung skor yang diperoleh pada masing-masing subtes maka diperoleh skor mentah (*raw score*) pada setiap subtes IST, yang selanjutnya dengan menggunakan norma tes IST, *raw score* tersebut diubah menjadi *weighted score* yang selanjutnya dikonversi lagi menjadi nilai inteligensi seseorang dalam bentuk angka *Intelligence Quotient* (IQ).

Subtes RA berisi soal-soal di mana partisipan diminta untuk melakukan hitungan sederhana dengan waktu penyelesaian s10 (sepuluh) menit. Dalam subtes ini aspek yang dilihat adalah kemampuan berpikir praktis dalam berhitung, berpikir matematis, logis, dan lugas, bernalar, berpikir runtun, berpikir induktif, reasoning, dan kemampuan mengambil kesimpulan. Berikut ini adalah contoh soal subtes RA.

SUBTES 5 RECHENAUFGABEN (RA)

Instruksi: Soal berikut ialah soal hitungan.

Sebatang pensil harganya 25 rupiah. Berapakah harga 3 batang?

1 2 3 4 5 6 7 8 9 0

Jawaban yang tepat ialah 75.

Sementara subtes ZR berisi soal-soal di mana partisipan diminta untuk melanjutkan deret angka dengan waktu penyelesaian selama 10 (sepuluh) menit. Dalam subtes ini, akan dilihat bagaimana cara berpikir teoritis dengan hitungan, berpikir induktif dengan angka-

angka, kelincahan fleksibilitas, dan kemampuan berpikir. Dengan mengubah atau menggantikan cara atau pendekatan, dan mengukur komponen ritmis berirama. Berikut ini adalah contoh soal subtes ZR.

Subtes 6 ZAHLEN REIHEN (ZR)

Instruksi: Pada persoalan berikut, Anda akan dihadapkan dengan deret angka. Setiap deret tersusun menurut suatu aliran atau pola tertentu. Angka terakhir dari deret tersebut kosong. Anda diminta untuk mengisi angka yang kosong dari deret tersebut dengan cara mencoret angka yang merupakan pilihan jawaban Anda. Urutan angka jawaban tidak perlu diperhatikan. Coretlah jawaban Anda di halaman jawaban yang disediakan.

2 4 6 8 10 12 14 ?

1 2 3 4 5 6 7 8 9 0

Jawaban yang tepat adalah 16.

Studi dalam mengidentifikasi properti psikometrik terhadap subtes RA dan ZR akan dilakukan dalam tiga tahap penting, yaitu: (1) analisis item dimana parameter diestimasi dengan menggunakan pendekatan *Item Response Theory* (IRT) dengan metode estimasi *Marginal Maximum Likelihood*; (2) uji validitas; dan (3) uji reliabilitas.

Tahap pertama adalah analisis item untuk mengidentifikasi kualitas setiap item dalam membedakan partisipan kemampuan numeriknya. Guna mencapainya, maka dilakukan metode analisis parameter aitem, analisis statistik fit aitem, dan analisis *differential item function* (DIF). Analisis tersebut dilakukan sebagai masukan empiris mengenai aitem mana saja yang masih berkualitas dan yang tidak berkualitas. Analisis pada tahap ini dilakukan dengan bantuan perangkat lunak jMetrik.

Adapun model IRT yang digunakan adalah model logistik 3 (tiga) parameter, di mana parameter aitem yang digunakan terdiri dari daya beda (a), tingkat kesulitan (b), dan peluang menebak aitem (c). Peluang menebak diperhitungkan dalam analisis ini karena dengan keterbatasan waktu tes dimungkinkan

tidak semua partisipan memiliki kemampuan yang cepat untuk memberikan jawaban yang benar secara utuh. Terlebih dalam proses pengerjaan subtes RA dan ZR ini dibutuhkan kemampuan menghitung yang cepat sehingga bisa diasumsikan terjadinya *blind guessing*. *Blind guessing* adalah kondisi dimana seorang partisipan memilih jawaban secara acak dari beberapa alternatif yang ditawarkan (Obinne, 2012). Selain itu, pada model IRT 3PL ini diperlukan pengaturan nilai $D = 1,702$ agar dapat mengalihkan perhitungan model *ogive* normal ke perhitungan model logistik (Rahmawati 2014). Selanjutnya, yaitu tahap kedua dilakukan uji validitas dengan *convergen-discriminant validity*. Dan pada tahap terakhir adalah uji reliabilitas dengan membagi alat tes yang telah melalui proses analisis item ke dalam dua belahan yang setara atau disebut dengan *split-half*.

HASIL DAN PEMBAHASAN

Data subyek penelitian meliputi skor tes IST subtes RA dan ZR partisipan yang berjumlah 3,662 orang. Tabel 1 menunjukkan demografi para partisipan tes. Sementara tabel 2 merangkum hasil tes IST subtes RA dan ZR.

Tabel 1.
Demografi partisipan penelitian

Usia (tahun)	Jenis Kelamin	Jumlah
11-30	Laki-laki	1,286 (62%)
	Perempuan	778 (38 %)

Tabel 2.
Statistik deskriptif hasil tes IST

	Mean	Modus	Min	Max
RA	4.154	2	0	20
ZR	8.085	5	0	20

Berdasarkan tabel 2 di atas, diketahui bahwa rata-rata skor subtes RA lebih rendah daripada subtes ZR. Adapun pada tabel 3, ditampilkan hasil yang lebih rinci, yaitu persentase jawaban partisipan tes untuk setiap butir soal.

