Jurnal Muara Ilmu Sosial, Humaniora, dan Seni
Vol. 5, No. 1, April 2021: hlm 63-72

ISSN 2579-6348 (Versi Cetak)
ISSN-L 2579-6356 (Versi Elektronik)

PROPERTI PSIKOMETRI STRUKTUR INTELIGENSI IST SUBTES VERBAL (*SATZERGAENZUNG*, *WORTAUSWAHL*, DAN *ANALOGIEN*) BERBAHASA INDONESIA

Medianta Tarigan[1], Fadillah[2]

[1]Departemen Psikologi, Universitas Pendidikan Indonesia
Email: medianta@upi.edu
[2]Program Studi Desain Komunikasi Visual, Institut Teknologi Bandung
Email: fadillah@itb.ac.id

Masuk : 13-10-2020, revisi: 09-03-2021, diterima untuk diterbitkan : 21-03-2021

ABSTRACT

Intelligence as one of the individual abilities that is widely used in everyday life has been extensively studied and measured using psychological measurement tools. One of them is the Intelligenz Structure Test (IST). However, at this time IST has leakage through discussions made by many parties. Moreover, the process of IST adaptation to the Indonesian version which tends to translate each word allegedly results in a bias of meaning that can affect the validity of this measurement tools. Therefore, this study is aimed to evaluating the current quality of IST by testing the feasibility of the Indonesian version of IST items for verbal ability, namely SE (Satzergaenzung), WA (Wortauswahl), and AN (Analogien). Item Response Theory (IRT) is used as a research method. The data were collected from 2.064 participants who live in Bandung. The results of the analysis revealed that the SE, WA, and AN subtest are still valid. Based on 60 items analyzed, 71.67% of the items have good quality, i.e. 43 of the 60 items have estimation of discriminant (a) parameter is acceptable. In addition, based on the fit item statistics it was also known that 78.33% of significant items followed the IRT model. Furthermore, based on statistics of item fit, it is also known that 78.33% of items fit the IRT model. This shows that the Indonesian version of IST is still valid to be used particularly in measuring verbal comprehension (V) through 3 subtests (SE, WA, and AN). However, it is necessary to revise the items that have been infected with DIF, in which 25% of items were declared to have a gender bias.

Keywords: *Intelligenz struktur test (ist); item response theory; verbal comprehension*

ABSTRAK

Inteligensi sebagai salah satu kemampuan individu yang banyak berperan dalam kehidupan sehari-hari telah banyak diteliti dan diukur menggunakan alat ukur psikologi. Salah satunya adalah Intelligenz Struktur Test (IST). Namun, saat ini IST telah mengalami kebocoran melalui pembahasan yang dibuat oleh banyak pihak. Selain itu, proses adaptasi IST ke bahasa Indonesia yang cenderung menerjemahkan setiap kata secara langsung diduga mengakibatkan terjadinya bias makna yang dapat mempengaruhi keabsahan alat ukur ini. Oleh karena itu, penelitian ini ditujukan untuk mengevaluasi kualitas terkini IST dengan menguji kelayakan butir soal IST Bahasa Indonesia untuk kemampuan verbal, yaitu SE (Satzergaenzung), WA (Wortauswahl), dan AN (Analogien). Item Response Theory (IRT) digunakan sebagai metode penelitian ini. Data penelitian ini diperoleh dari 2.064 partisipan yang berdomisili di kota Bandung. Adapun penelitian ini menunjukkan hasil bahwa subtes SE, WA, dan AN masih tergolong valid. Berdasarkan 60 item yang dianalisis, 71,67% item memiliki kualitas yang cukup baik, yaitu 43 dari 60 item memiliki estimasi daya beda yang dapat diterima. Selain itu, berdasarkan statistik item fit juga diketahui 78,33% item signifikan mengikuti model IRT. Hal ini menunjukkan bahwa IST Bahasa Indonesia masih valid untuk digunakan terutama dalam mengukur verbal comprehension (V) melalui 3 subtes (SE, WA, dan AN). Namun, perlu dilakukan revisi terhadap item soal yang terjangkit DIF, di mana 25% butir soal dinyatakan mempunyai bias jenis kelamin.

Kata Kunci: Intelligenz struktur test (ist); teori jawaban butir soal; verbal comprehension

1. PENDAHULUAN

Latar Belakang

Inteligensi memiliki sejarah penelitian dan diskusi yang panjang. Sejauh ini, masih belum ada definisi standar mengenai inteligensi. Berbagai definisi diusulkan oleh kelompok atau organisasi dari para psikolog. Dalam hal definisi, sulit untuk berpendapat bahwa ada pengertian obyektif di

mana satu definisi dapat dianggap sebagai definisi yang benar. Meskipun demikian, berbagai definisi tersebut sangat terkait satu sama lain sehingga bila diamati dengan seksama, didapatkan pengertian-pengertian kunci mengenai inteligensi. Inteligensi dapat dikatakan sebagai sebuah konsep global yang melibatkan kemampuan individu untuk bertindak dengan sengaja, berpikir rasional, dan menangani lingkungan secara efektif (Webster & Wechsler, 1958). Kecerdasan merupakan pusat dari kumpulan energi yang diperlukan untuk semua tugas kognitif (Spearman, 1904). Pada teorinya, Spearman mengemukakan perihal konsep faktor g atau biasa dikenal dengan kecerdasan umum. Ia kemudian mengembangkan teknik statistik yang dikenal sebagai analisis faktor, yang memungkinkan peneliti menggunakan sejumlah item tes yang berbeda untuk mengukur kemampuan umum. Spearman menyarankan bahwa faktor g ini bertanggung jawab atas kinerja keseluruhan pada tes kemampuan mental. Ia mencatat bahwa meskipun orang pasti bisa dan sering berhasil unggul di bidang tertentu, orang yang berhasil baik di satu bidang cenderung juga berhasil di bidang lain. Salah satu pengukuran inteligensi yang mengikuti dasar teori Spearman adalah pengukuran yang dilakukan oleh Binet (Becker, 2003).

Gagasan bahwa kecerdasan dapat diukur dan diringkas dengan satu angka pada tes IQ dianggap kontroversial pada jamannya. Banyak ahli berpendapat bahwa faktor g hanyalah salah satu cara berpikir tentang kecerdasan. Salah satu yang menentang padangan faktor g adalah Thurstone. Pada karyanya, ia mempresentasikan metode analisis faktor untuk menjelaskan korelasi antara hasil skor dalam tes psikologis (S. & Thurstone, 1938).

Thurstone menolak gagasan tentang satu faktor yang memiliki aplikasi umum daripada yang lainnya. Menurutnya, kecerdasan tidak dapat digeneralisasikan menjadi satu faktor saja. Terdapat beberapa faktor dari kecerdasan, di mana setiap orang memiliki faktor-faktor tersebut namun bisa saja satu faktor menonjol dan faktor lain tidak. Thurstone mengidentifikasi sejumlah aspek kecerdasan yang disebutnya sebagai *Primary Mental Abilities* (PMA). PMA adalah teori inteligensi yang memaparkan bahwa manusia memiliki tujuh kemampuan dasar yang saling terkait satu sama lain. Ketujuh faktor yang diidentifikasi oleh Thurstone, sesuai dengan proporsi perbedaan individu tersebut adalah: 1) *Verbal Comprehension* (V); 2) *Spatial Orientation (S)*; 3) *Inductive Reasoning (R or I)*; 4) *Number (N)*; 5) *Word Fluency (W)*; 6) *Associative Memory (M)*; 7) *Perceptual Speed (P)* (Tracy et al., 2007). Selain terkait faktor g, Thurstone juga menolak konsep umur mental yang selama ini digunakan oleh Binet. Beliau menganjurkan penggunaan peringkat persentil untuk membandingkan kemampuan antar kelompok subjek (Guilford, 1972).

Salah satu pengukuran kecerdasan yang dikembangkan, mengacu pada teori inteligensi yang dikemukakan Thurstone terkait PMA adalah alat ukur *Intelligence Structure Test* (IST). Pada tahun 1953, IST disusun dan diciptakan oleh Rudolf Amthauer kemudian diterbitkan di bawah naungan Hogrefe Verlag Göttingen. Tahun 1970, Amthauer menerbitkan revisi alat ukur ini dan dinamakan IST-70. Alat ini diperuntukkan kelompok usia antara 12-60 tahun. IST mengukur tingkat kecerdasan umum individu dan memetakan struktur kecerdasan serta menentukan tingkat kecerdasan individu berdasarkan standar kelompok.

Pada penelitian sebelumnya menunjukkan bahwa IST-70 versi Jerman memiliki validitas yang cukup baik untuk memprediksi performa akademis (Schmidt-Atzert & Deter, 1993). Adapun IST sebagai alat ukur inteligensi di Indonesia pertama kali digunakan oleh Pusat Psikologi TNI AD. IST juga banyak diaplikasikan dalam proses rekrutmen, seleksi di perusahaan, dan penjurusan sekolah. Hal ini tidak terlepas dari beberapa penelitian yang telah dilakukan sebelumnya dan memberi kesimpulan bahwa IST merupakan alat tes yang reliabel dalam mengukur inteligensi. Penelitian di Indonesia yang berkaitan dengan IST telah dilakukan, diantaranya oleh Adinugroho

 https://doi.org/10.24912/jmishumsen.v5i1.9623

ISSN 2579-6348 (Versi Cetak)
ISSN-L 2579-6356 (Versi Elektronik)