Berdasarkan tabel 3 diketahui bahwa pada IST subtes RA, persentase tertinggi partisipan menjawab benar adalah pada aitem soal ke-1 dan terendah pada soal ke-17. Sedangkan pada subtes ZR persentase tertinggi untuk jawaban benar adalah pada soal ke-20 dan terendah pada soal ke-1.

Tabel 3.
Persentase jawaban partisipan

Item	Persentase RA (%)		Item	Persentase ZR (%)	
	Salah	Benar		Salah	Benar
RA1	6.52	93.42	ZR1	7.30	92.70
RA2	28.96	70.99	ZR2	16.70	83.30
RA3	71.15	28.79	ZR3	22.30	77.70
RA4	77.76	22.19	ZR4	21.20	78.80
RA5	47.38	52.57	ZR5	42.80	57.20
RA6	76.28	23.66	ZR6	39.30	60.70
RA7	55.62	44.32	ZR7	61.50	38.50
RA8	81.63	18.31	ZR8	69.50	30.50
RA9	90.94	9.01	ZR9	67.20	32.80
RA10	95.01	4.94	ZR10	56.70	43.30
RA11	84.28	15.67	ZR11	61.20	38.80
RA12	92.47	7.48	ZR12	62.50	37.50
RA13	94.90	5.05	ZR13	69.80	30.20
RA14	94.73	5.21	ZR14	79.40	20.60
RA15	96.51	3.44	ZR15	75.70	24.30
RA16	96.10	3.85	ZR16	83.70	16.30
RA17	98.83	1.12	ZR17	86.10	13.90
RA18	98.28	1.66	ZR18	87.10	12.90
RA19	97.98	1.97	ZR19	90.00	10.00
RA20	98.42	1.53	ZR20	91.60	8.40

Selanjutnya, pada tabel 4 disajikan rekapitulasi rata-rata persentase jawaban partisipan berdasarkan subtes.

Tabel 4.
Tabel rekap rata-rata persentase jawaban partisipan

Subtes	Rata-rata Persentase Jawaban (%)	
	Benar	Salah
RA	20.8	79.2
ZR	40.4	59.6

Kemudian, untuk menginterpretasi model IRT 3PL yang digunakan, terlebih dahulu perlu diketahui beberapa ketentuan berdasarkan pendapat beberapa ahli. Ketentuan tersebut telah dirangkum dalam tabel 5 hingga tabel 7. Dalam mendefinisikan parameter aitem daya beda dan peluang menebak digunakan kriteria menurut Baker (2001). Adapun dalam klasifikasi tingkat kesulitan didasarkan pada kriteria yang dikemukakan oleh De Ayala (2008).

Selain estimasi parameter, informasi yang bisa didapatkan dari IRT adalah melihat kualitas butir dengan model atau kesesuaian

butir (fit aitem). Fit aitem menjelaskan apakah butir soal berfungsi normal, melakukan pengukuran atau tidak (Sumintono & Widhiarso, 2015). Butir soal yang sesuai (aitem fit) berarti soal tersebut 'berperilaku' secara konsisten dengan apa yang diharapkan oleh model (Benyamin, 1998). Sedangkan butir soal tidak fit merupakan indikasi bahwa terjadi miskonsepsi pada partisipan tes terhadap butir tersebut. Informasi miskonsepsi pada partisipan sangat berguna untuk memperbaiki kualitas butir sehingga miskonsepsi bisa dihindari kembali.

Tabel 5.
Kriteria daya beda (a)

Sumber	Daya Beda (a)	Kategori
De Ayala (2008)	0.8 – 2.5	Bagus
	0	Tidak dapat membedakan
	0.01–0.34	Sangat rendah
	0.35–0.64	Rendah
	0.65–1.34	Sedang
Baker (2001)	1.35–1.69	Tinggi
	>1.7	Sangat tinggi
	+ <i>infinity</i> (tak terhingga)	Sempurna
Hambleton, dkk (1991)	0 – 2	Normal

Tabel 6.
Kriteria tingkat kesulitan (b)

Sumber	Tingkat Kesulitan (b)	Kategori
De Ayala (2008)	- <i>infinity</i> sampai + <i>infinity</i>	Rentang teoritis
	-3 sampai 3	Rentang khas
	kurang dari -2	Mudah
	-2 sampai 2	Rata-rata
	lebih dari 2	Sulit
Baker (2001)	-4 sampai 4	Rentang teoritis
	-3 sampai 3	Rentang khas
	-2.8 sampai 2.8	Rentang normal
Hambleton, dkk (1991)	kurang dari -2	Sangat rendah
	-2 sampai 2	Sedang
	lebih dari 2	Sangat sulit

Tabel 7.
Kriteria peluang menebak (c)

Sumber	Peluang Menebak (c)	Kategori
Baker (2001)	0 sampai 0.35 di atas 0.35	Dapat diterima Tidak dapat diterima

Adapun statistik fit aitem memiliki karakteristik pengujian sebagai berikut:

1) Hipotesis uji:

H_0 : Item mengikuti model IRT

H_1 : Item tidak mengikuti model IRT

2) Daerah kritis:

H_0 ditolak jika $p\text{-value} < \alpha$. Tingkat kepercayaan (α) yang digunakan dalam penelitian ini sebesar 95%. Sehingga, H_0 ditolak jika $p\text{-value} < 0.05$.