(2016) yang khusus membahas mengenai IST subtes *Auswahl* (FA) yang mengukur kemampuan spasial dua dimensi. Selain itu, studi komparasi yang dilakukan oleh Kumolohadi dan Suseno (2012) yang meneliti psikometri IST dan *Standard Progressive Matrices* (SPM) di mana dihipotesiskan mengukur suatu tingkat inteligensi yang sama. Hanya saja pada perjalanannya, pengembangan IST-70 dinilai oleh banyak ahli tidak memiliki dasar teoritis yang sistematis dan penjelasan psikometrik terkait model hubungan antar faktornya (Brocke et al., 1998). Penelitian juga banyak dilakukan untuk menelaah masing-masing subtes. Salah satu penelitian dilakukan untuk melihat validitas subtes pada IST-70 yaitu subtes 8. Hasil penelitian menunjukkan bahwa subtes tersebut tidak mengukur imajinasi spasial secara valid (Gittler, 1984). Di Indonesia sendiri, tren yang terjadi saat ini adalah alat tes IST ini ternyata mudah sekali bocor, karena sering sekali dibahas baik dalam buku soal-soal psikotes yang menyajikan cara pengerjaan tes maupun menyajikan kunci jawaban dari tes IST. Individu menjadi mudah memanipulasi jawaban, terutama pada subtes yang berbasis kemampuan verbal karena relatif mudah dihafalkan jawabannya. Pada akhirnya, hasil dari pengetesan tersebut bukan kemampuan sebenarnya dari individu yang menjalankan.

Salah satu hal yang menentukan dalam evaluasi karakteristik psikometri sebuah tes adalah pendekatan yang digunakan. Pendekatan yang modern dalam analisis *item* adalah *Item Response Theory* (IRT) (Sadhu & Laksono, 2018). IRT atau Teori Responsi Butir dinamai juga sebagai Teori Ciri Laten (*Latent Trait Theory* disingkat LTT) atau Lengkungan Karakteristik Butir (*Item Characteristic Curve* disingkat ICC) (Fatkhudin et al., 2016; Sudaryono, 2011). Konsep teori respon butir dimulai pada tahun 1970 oleh psikometri untuk menyelesaikan masalah & kelemahan teori tes klasik karena tes ini tidak dapat menilai pengukuran kemampuan peserta tes yang sebenarnya (Pathak et al., 2013).

Dalam IRT, hubungan antara kinerja (performa subjek tes) dan kemampuan item menjadi fokus utama, di mana kemampuan mengacu pada sifat laten unidimensi yang berkaitan dengan dimensi psikologis yang diukur oleh tes tersebut (Veldhuis et al., 2014). Sifat laten lebih jauh dijelaskan oleh Ilhan dan Guler (2018) adalah sebagai dimensi/konstruksi yang diukur dengan tes tetapi tidak diamati secara langsung. Di samping itu, menurut Xia et al. (2019), perbedaan penting antara IRT dan teori uji klasik adalah bahwa IRT mendefinisikan skala untuk variabel potensial yang diukur oleh satu set *item*, dan *item* dikalibrasi untuk skala yang sama sehingga menggunakan metode IRT dapat dengan mudah mengkalibrasi dua penilaian dengan panjang yang berbeda. Oleh karena itu, IRT telah banyak digunakan untuk mengevaluasi kuesioner yang diterapkan di bidang kesehatan, bidang psikometri dan pendidikan, dalam pemasaran, survei, dan diagnosis kognitif (da Silva et al., 2020). Selama beberapa dekade terakhir, penilaian pendidikan telah menggunakan semakin banyak teknik berbasis IRT untuk mengembangkan tes. Saat ini, semua tes pendidikan utama, seperti *Scholastic Aptitude Test* (SAT) dan *Graduate Record Examination* (GRE), dikembangkan dengan menggunakan teori respons *item*, karena metodologi ini dapat secara signifikan meningkatkan akurasi dan keandalan pengukuran sambil memberikan potensi pengurangan yang signifikan dalam waktu penilaian (An & Yung, 2014). Selain itu, IRT juga metode yang efisien dan bermanfaat untuk menganalisis tidak hanya data *testing*, tetapi juga kuesioner, pengukuran, dan berbagai bentuk data lainnya (Garcia et al., 2018).

IRT memiliki beberapa asumsi yang harus dipenuhi, yaitu unidimensi, independensi lokal, dan *item characteristic curve* (ICC) (Le, 2013; Rahmawati, 2014). Hal yang paling penting dalam asumsi unidimensional adalah adanya satu komponen dominan yang memengaruhi performansi subjek (Diputera, 2018). Independensi lokal juga diartikan sebagai kondisi di mana respon pada *item* yang satu bebas dari pengaruh respon pada *item* lain jika kemampuan yang mempengaruhi

performansi dibuat konstan sehingga jika kemampuan disamakan pada *item*, *item* tidak saling berhubungan (Hambleton et al., 1991). Selain itu, *item characteristic curve* (ICC) adalah kurva yang merefleksikan hubungan yang sebenarnya antara kemampuan dan respon peserta terhadap *item* tes (Rohmah et al., 2018), sehingga parameter *item* dan parameter peserta harus invarian (Aliyu, 2015; Naga, 1992 dalam Rahmawati, 2014). Adapun invarian parameter *item* ini dijelaskan oleh Foster et al. (2017) adalah suatu hal yang penting karena memungkinkan peneliti untuk menggeneralisasi bagaimana *item* bekerja di seluruh populasi.