Tabel 8 dan Tabel 9 masing-masing menyajikan hasil analisis setiap subtes IST menggunakan metode *Marginal Maksimum Likelihood*.

Tabel 8.

Hasil analisis parameter dan statistik fit aitem RA

Butir	Analisis Parameter Aitem					Statistik <i>Fit</i> Aitem		
	Daya Beda (a)	Keterangan	Tingkat Kesulitan (b)	Keterangan	Peluang Menebak (c)	Keterangan	<i>p</i> - <i>value</i>	Keterangan
RA1	0.41	Rendah	-3.01*	Mudah	0.50*	Ditolak	0.0000	Tidak Signifikan
RA2	0.73	Sedang	-0.77	Rata-rata	0.11	Diterima	0.0079	Tidak Signifikan
RA3	0.79	Sedang	0.96	Rata-rata	0.01	Diterima	0.0000	Tidak Signifikan
RA4	0.9	Sedang	1.20	Rata-rata	0.01	Diterima	0.4227	Signifikan
RA5	0.94	Sedang	-0.03	Rata-rata	0.03	Diterima	0.0000	Tidak Signifikan
RA6	1.43	Tinggi	0.92	Rata-rata	0.01	Diterima	0.3054	Signifikan
RA7	1	Sedang	0.33	Rata-rata	0.04	Diterima	0.0002	Tidak Signifikan
RA8	1.48	Tinggi	1.15	Rata-rata	0.01	Diterima	0.0003	Tidak Signifikan
RA9	1.19	Sedang	1.83	Rata-rata	0.01	Diterima	0.7769	Signifikan
RA10	1.39	Tinggi	2.14	Sulit	0.01	Diterima	0.0541	Signifikan
RA11	1	Sedang	1.34	Rata-rata	0.01	Diterima	0.8963	Signifikan
RA12	1.87	Sangat Tinggi	1.72	Rata-rata	0.01	Diterima	0.4681	Signifikan
RA13	2.44	Sangat Tinggi	1.83	Rata-rata	0.01	Diterima	0.1549	Signifikan
RA14	2.18	Sangat Tinggi	1.83	Rata-rata	0.00	Diterima	0.8509	Signifikan
RA15	2.16	Sangat Tinggi	2.07	Sulit	0.01	Diterima	0.3760	Signifikan
RA16	2.04	Sangat Tinggi	2.05	Sulit	0.01	Diterima	0.7747	Signifikan
RA17	2.19	Sangat Tinggi	2.63	Sulit	0.00	Diterima	0.6967	Signifikan
RA18	2.29	Sangat Tinggi	2.31	Sulit	0.00	Diterima	0.2814	Signifikan
RA19	2.15	Sangat Tinggi	2.31	Sulit	0.00	Diterima	0.8179	Signifikan
RA20	2.18	Sangat Tinggi	2.70	Sulit	0.01	Diterima	0.2539	Signifikan

Tabel 9.
Hasil analisis parameter dan statistik fit aitem ZR

Butir	Analisis Parameter Aitem						Statistik <i>Fit</i> Aitem	
	Daya Beda (a)	Keterangan	Tingkat Kesulitan (b)	Keterangan	Peluang Menebak (c)	Keterangan	<i>p-value</i>	Keterangan
ZR1	0.91	Sedang	-2.15	Mudah	0.10	Diterima	0.000	Tidak Signifikan
ZR2	0.92	Sedang	-1.40	Rata-rata	0.04	Diterima	0.000	Tidak Signifikan
ZR3	0.92	Sedang	-1.08	Rata-rata	0.02	Diterima	0.000	Tidak Signifikan
ZR4	1.12	Sedang	-1.04	Rata-rata	0.02	Diterima	0.000	Tidak Signifikan
ZR5	0.84	Sedang	-0.18	Rata-rata	0.04	Diterima	0.354	Signifikan
ZR6	1.16	Sedang	-0.30	Rata-rata	0.02	Diterima	0.067	Signifikan
ZR7	1	Sedang	0.47	Rata-rata	0.02	Diterima	0.236	Signifikan
ZR8	1.5	Tinggi	0.69	Rata-rata	0.02	Diterima	0.089	Signifikan
ZR9	1.2	Sedang	0.67	Rata-rata	0.02	Diterima	0.595	Signifikan
ZR10	1.56	Tinggi	0.27	Rata-rata	0.02	Diterima	0.003	Tidak Signifikan
ZR11	2	Sangat Tinggi	0.40	Rata-rata	0.02	Diterima	0.002	Tidak Signifikan
ZR12	1.74	Tinggi	0.46	Rata-rata	0.02	Diterima	0.075	Signifikan
ZR13	2	Sangat Tinggi	0.69	Rata-rata	0.03	Diterima	0.000	Tidak Signifikan
ZR14	2.54	Sangat Tinggi	0.94	Rata-rata	0.01	Diterima	0.001	Tidak Signifikan
ZR15	2.53	Sangat Tinggi	0.83	Rata-rata	0.02	Diterima	0.165	Signifikan
ZR16	2.36	Sangat Tinggi	1.11	Rata-rata	0.01	Diterima	0.123	Signifikan
ZR17	1.55	Tinggi	1.36	Rata-rata	0.01	Diterima	0.299	Signifikan
ZR18	1.81	Tinggi	1.39	Rata-rata	0.02	Diterima	0.968	Signifikan
ZR19	2.32	Sangat Tinggi	1.43	Rata-rata	0.01	Diterima	0.690	Signifikan
ZR20	1.66	Tinggi	1.67	Rata-rata	0.01	Diterima	0.554	Signifikan