Disamping asumsi yang harus dipenuhi, IRT dikategorikan menjadi beberapa model. Model-model *item response theory* yang sering digunakan adalah model logistik 1, 2, dan 3 parameter (Bichi et al., 2019; Hambleton et al., 1991).

Tabel 1. *Model IRT*

	IRT Model Logistik		
	1 Parameter (IRT 1 PL)	**2 Parameter (IRT 2 PL)**	**3 Parameter (IRT 3PL)**
Persamaan	$P(\theta) = \frac{1}{1 + e^{-1(\theta - b)}}$	$P(\theta) = \frac{1}{1 + e^{-L}} = \frac{1}{1 + e^{-a(\theta - b)}}$	$P(\theta) = c + (l - c)\frac{l}{l + e^{-a(\theta - b)}}$
Jumlah parameter yang diestimasi	1	2	3
Parameter yang diestimasi	● parameter tingkat kesulitan (*difficulty*) (b)	● parameter tingkat kesulitan (*difficulty*) (b) ● parameter daya beda (*discriminant*) (a)	● parameter tingkat kesulitan (*difficulty*) ● parameter daya beda (*discriminant*) (a) ● parameter *guessing* (c)

Di dalam tulisannya, Erguven (2013) menyatakan bahwa CTT dapat didefinisikan sebagai "berbasis tes" sedangkan IRT dapat didefinisikan sebagai "berbasis *item*". Adapun IRT melengkapi teori tes klasik dengan mengatasi ketergantungan ukuran ciri peserta terhadap ciri *item*, serta ketergantungan ukuran ciri butir soal terhadap peserta tes. Oleh karena itu, dapat menghasilkan pengukuran yang ekuivalen pada peserta dari berbagai kelompok eksperimen yang berbeda (Rahmawati, 2014). Oleh karena itu, IRT adalah pendekatan yang baik untuk mengukur IST mengingat memiliki peranan yang sangat penting untuk melihat perbedaan pada suatu atribut dalam penelitian lintas budaya seperti di Indonesia ini.

Rumusan Masalah

Bagaimana kualitas *item* pada alat ukur inteligensi IST aspek verbal (subtes SE, WA, dan AN) berbahasa Indonesia dengan mengukur menggunakan metode IRT?

2. METODE PENELITIAN

Jenis penelitian yang digunakan dalam penelitian ini adalah metode penelitian kuantitatif. Adapun partisipan atau subyek dalam penelitian ini adalah masyarakat umum yang merupakan lulusan tingkat SMA hingga Strata 1 (S1) di Kota Medan, Palembang, Jakarta, Bandung, Yogyakarta, Semarang, Samarinda dan Balikpapan. Jumlah partisipan yang terlibat adalah 2.064 partisipan.

Pengambilan data subtes SE, WA, dan AN dilakukan dalam metode administrasi tes IST yang terstruktur sehingga seluruh standar, ketentuan, dan batasan waktu yang disediakan di setiap subtes mengikuti panduan standar IST. Karakteristik demografi partisipan pada penelitian ini, yaitu: (a)

https://doi.org/10.24912/jmishumsen.v5i1.9623

ISSN 2579-6348 (Versi Cetak)
ISSN-L 2579-6356 (Versi Elektronik)

berusia 14-30 tahun; (b) partisipan laki-laki sebanyak 1.286 (62%) dan perempuan 778 (38%); (c) periode tes tahun 2016-2018.

Pada penelitian ini, dilakukan tes IST dengan *setting* klasikal. Administrasi tes mencakup buku soal, lembar jawaban, pensil 2B (jika dikerjakan dalam LJK) atau bolpoin. Untuk instruksi tes, *tester* cukup membacakan atau menjelaskan instruksi pengerjaan sesuai dengan yang tercantum pada buku soal. Setiap subtes memiliki metode pengerjaan yang berbeda. Berikut penjelasan metode pengerjaan setiap subtes.

Tabel 2. *IST Subtes Verbal*

Satzergaenzung	***Wortauswahl***	***Analogien***
Pada tes ini terdapat kalimat yang belum lengkap. Peserta diminta memilih satu jawaban di antara lima pilihan jawaban yang diberikan sehingga kalimat menjadi sempurna.	Pada tes ini terdapat 5 (lima) kata dalam tiap soal di mana 4 (empat) kata di antaranya mempunyai kesamaan sehingga peserta diminta mencari satu kata yang tidak memiliki kesamaan dari kelima kata tersebut.	Pada tes ini, dalam setiap item terdapat dua pasang kata, pasangan kata yang kedua berkaitan dengan pasangan kata sebelumnya. Tugas peserta adalah mencari pasangan untuk pasangan kata yang kedua dari pilihan jawaban yang diberikan.

Sementara itu, data yang digunakan dalam penelitian ini adalah data dikotomi (benar = 1 dan salah = 0). Data yang digunakan adalah data *raw score* IST karena yang akan dianalisis adalah tingkat kesukaran per butir soal sehingga skor yang akan ditelaah adalah konsistensi skor benar atau salah pada tiap butir soal dan tidak terkait dengan penormaan skor total kelompok. Parameter diestimasi dengan menggunakan pendekatan IRT dengan metode estimasi *marginal maximum likelihood*. Validasi instrumen tes *Intelligenz Structure Test* yang dilakukan dalam penelitian ini meliputi analisis parameter butir soal, analisis statistik *fit* butir soal, dan *analysis differential item functioning* (DIF). Adapun analisis data ini dilakukan dengan menggunakan bantuan perangkat lunak jMetrik.

Setelah semua asumsi dasar dipenuhi, dipilih model IRT yang akan digunakan. Untuk menentukan model yang tepat, perlu dibuat asumsi tentang fungsi karakteristik *item* untuk membantu menentukan jumlah parameter yang dibutuhkan dalam model yang digunakan. Kemudian parameter tersebut digunakan untuk memperoleh ICC. Model IRT yang digunakan adalah model logistik 2 parameter, yaitu pada model ini diperlukan nilai D = 1,702 agar mengalihkan perhitungan model *ogive* normal ke perhitungan model logistik (Rahmawati, 2014).

3. HASIL DAN PEMBAHASAN

Dalam penelitian ini telah dilakukan analisis butir soal dari alat ukur inteligensi IST dengan menggunakan metode *Item Response Theory* 2 parameter (IRT 2PL). Tujuannya adalah untuk mengukur tingkat kelayakan alat ukur inteligensi IST berbahasa Indonesia terkhusus untuk aspek verbal (*verbal comprehension*). Analisis butir soal IST aspek verbal ini terdiri dari tiga subtes verbal, dengan masing-masing subtes terdiri dari dua puluh butir soal dan diujikan selama sembilan belas menit. Adapun jumlah partisipan dalam penelitian ini adalah 2.064 peserta. Berikut ini adalah gambaran data karakteristik skor IST subtes verbal dari seluruh subjek penelitian. Adapun pada tabel 5 ditampilkan hasil analisis butir soal menggunakan metode IRT.

Tabel 3. *Deskriptif Skor IST (SE, WA, & AN)*

	SE	WA	AN
Rata-rata	8,30	10,77	7,44
Modus	7	10	6
Standar deviasi	3,166	3,232	3,285
Varians	10,021	10,445	10,793
Minimum	0	0	0
Maximum	20	20	20

Hasil analisis butir soal pada subtes SE menunjukkan bahwa pada subtes ini tidak ditemukan butir soal yang memiliki indeks daya beda lebih dari 2,00 atau kurang dari 0,00 dan ditemukan tujuh butir soal yang dinyatakan ditolak, yaitu terdiri dari empat butir soal (SE butir soal ke-2, 3, 4, dan 6) yang memiliki tingkat kesukaran dibawah -2,00 dan tiga butir soal (SE 12, 15, dan 16) dengan tingkat kesukaran diatas 2,00. Adapun hasil analisis statistik *item fit* pada penelitian ini menggunakan tingkat kepercayaan 99% yang berarti nilai $\alpha = 0,01$. Sementara hasil analisis statistik *item fit* menunjukkan bahwa terdapat tiga butir soal subtes SE yang tidak signifikan atau dapat dikatakan tidak mengikuti model IRT, yaitu butir soal ke-2, 11, dan 14.
Selanjutnya, hasil analisis butir soal menunjukkan bahwa pada subtes WA tidak ditemukan butir soal yang memiliki indeks daya beda lebih dari 2,00 atau kurang dari 0,00 tetapi ditemukan tiga butir soal yang ditolak (WA butir ke-2, 3, dan 8) karena memiliki tingkat kesukaran dibawah -2,00. Sedangkan hasil analisis statistik *item fit* menunjukkan bahwa terdapat empat butir soal pada subtes WA yang tidak mengikuti model IRT, yaitu butir soal ke-2, 4, 9, dan 18.

Hasil analisis butir soal subtes AN memberi informasi bahwa tidak terdapat butir soal yang memiliki daya beda lebih dari 2,00 atau kurang dari 0,00 tetapi ditemukan satu butir soal yang ditolak karena memiliki tingkat kesukaran dibawah -2,00 (AN butir ke-1) dan enam butir soal yang memiliki tingkat kesukaran diatas 2,00 (AN butir ke- 12, 13, 15,16, 17, dan 19). Sementara untuk hasil analisis statistik *item fit* subtes AN menunjukkan bahwa terdapat enam butir soal tidak signifikan sehingga dapat dikatakan butir soal tersebut tidak mengikuti model IRT, yaitu butir soal ke-7, 11, 12, 14, 17, dan 18.