Berdasarkan hasil yang disajikan dalam tabel 8, pada subtes RA ditemukan bahwa nilai daya beda aitem (a) berkisar pada nilai 0.41 sampai dengan 2.44. Terdapat satu aitem soal yang termasuk kategori “Rendah”, tujuh aitem soal berkategori “Sedang”, tiga aitem yang tergolong pada kategori “Tinggi”, dan sembilan aitem tergolong berdaya beda “Sangat Tinggi”. Nilai b, yaitu tingkat kesulitan aitem soal, berkisar -3.01 sampai dengan 2.7. Terdapat satu aitem soal yang

keluar dari rentang normal karena memiliki nilai b kurang dari -2.8. Terdapat satu aitem soal tergolong “mudah”, 12 aitem dengan tingkat kesulitan “Rata-rata”, dan 7 aitem “Sulit”. Kedua subtes memiliki peluang menebak aitem (c) berkisar 0.00 sampai dengan 0.50. Ditemukan bahwa terdapat satu aitem (RA butir 1) yang ditolak karena memiliki nilai c lebih dari 0.35.

Selanjutnya, hasil analisis statistik aitem pada penelitian ini menggunakan tingkat

kepercayaan 95% sehingga nilai $\alpha = 0,05$. Ditemukan terdapat 6 aitem (RA butir 1, 2, 3, 5, 7, 8) yang tidak signifikan atau dapat dikatakan aitem tidak mengikuti model IRT dan diindikasikan terjadi miskonsepsi pada partisipan tes terhadap aitem-aitem soal tersebut.

Sementara itu, untuk hasil analisis subtes ZR yang ditampilkan pada tabel 9. Berdasarkan analisis yang sudah dilakukan ditemukan bahwa pada subtes ZR ditemukan nilai a berkisar dari 0.84 sampai dengan 2.54. Terdapat delapan aitem yang termasuk dalam kategori “Sedang”, enam aitem kategori “Tinggi”, dan enam aitem kategori “Sangat Tinggi”. Nilai tingkat kesulitan (b) berkisar pada nilai -2.15 sampai dengan 1.67. Adapun dalam subtes ZR ini, tidak terdapat aitem soal yang keluar dari rentang normal karena tidak

ada nilai yang kurang dari -2.8 ataupun lebih dari 2.8. Terdapat satu aitem soal yang termasuk klasifikasi “Mudah” (ZR butir 1) dan 19 aitem soal lainnya termasuk pada kategori tingkat kesulitan “Rata-rata”. Adapun peluang menebak (c) berkisar dari 0.01 sampai dengan 0.1. Pada subtes ini, tidak terdapat aitem soal yang memiliki nilai c lebih dari 0.35 sehingga peluang tebakan benar diperoleh 100% memiliki estimasi parameter yang baik (nilai estimasi c kurang dari 0.35). Selanjutnya, pada hasil analisis statistik fit aitem ditemukan terdapat delapan aitem soal (ZR butir 1, 2, 3, 4, 10, 11, 13, dan 14) yang tidak signifikan, yaitu aitem yang tidak mengikuti model IRT 3PL.

Adapun hasil analisis aitem pada masing-masing subtes kami rangkum sebagai berikut

Tabel 10.
Hasil estimasi parameter aitem.

Subtes	Daya Beda (a)				Tingkat Kesulitan (b)		Peluang Menebak (c)	
	Rendah	Sedang	Tinggi	Sangat Tinggi	Dalam Rentang Normal	Diluar Rentang Normal	Diterima	Ditolak
RA	1 (5%)	7 (35%)	3 (15%)	9 (45%)	19 (95%)	1 (5%)	19 (95%)	1 (5%)
ZR	-	8 (40%)	6 (30%)	6 (30%)	20 (100%)	-	20 (100%)	-

Tabel 11.
Hasil analisis statistik fit aitem

Subtes	<i>p-value</i>	
	Signifikan ($\alpha > 0.05$)	Tidak Signifikan ($\alpha < 0.05$)
RA	14 (70%)	6 (30%)
ZR	12 (60%)	8 (40%)

Sebagai analisis lanjutan dari analisis aitem menggunakan IRT, kami menyajikan hasil analisis yang dilakukan dengan metode *Differential Item Functioning* (DIF). DIF untuk mendeteksi adanya bias respon yang disebabkan perbedaan karakteristik antara

suatu kelompok dengan kelompok yang lain. Aitem dikatakan terdeteksi DIF apabila partisipan tes dengan karakteristik sama, namun dari kelompok berbeda memiliki peluang yang berbeda dalam menjawab atau merespon aitem tersebut.

Meyer (2014) menjelaskan skema klasifikasi DIF (A, B, C) sebagai berikut.

- A : *Aitem bagus*, dimana Cochran-Mantel-Haenszel (CMH) Statistik Chi-square $p\text{-value} > 0.05$ atau Rasio odds antara 0,65 sampai 1,53.
- B : *Aitem yang dipertanyakan*. Aitem ini bukan termasuk klasifikasi A dan C.
- C : *Aitem yang buruk*. Rasio odds < 0.53 dan batas atas CI < 0.65 , dimana baik aitem dan interval kepercayaan sangat rendah atau Rasio odds > 1.89 dan batas bawah CI > 1.53 , dimana item dan interval kepercayaan sangat tinggi.

Sementara itu, tanda ‘+’ pada *class B* atau *C* menunjukkan aitem lebih mudah bagi kelompok *focal*, sedangkan tanda ‘-’ menunjukkan aitem lebih mudah bagi

kelompok *reference*. E.S menunjukkan *common odds ratio*, yaitu berapa kali aitem tersebut lebih mudah bagi kelompok *focal*. Berikut hasil analisis DIF untuk subtes RA dan ZR dengan bias aitem yang digunakan adalah jenis kelamin, yaitu perempuan sebagai *focal* dan laki-laki sebagai *reference*.

Berdasarkan tabel 12, diketahui bahwa dari 20 aitem, hanya terdapat satu aitem yang teridentifikasi DIF (RA butir 11). Dimana aitem tersebut 0.47 kali lebih mudah untuk kelompok perempuan. Dan 15 aitem termasuk pada aitem yang bagus.

Tabel 13 menampilkan hasil analisis DIF. Berdasarkan tabel tersebut, diketahui dari 20 aitem, hanya terdapat satu aitem yang teridentifikasi DIF (ZR butir 20) di mana aitem tersebut 0.43 kali lebih mudah untuk kelompok perempuan. Serta 18 aitem termasuk pada aitem yang bagus.

Tabel 12.
Hasil analisis DIF subtes RA

Aitem	χ^2	<i>p-value</i>	N	E.S.	(95% C.I.)	<i>Cl</i>
RA1	0.38	0.54	3558	1.11	(0.80, 1.55)	A
RA2	4.95	0.03	3558	0.81	(0.67, 0.98)	A
RA3	39.19	0	3558	1.77	(1.48, 2.12)	B-
RA4	30.63	0	3558	1.74	(1.43, 2.11)	B-
RA5	11.15	0	3558	0.74	(0.62, 0.88)	A
RA6	0.68	0.41	3137	1.09	(0.88, 1.35)	A
RA7	0.06	0.81	3558	1.02	(0.87, 1.20)	A
RA8	3.97	0.05	3558	0.8	(0.64, 1.00)	A
RA9	6.56	0.01	3558	1.41	(1.08, 1.83)	A
RA10	9.71	0	3137	1.71	(1.22, 2.39)	B-
RA11	44.75	0	3558	0.47	(0.37, 0.59)	C+
RA12	0.42	0.52	3137	0.91	(0.68, 1.22)	A
RA13	0.27	0.6	3558	1.1	(0.78, 1.54)	A
RA14	0.35	0.55	3137	1.11	(0.79, 1.55)	A
RA15	0	0.97	3137	1.01	(0.68, 1.48)	A
RA16	5.71	0.02	3558	0.64	(0.44, 0.92)	B+
RA17	0.16	0.69	3137	1.14	(0.60, 2.14)	A
RA18	2.76	0.1	1821	1.59	(0.92, 2.75)	A
RA19	1.4	0.24	3137	0.74	(0.45, 1.22)	A
RA20	0.42	0.51	3137	1.19	(0.70, 2.04)	A

Keterangan: $\chi^2 = \text{Chi-square}$, $N = N \text{ valid}$; $Cl = \text{Class}$.

Tabel 13.
Hasil analisis DIF subtes ZR

Aitem	χ^2	<i>p-value</i>	N	E.S.	(95% C.I.)	<i>CI</i>
ZR1	0.16	0.69	3645	1.06	(0.79, 1.44)	A
ZR2	0.4	0.52	3645	1.08	(0.86, 1.35)	A
ZR3	0.51	0.47	3645	1.08	(0.88, 1.32)	A
ZR4	0.21	0.65	3645	0.95	(0.76, 1.19)	A
ZR5	5.66	0.02	3645	1.22	(1.03, 1.43)	A
ZR6	0.47	0.49	3645	0.94	(0.78, 1.13)	A
ZR7	6.33	0.01	3645	1.25	(1.05, 1.48)	A
ZR8	26.81	0	3417	1.7	(1.39, 2.08)	B-
ZR9	0.52	0.47	3417	1.07	(0.89, 1.28)	A
ZR10	8.04	0	3645	0.76	(0.63, 0.92)	A
ZR11	1.62	0.2	3417	1.14	(0.93, 1.40)	A
ZR12	7.32	0.01	3645	0.77	(0.63, 0.93)	A
ZR13	0.36	0.55	3645	0.94	(0.77, 1.15)	A
ZR14	0.37	0.54	3417	0.92	(0.72, 1.19)	A
ZR15	1.56	0.21	3645	0.87	(0.69, 1.09)	A
ZR16	0.08	0.78	3365	1.04	(0.80, 1.35)	A
ZR17	0.41	0.52	3417	1.08	(0.85, 1.37)	A
ZR18	1.07	0.3	3645	0.88	(0.69, 1.12)	A
ZR19	2.68	0.1	3137	0.78	(0.58, 1.05)	A
ZR20	32.91	0	3417	0.43	(0.32, 0.58)	C+