Penggolongan parameter daya beda, dan tingkat kesulitan pada penelitian ini mengacu pada kriteria yang dirumuskan oleh Hambleton et al. (1991), yaitu daya beda butir soal yang normal (baik) adalah berada dalam rentang 0 sampai dengan 2, serta tingkat kesulitan sedang (tidak terlalu mudah ataupun terlalu sulit) yaitu berada pada rentang -2 sampai dengan 2. Berdasarkan kriteria ini, diperoleh hasil bahwa dari keseluruhan butir soal yang termasuk ke dalam aspek verbal, seluruhnya termasuk dalam butir soal dengan daya beda yang baik. Ini berarti bahwa setiap butir soal mampu membedakan subjek tes yang memiliki kemampuan verbal tinggi dan rendah.
Di samping itu, untuk tingkat kesulitan butir soal IST berbahasa Indonesia aspek verbal, sebanyak 43 butir soal atau setara dengan 71,67% dari jumlah seluruh butir soal, tergolong kategori baik dan sisanya, yaitu 17 butir atau setara dengan 28,33% dari jumlah keseluruhan butir soal, termasuk pada kategori tidak baik. Dengan kata lain, butir soal aspek verbal dari alat ukur inteligensi IST yang memiliki tingkat kesulitan dalam taraf sedang (tidak terlalu mudah dan tidak terlalu sulit) terdapat sebanyak 71,67% dari jumlah keseluruhan dan sisanya adalah butir soal yang memiliki kesulitan yang rendah (terlalu mudah) atau memiliki kesulitan tinggi (terlalu sulit).

 https://doi.org/10.24912/jmishumsen.v5i1.9623

ISSN 2579-6348 (Versi Cetak)
ISSN-L 2579-6356 (Versi Elektronik)

Tabel 4. *Rangkuman Hasil Estimasi Parameter Butir Soal*

Parameter	Batasan Nilai	Keterangan	SE	WA	AN	Total Verbal
A (daya beda)	$0 \leq a \leq 2$	Baik	20 (100%)	20 (100%)	20 (100%)	60 (100%)
	a < 0, a > 2	Tidak Baik	-	-	-	-
B (tingkat kesulitan)	$-2 \leq b \leq 2$	Baik	13 (65%)	17 (85 %)	13 (65%)	43 (71,67%)
	b < 2, b > 2	Tidak Baik	7 (35%)	3 (15%)	7 (35%)	17 (28,33%)

Tabel 5. *Rangkuman Hasil Estimasi Berdasarkan Kedua Parameter Item*

Parameter	Keterangan	SE	WA	AN	Total Verbal
$-2 \leq b \leq 2; 0 \leq a \leq 2$	Baik	13 (65%)	17 (85%)	13 (65%)	43 (71,67%)
	Tidak Baik	7 (35%)	3 (15%)	7 (35%)	17 (28,33%)

Berdasarkan tabel diatas, dapat disimpulkan bahwa 43 butir soal IST berbahasa Indonesia aspek verbal mampu membedakan peserta dengan kemampuan verbal yang unggul dan kurang serta tingkat kesulitannya tidak terlalu mudah ataupun terlalu sulit. Berdasarkan statistik *Item Fit*, diperoleh ringkasan analisis untuk tiga subtes verbal sebagai berikut.

Tabel 6. *Rangkuman Hasil Statistik Item Fit*

Keterangan	SE	WA	AN	Total Verbal
Signifikan	17 (85%)	16 (80%)	14 (70%)	47 (78,33%)
Tidak Signifikan	3 (15%)	4 (20%)	6 (30%)	13 (21,67%)

Sebagai analisis tambahan, dilakukan analisis DIF. Chiesi et al. (2018) menyebutkan bahwa analisis DIF digunakan untuk mempelajari kinerja/performa *item* dalam skala, dan memeriksa apakah setiap *item* memiliki peluang yang sama di seluruh subkelompok yang sesuai dengan sifat yang diukur. Oleh karena itu, pada analisis ini, satu butir soal dikatakan teridentifikasi DIF jika terjadi probabilitas yang tidak sama dalam menjawab benar sebuah butir soal pada dua kelompok peserta tes dengan kemampuan sama setelah berada pada kontinum kemampuan yang sama (Ridho, 2013). DIF terjadi bila nilai *Chi-square* signifikan (*p-value* $\leq 0,05$). Adapun kolom *Class* membantu untuk mempermudah identifikasi apakah terjadi DIF atau tidak. Bila tidak terjadi DIF, butir soal termasuk *class* A ($p > 0,05$) dan DIF terjadi pada *class* B (sedang), dan C (tinggi). Tanda '+' pada *class* B atau C menunjukkan butir soal lebih mudah bagi kelompok *focal*, sedangkan tanda '-' menunjukkan butir soal tersebut lebih mudah bagi kelompok *reference*. E.S menunjukkan *common odds ratio*, yaitu menunjukkan berapa kali butir soal tersebut lebih mudah bagi kelompok *focal*.

Adapun dalam penelitian ini, bias *item* dalam dilihat berdasarkan jenis kelamin, yaitu perempuan sebagai *focal* dan laki-laki sebagai *reference*. Pada penelitian ini, diperoleh hasil bahwa IST berbahasa Indonesia subtes verbal terjangkit DIF dengan bias jenis kelamin berjumlah 15 butir soal atau setara dengan 25% dari jumlah keseluruhan.