Sementara itu, berdasarkan hasil yang ditunjukkan tabel 14, diketahui bahwa jumlah persentase butir soal dari subtes RA dan ZR yang termasuk kategori aitem bagus (tidak

terjangkit DIF). Dengan bias jenis kelamin berjumlah 33 aitem atau sama dengan 82.5% dan sisanya tujuh aitem atau 17.5% terjangkit DIF.

Tabel 14.
Hasil analisis DIF IST RA & ZR

Sub-tes	Kategori	Soal No ke-	Total Aitem	Persentase (%)
RA	Tidak terjadi DIF (Aitem bagus)	1, 2, 5, 6, 7, 8, 9, 12, 13, 14, 15, 17, 18, 19, 20	15	75
	DIF sedang	3, 4, 10, 16	4	20
	DIF tinggi	11	1	5
ZR	Tidak terjadi DIF (Aitem bagus)	1, 2, 3, 4, 5, 6, 7, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19	18	90
	DIF sedang	8	1	5
	DIF tinggi	20	1	5
RA & ZR	Tidak terjadi DIF (Aitem bagus)		33	82.5
	DIF sedang		5	12.5
	DIF tinggi		2	5

Selanjutnya, dalam penelitian ini dilakukan uji normalitas untuk menguji apakah data berdistribusi normal atau tidak. Hasil uji

normalitas ini sebagai langkah dalam menentukan teknik korelasi yang akan digunakan selanjutnya.

Tabel 15.

Hasil uji *kolmogorov smirnov* pada distribusi frekuensi IST subtes RA & ZR

Distribusi Frekuensi	sig. (W)
Subtes RA	0.0000
Subtes ZR	0.0000

Hasil uji normalitas dengan metode uji *kolmogorov-smirnov* menunjukkan bahwa kedua subtes tidak berdistribusi normal. Sehingga dalam pengujian korelasi pada *convergent-discriminant* serta uji reliabilitas *split half* akan menggunakan teknik *spearman rank order*. Teknik ini dapat digunakan untuk persebaran skor yang tidak membentuk distribusi normal. (Corder & Foreman, 2009).

Uji *convergent* dilakukan guna mengidentifikasi kemampuan subtes RA dan ZR dalam mengukur kemampuan numerik. Uji *convergent* bertujuan untuk membandingkan dua alat ukur yang mengukur konstruk serupa. Sedangkan uji *discriminant* bertujuan untuk membandingkan dua alat ukur yang mengukur konstruk yang berbeda (Cohen & Swerdlik, 2009). Dalam pengujian *convergent validation*, subtes RA akan dikorelasikan dengan subtes ZR. Kedua subtes ini sama-sama mengukur kemampuan dasar numerik. Adapun dalam pengujian *discriminant*, subtes RA dan ZR akan dikorelasikan dengan subtes FA dimana subtes FA ini mengukur kemampuan spasial tiga dimensi.

Tabel 16.

Hasil uji *convergent-discriminant validity* pada IST subtes RA & ZR

	Subtes RA	Subtes ZR	Subtes FA
Subtes RA	1	0.448	0.354
Subtes ZR	0.448	1	0.450
Subtes FA	0.354	0.450	

*Signifikan pada $p < 0.01$.

Berdasarkan tabel 16 dapat diketahui bahwa korelasi antara subtes RA dengan subtes ZR ($r = 0.448$) lebih kuat dibandingkan dengan

korelasi antara subtes RA dengan FA ($r = 0.354$). Nilai *convergent* diidentifikasi lebih kuat karena subtes RA dan ZR mengukur kemampuan yang serupa, yaitu kemampuan numerik. Akan tetapi, korelasi antara subtes ZR dan subtes RA ($r=0.448$) tidak lebih kuat dibandingkan dengan korelasi antara subtes ZR dengan subtes FA ($r=0.450$). Hal ini bukan berarti subtes RA dan ZR tidak mengukur kemampuan yang serupa perbedaan nilai korelasi antar RA dan ZR dengan ZR dan FA hanya berbeda sedikit (0.002). Ada kemungkinan bahwa kemampuan numerik yang diukur dalam subtes ZR tidak dapat dilepaskan dari kemampuan keruangan yang diukur dalam subtes FA. Adapun hasil signifikan pada korelasi *discriminant* memberikan pengetahuan tambahan bahwa kemampuan numerik memiliki hubungan positif dengan kemampuan keruangan.