4. KESIMPULAN DAN SARAN

IST adalah salah satu alat tes untuk mengukur inteligensi yang masih banyak digunakan di Indonesia terutama dalam proses seleksi karyawan. Oleh karena itu, IST harus terus diperhatikan dan dikembangakan agar tetap mampu menjalankan perannya dengan baik sesuai tujuan awal disusunnya. Berdasarkan hasil analisis karakteristik dengan metode *Marginal Maksimum Likelihood* menggunakan pendekatan *Item Response Theory* (IRT) bahwa dari kedua estimasi yang dilakukan, subtes SE, WA, dan AN masih tergolong baik, karena dari 60 butir soal yang dianalisis, 71,67% memiliki kualitas yang cukup baik dengan persentase paramater diskriminasi 100% baik, dan persentase parameter kesukaran 71,67%. Adapun *item* yang termasuk kurang baik daya bedanya adalah *item* se2, se3, se4, se6, se12, se15, se16, wa2, wa3, wa8, an1, an12, an13, an15, an16, an17, dan an19.

Sementara berdasarkan statistik *item fit* juga diketahui alat ukur ini memiliki kualitas yang cukup baik, karena dari 60 *item* yang dianalisis, 78,33% *item* mengikuti model IRT. Adapun *item* yang tidak signifikan itu adalah se2, se11, se14, wa2, wa4, wa9, wa18, an7, an11, an12, an14, an17, dan an18. Hal ini menunjukkan meskipun terdapat dugaan bahwa alat tes IST ini sudah bocor di masyarakat tetapi ternyata kegunaan IST mengukur *verbal comprehension* (V) melalui 3 subtes, yaitu SE (*Satzergaenzung*), WA (*Wortauswahl*), dan AN (*Analogien*) masih dikatakan valid.

Disamping itu, secara umum, *item* pada IST subtes SE, WA, dan AN terindikasi bias jenis kelamin yang relatif rendah. Pada subtes SE, WA, dan AN dengan bias jenis kelamin menunjukkan persentase tidak terjadi DIF sebesar 75%. Adapun *item* yang terindikasi DIF dengan bias jenis kelamin adalah *item* se4, se9, se14, se13, se18, wa1, wa6, wa13, wa5, an1, an13, an16 (DIF sedang), se1, se5, dan wa9 (DIF tinggi). Dengan demikian, dapat dikatakan bahwa alat ukur inteligensi IST berbahasa Indonesia subtes SE, WA, dan AN masih layak digunakan untuk mengukur kemampuan aspek verbal IST berbahasa Indonesia.

Adapun berdasarkan hasil penelitian ini, maka yang dapat disarankan adalah melakukan revisi pada *item-item* IST khususnya subtes SE, WA, dan AN yang memiliki kualitas yang tidak baik, baik dari parameter diskriminasi maupun kesukaran, serta terjangkit DIF agar kedepannya alat tes IST lebih valid lagi dalam mengukur intelegensi aspek verbal. Selain itu, melakukan analisis lanjutan DIF dengan bias yang lain, mengingat peserta tes IST ini dilakukan pada jenjang usia yang kompleks juga khususnya di Indonesia memiliki lintas budaya yang sangat berbeda.

REFERENSI

Aliyu, R. T. (2015). Construct validity of mathematics test items using the rasch model. *International Journal of Social Science and Humanities Research*, *3*(2), 22–28. www.researchpublish.com

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

Becker, K. A. (2003). Stanford-Binet intelligence scales, assessment service bulletin number 1 history of the Stanford-Binet intelligence scales: Content and psychometrics. *Intelligence*, *1*, 14.

Bichi, A. A., Embong, R., Talib, R., Salleh, S., & Bin Ibrahim, A. (2019). Comparative analysis of classical test theory and item response theory using chemistry test data. *International Journal of Engineering and Advanced Technology*, *8*(5 C), 1260–1266. https://doi.org/10.35940/ijeat.E1179.0585C19

ISSN 2579-6348 (Versi Cetak)
ISSN-L 2579-6356 (Versi Elektronik)

Brocke, B., Beauducel, A., & Tasche, K. (1998). Der intelligenz-struktur-test: Analysen zur theoretischen grundlage und technischen güte. *Diagnostica*, *44*(2).

Chiesi, F., Morsanyi, K., Donati, M. A., & Primi, C. (2018). Applying item response theory to develop a shortened version of the need for cognition scale. *Advances in Cognitive Psychology*, *14*(3), 75–86. https://doi.org/10.5709/acp-0240-z

da Silva, C. A. O., Cavalcanti, A. P. R., Lima, K. da S., Cavalcanti, C. A. M., Valente, T. C. de O., & Büssing, A. (2020). Item response theory applied to the spiritual needs questionnaire (SPNQ) in Portuguese. *Religions*, *11*(3). https://doi.org/10.3390/rel11030139

Diputera, A. M. (2018). *Analisis IRT menggunakan Wingen 3: Teori respon butir & aplikasi*. Uwais Inspirasi Indonesia.