Pengujian reliabilitas dilakukan dengan metode psikometri *split half* yang membagi subtes menjadi dua belahan (Cohen & Swerdlik, 2009). Untuk mengidentifikasi nilai koefisien korelasi antar belahan dalam penelitian ini menggunakan metode korelasi *spearman-brown*. formula ini digunakan untuk mengidentifikasi reliabilitas keseluruhan subtes yang dibelah menjadi dua belahan (Crocker & Algina, 1986; Cohen & Swerdlik, 2009). Berikut hasil uji reliabilitas subtes RA dan ZR.

Tabel 17.

Hasil Uji Reliabilitas Spearman-Brown Coefficient

	Subtes RA	Subtes ZR
<i>Spearman-Brown Coefficient</i>	0.789	0.675

Hasil koefisien korelasi ini mengindikasikan bahwa 78.9% varians *observed score* pada subtes RA dan 67.5% *variens observed score* pada subtes ZR disebabkan oleh varians *true score*. *True score* menunjukkan nilai murni dari partisipan tanpa

dipengaruhi oleh error, sedangkan *observed score* adalah nilai *true score* yang dipengaruhi oleh nilai error (Crocker & Algina, 1986). Artinya, indikasi error yang terjadi dan mempengaruhi nilai individu di subtes RA berkisar sebesar 21.1% sedangkan pada subtes ZR berkisar sebesar 32.5%. Selain itu, nilai koefisien reliabilitas subtes RA sebesar 0.789 dan subtes ZR sebesar 0.675 menunjukkan bahwa subtes RA dan ZR memiliki reliabilitas yang tinggi. Sehingga mampu menggambarkan konsistensi pengukuran terhadap kemampuan partisipan.

SIMPULAN

Berdasarkan hasil analisis butir soal dengan pendekatan item response theory 3PL menunjukkan bahwa subtes RA dan ZR memiliki butir soal yang baik. Hal ini dapat dilihat dari hasil estimasi parameter aitem dan hasil estimasi statistik fit aitem pada subtes RA dan ZR. Pada subtes RA 100% aitem termasuk pada kategori nilai a (daya beda) bagus. Pada nilai b (tingkat kesulitan), hanya terdapat 1 aitem soal (RA butir 1) yang keluar dari rentang normal. Dan nilai c (peluang menebak) hanya memiliki 1 aitem soal (RA butir 1) yang ditolak. Adapun hasil statistik fit aitem subtes RA menunjukkan 70% aitem signifikan atau mengikuti model IRT. Selanjutnya, hasil estimasi parameter subtes ZR baik nilai a (daya beda), b (tingkat kesulitan), maupun c (peluang menebak) menunjukkan bahwa 100% aitem termasuk pada rentang normal. Adapun hasil statistik fit aitem subtes ZR menunjukkan 60% aitem signifikan mengikuti model IRT. Sementara hasil analisis DIF dengan bias jenis kelamin menunjukkan bahwa subtes RA dan ZR pada masing-masing subtes hanya 1 butir aitem yang terjangkit DIF tinggi. Terkait dengan uji properti psikometrik validitas dan reliabilitas, subtes RA dan ZR berada pada posisi yang cukup baik. Hal ini dapat dilihat melalui hasil korelasi yang positif antara RA dan ZR. Selain itu, hasil reliabilitas menunjukkan bahwa error yang terjadi pada

subtes RA dan ZR cenderung kecil (21.1% untuk subtes RA dan 32.5% untuk subtes ZR). Hal ini berarti bahwa aitem-aitem dengan pilihan jawaban pada subtes RA dan ZR mampu memberikan stimulasi kepada partisipan tes untuk mengukur kemampuan numeriknya. Dengan demikian, dapat disimpulkan bahwa subtes RA dan ZR pada tes IST masih berfungsi dan valid dalam menghasilkan pengukuran inteligensi yang berkualitas. Hal ini dikhususkan dalam mengukur kemampuan numerik.

Berdasarkan hasil penemuan dari penelitian ini, guna melengkapi hasil analisis karakteristik psikometri dalam mengevaluasi IST khususnya subtes RA dan ZR. Maka dapat dilakukan analisis distraktor dan analisis lanjutan DIF dengan bias yang lain. Mengingat partisipan tes IST ini dilakukan pada jenjang usia yang kompleks serta dilakukan pada partisipan yang jelas memiliki latar belakang pendidikan dan lintas budaya yang berbeda.

DAFTAR PUSTAKA

- Adinugroho, I. (2016). *Pengujian Properti Psikometrik Intelligenz Struktur Test Subtes Kemampuan Spasial Dua Dimensi (Form AUSWAHL): Studi pada Dua SMA Swasta di Jakarta*. 5(2), 165–180.
- Aminah, Nonoh Siti. (2013). The Characteristics Of Test Equating Methods For Dichotomous Data. *Jurnal Penelitian dan Evaluasi Pendidikan*, Pg 88-100.
- An, X., & Yung, Y. (2014). Item Response Theory : What It Is and How You Can Use the IRT Procedure to Apply It. *SAS Institute Inc.*, 1–14. <https://support.sas.com/resources/papers/proceedings14/SAS364-2014.pdf>.
- Baker, F. B. (2001). *Item Response Theory the Basics of Item Response Theory the* (C. Boston & L. Rudner (eds.); Second Edi). ERIC Clearinghouse on Assessment and Evaluation.
- Benyamin, J. C. (1998). *Analisis kualitas soal Ebtanas PPKn SMU tahun pelajaran 1996/1997 dengan Pendekatan model rasch di Provinsi Nusa Tenggara Timur*.

- Yogyakarta: Fakultas Psikologi Universitas Gadjah Mada Yogyakarta.
- Cohen, R. J., & Swerdlik, M. E. (2010). *Psychological testing and assessment: An introduction to tests and measurement (7th.ed)*. Penn Plaza: McGrawHill.
- Corder, Gregory W. & Foreman, Dale I. (2009). *Nonparametric Statistics for Non-Statisticians: A Step-by-Step Approach 1st Edition*. Hoboken, New Jersey: Wiley
- Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory-Procedures for Estimating Reliability*. Holt, Rinehart and Winston, Inc.
- De Ayala, R. J. (2009). *The Theory and Practice of Item Response Theory*. New York: The Guilford Press.
- Demars, Christine. (2010). *Item Response Theory: Understanding Statistics Measurement*. New York: Oxford University Press, Inc.
- Embretson, S.E. & Reise, S.P. (2000). *Item Response Theory for Psychologists*. New Jersey: Lawrence Erlbaum Associates, Publishers.
- Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of Item Response Theory*. California, USA: Sage Publication.
- Hambleton, Ronald K. (1990). Item Response Theory: Introduction and Bibliography. *Psicothema* 1990, 2(1), 97-107.
- Jonas, N. (2018). *DIRECTORATE FOR EDUCATION AND SKILLS NUMERACY PRACTICES AND NUMERACY SKILLS AMONG ADULTS OECD Education Working Paper No. 177*. www.oecd.org/edu/workingpapers
- Loe, Aiden. (2018). *Intro to IRT*. <https://aidenloe.github.io/introToIRT.html>. Diakses pada tanggal 28 November 2019 pukul 16.00 WIB.
- Mardapi, Djemari. (2017). *Pengukuran Penilaian Dan Evaluasi Pendidikan Edisi 2*. Yogyakarta: Parama publishing
- McCowan, Richard J. & McCowan Sheila J. (1999). *Item Analysis for Criterion Referenced Tests*. USA: Center for Development of Human Services.
- Meyer, J.Patrick. (2014). *Applied Measurement with Jmetrik*. New York: Routledge.
- Naga, Dali S. (1992). *Pengantar Teori Sekor pada Pengukuran Pendidikan*. Depok: Universitas Gunadarma.
- Nayla Amalia, A., & Widayati, A. (2012). Analisis butir soal tes kendali mutu kelas XII SMA mata pelajaran Ekonomi Akuntansi di Kota Yogyakarta. *Jurnal Pendidikan Akuntansi Indonesia*, X(1), 1–26.
- Obinne, A.D.E. 2012. Using IRT in Determining Test Item Prone to Guessing. *World Journal of Education*, 2(1).
- Permana. (2017). *Panduan Praktis Personality Assessment*. Jakarta: Raih Asa Sukses.
- Polhaupessy, L.F. (1993). *Intelligenz Struktur Test (IST) dari Rudolf Amthauer. Suatu Pengantar. Lecture notes in Department of Psychology UNPAD*. Bandung: Universitas Padjajaran.
- Purwanto. (2010). Intelegensi: Konsep dan Pengukurannya. *Jurnal Pendidikan dan Kebudayaan*, 16(4), 477-485.
- Rahmawati, Etti. (2014). *Proceeding. Seminar Nasional Psikometri*. Surakarta: Fakultas Psikologi, Universitas Muhammadiyah Surakarta.
- Ridho, Ali. (2013). *Evaluasi Implementasi Kurikulum 2013 dan Sistem Penilaian*. Manado: Univesitas Negeri Manado.
- Sari, Kiki Fatmala. (2011). Karakteristik Psikometri Subtes Rechenaufgaben (RA) Pada Intelligenz Struktur Test (IST). *Skripsi*. Medan: Fakultas Psikologi Universitas Sumatra Utara.
- Spearman, C. (1927). *The Abilities of Man*. New York: Macmillan.
- Sudaryono. (2011). Implementasi Teori Responsi Butir (Item Response Theory) pada Penilaian Hasil Belajar Akhir di Sekolah. *Jurnal Pendidikan dan Kebudayaan*, 17(6), 719-732.
- Sumintono, B., & Widhiarso, W. (2015). *Aplikasi Pemodelan Rasch Pada Assesment Pendidikan*. Cimahi: Trim Komunikata.
- Suryani, Yulinda E. (2018). Aplikasi Rasch Model dalam Mengevaluasi Intelligenz Structure Test (IST). *Psikohumaniora*:

Jurnal Penelitian Psikologi, 3(1), 73–100.
Syah, Muhibin. 2005. *Psikologi Belajar*. Jakarta: PT Raja Grafindo Persada.
Tarigan, M., & Fadillah, F. (2021). Properti

Psikometri Struktur Inteligensi IST Subtes Verbal (Satzergaenzung, Wortauswahl, dan Analogien). *Jurnal Muara Ilmu Sosial, Humaniora, dan Seni*, 5(1), 63–72.