Erguven, M. (2013). Two approaches in psychometric process: Classical test theory & item response theory. *Journal of Education*, *2*(2), 23–30.

Fatkhudin, A., Surarso, B., & Subagio, A. (2016). Item response theory model empat parameter logistik pada computerized adaptive test. *Jurnal Sistem Informasi Bisnis*, *4*(2), 121–129. https://doi.org/10.21456/vol4iss2pp121-129

Foster, G. C., Min, H., & Zickar, M. J. (2017). Review of item response theory practices in organizational research: Lessons learned and paths forward. *Organizational Research Methods*, *20*(3), 465–486. https://doi.org/10.1177/1094428116689708

Garcia, E., Aryal, S., Spence-Almaguer, E., Rohr, D., & Walters, S. T. (2018). Use of the IRT model to validate test items from a technology assisted health coaching program. *Open Journal of Statistics*, *8*(3), 519–532. https://doi.org/10.4236/ojs.2018.83034

Gittler, G. (1984). Entwicklung und erprobung eines neuen testinstruments zur messung des räumlichen vorstellungsvermögens. *Zeitschrift Für Differentielle Und Diagnostische Psychologie*, *5*(2).

Guilford, J. P. (1972). Thurstone's primary mental abilities and structure-of-intellect abilities. *Psychological Bulletin*, *77*(2). https://doi.org/10.1037/h0032227

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of Item Response Theory Library*. SAGE Publications.

Ilhan, M., & Guler, N. (2018). A comparison of difficulty indices calculated for open-ended items according to classical test theory and many facet rasch model. *Egitim Arastirmalari - Eurasian Journal of Educational Research*, *2018*(75), 99–114. https://doi.org/10.14689/ejer.2018.75.6

Kumolohadi, R., & Suseno, M. N. (2012). Intelligenz struktur test dan standard progressive matrices: (dari konsep inteligensi yang berbeda menghasilkan tingkat inteligensi yang sama). *Jurnal Inovasi Dan Kewirausahaan*, *1*(2), 79–85.

Le, D. T. (2013). Applying item response theory modeling in educational research. [Disertasi, IOWA State University].

Pathak, A., Patro, K., Pathak, M., & Valecha, M. (2013). Item response theory. *International Journal of Computer Science and Mobile Computing, 2*(11), 7-11.

Rahmawati, E. (2014). Evaluasi karakteristik psikometri intelligenz struktur test (IST). *Proceeding Seminar Nasional Psikometri*, 270–282.

Ridho, A. (2013). Differential item functioning potensi akademik pada kelompok SMA-MA. *Prosiding konferensi ilmiah nasional himpunan evaluasi pendidikan Indonesia (HEPI): Evaluasi Implementasi Kurikulum 2013 dan Sistem Penilaian*. (pp.192-204). Himpunan Evaluasi Pendidikan Indonesia.

Rohmah, S., Kaniawati, I., & Ramalis, T. R. (2018). Analysing PISA-like assessment test measuring scientific literacy using three-parameter logistic (3PL) of IRT-2018. *Journal of Physics: Conference Series*, *1108*(1). https://doi.org/10.1088/1742-6596/1108/1/012084

S., C. E., & Thurstone, L. L. (1938). Primary mental abilities. *The Mathematical Gazette*, *22*(251). https://doi.org/10.2307/3607923

Sadhu, S., & Laksono, E. W. (2018). Development and validation of an integrated assessment for measuring critical thinking and chemical literacy in chemical equilibrium. *International Journal of Instruction*, *11*(3), 557–572. https://doi.org/10.12973/iji.2018.11338a

Schmidt-Atzert, L., & Deter, B. (1993). Intelligenz und ausbildungserfolg: Eine untersuchung zur prognostischen validität des I-S-T 70. *Zeitschrift Für Arbeits- Und Organisationspsychologie*, *37*(2).

Spearman, C. (1904). "General Intelligence", objectively determined and measured. *The American Journal of Psychology*, *15*(2). https://doi.org/10.2307/1412107

Sudaryono. (2011). Implementasi teori responsi butir (Item Response Theory) pada penilaian hasil belajar akhir di sekolah. *Jurnal Pendidikan Dan Kebudayaan*, *17*(16), 719–732.

Veldhuis, M., Matton, N., & Vautier, S. (2014). Using IRT to evaluate measurement precision of selection tests at the french pilot training. *International Journal of Aviation Psychology*, *22* (1), 18-29. https://doi.org/10.1080/10508414.2012.635123

Webster, A. S., & Wechsler, D. (1958). The measurement and appraisal of adult intelligence. *The Journal of Criminal Law, Criminology, and Police Science*, *49*(4). https://doi.org/10.2307/1141601

Xia, J., Tang, Z., Wu, P., Wang, J., & Yu, J. (2019). Use of item response theory to develop a shortened version of the EORTC QLQ-BR23 scales. *Scientific Reports*, *9*(1), 1–10. https://doi.org/10.1038/s41598-018-37965-x