JURNAL ILMU PERILAKU http://jip.fk.unand.ac.id
Volume 5, Nomor 1, 2021 : 92—110
ISSN (Online) : 2581-0421

Analisis Peluang *Guessing* pada *Culture Fair Intelligence Test* (CFIT) 3A dengan Metode IRT

Medianta Tarigan*[*1], Fadillah[2]
[1]Universitas Pendidikan Indonesia, [2]Institut Teknologi Bandung
e-mail: [*1]medianta@upi.edu

***Received**: 26th January 2021/**Revised**: 07th June 2021/**Accepted:** 29th July 2021*

Abstract. *Culture Fair Intelligence Test (CFIT) is a popular intelligence measuring instrument used for psychological assessment purposes. The figural form of a test with multiple choice as in CFIT can increase the chances of guessing behavior performed by the test subject. This behavior can affect the validity of test results in distinguishing quality and under-qualified test subjects. Research related to CFIT psychometric properties is still limited. This study aims to evaluate the validity of CFIT by analyzing the guessing behavior using the Item Response Theory (IRT) method. The sample of this study was 1,955 participants and the results showed that based on IRT 3 PL modeling, the items in the CFIT 3A are still feasible to use. However, CFIT 3A contains items with an unacceptable guessing rate, 12% of the total number of items. The result of model fit test using Confirmatory Factor Analysis (CFA) method as an additional analysis shows that validity construct of CFIT 3A is good. It is expected that the results of the study can provide up-to-date information on the quality of CFIT 3A items and become a reference for other relevant research in the future.*
Keywords: *culture fair intelligence test, item response theory, guessing*

Abstrak. *Culture Fair Intelligence Test* (CFIT) merupakan alat ukur inteligensi yang popular digunakan untuk kepentingan pemeriksaan psikologis. Bentuk tes yang berupa soal gambar dengan pilihan ganda pada CFIT dapat meningkatkan peluang munculnya perilaku menebak yang dilakukan subjek tes. Perilaku ini dapat mempengaruhi validitas hasil tes dalam membedakan subjek tes yang mampu menjawab dan yang tidak mampu menjawab. Penelitian terkait properti psikometrik CFIT masih terbatas, terutama terkait peluang *guessing.* Penelitian kali ini bertujuan untuk melakukan pengujian terhadap dugaan bahwa alat ukur CFIT 3A memiliki tingkat *guessing behavior* tinggi dengan menganalisis menggunakan metode *Item Response Theory* (IRT). Sampel penelitian ini adalah 1,955 partisipan dan hasil penelitian menunjukkan bahwa berdasarkan pemodelan IRT 3 PL, *item - item* dalam alat ukur CFIT 3A masih layak digunakan. Namun, CFIT 3A mengandung *item* dengan tingkat *guessing* yang tidak dapat diterima sebesar 12% dari jumlah *item* keseluruhan. Selain itu, hasil uji kecocokan model dengan teknik *Confirmatory Factor Analysis* (CFA) menunjukkan bahwa CFIT memiliki validitas konstrak yang baik. Diharapkan hasil penelitian dapat memberikan informasi terkini mengenai kualitas *item* CFIT 3A dan menjadi acuan bagi penelitian lain yang relevan di masa yang akan datang.
***Kata kunci*:** *culture fair intelligence test, item response theory, guessing*

Sejak kesuksesan besar Alfred Binet dalam merancang tes untuk membedakan kemampuan antarsubjek tes, instrumen psikometrik terkait inteligensi telah menjadi pembahasan penting dalam dunia psikologi di Eropa dan Amerika (Neisser dkk., 1996). Sebuah hasil tes yang terstandar dapat digunakan sebagai ukuran melihat selisih perbedaan kemampuan dari satu individu dengan individu lainnya (Pham & Chen, 2014). Hasil tes kognitif dapat memberikan informasi yang rinci terkait kemampuan individu serta dijadikan data pendukung dalam pembuatan keputusan, misalnya terkait penentuan program pembelajaran. Tes kognitif juga menjadi banyak digunakan tidak hanya untuk mengukur kecerdasan itu sendiri namun juga mengukur beberapa konstrak terkait misalnya bakat skolastik, prestasi di sekolah, atau kemampuan khusus terkait tugas dan bidang pekerjaan lainnya. Hal ini kemudian berimbas pada meningkatnya perhatian terhadap permasalahan terkait akuntabilitas sebuat alat tes, beberapa tahun terakhir.

Sejalan dengan berbagai konsep inteligensi yang berkembang, instrumen alat ukur yang mengungkap konsep laten inteligensi banyak bermunculan. Terdapat berbagai alat ukur inteligensi yang dibuat berdasarkan teori yang mendasarinya. Diantara alat ukur inteligensi yang populer digunakan adalah *Wechsler Adult Intelligence Scale* (WAIS), Binet, *Culture Fair Intelligence Test* (CFIT), *Wechsler Intelligence Scale for Children* (WISC), *Wechsler Preschool and Primary Scale of Intelligence* (WPPSI), dan *Intelligenz Struktur Test* (IST) (Andriani dkk., 2017).

Salah satu konsep inteligensi diajukan oleh Raymond B. Cattell, yaitu tentang inteligensi umum. Cattell menggolongkan inteligensi menjadi *fluid intelligence* dan *crystalized intelligence* (Cattell, 1963). *Fluid intelligence* adalah kemampuan terkait dengan kapasitas seseorang untuk belajar dan memecahkan masalah baru. Inteligensi ini dipengaruhi oleh faktor biologis, yaitu hereditas (Cattell, 2009). Sedangkan *crystalized intelligence* merupakan pengetahuan dan keterampilan yang diperoleh melalui pendidikan atau pengalaman dari lingkungan seperti melalui pemaparan pengalaman yang dapat menstimulasi intelektual, sekolah, dan sebagainya (Brown, 2016; Cattell, 1951; Suwartono dkk., 2017).

CFIT adalah alat ukur inteligensi yang dikembangkan dari konsep inteligensi Cattle (Larson, 1967). *Culture Fair Intelligence Test* (CFIT) merupakan alat ukur inteligensi yang berbentuk tes kognitif nonverbal yang seringkali dipakai di Indonesia dengan tujuan sebagai tes seleksi (Setiyowati, 2018). Pada tes nonverbal, pertanyaan atau permasalahan serta jawaban atau solusi tidak disampaikan dengan kata-kata, disebut juga sebagai *nonlanguage test* (American Psychological Association, n.d.). Perbedaan tes nonverbal dengan tes bentuk lainnya, menurut Bureau of Exceptional Education dan Student Services (2005), adalah bahwa tes ini memiliki sifat *language-reduced* yaitu subjek tes memahami instruksi tes tanpa harus bergantung pada pemahaman terhadap bahasa tertentu.

CFIT berbentuk *paper-pencil test* yang terdiri dari beberapa *item* soal berbentuk geometris (Nenty & Dinero, 1981), yang dapat digunakan pada subjek tes yang berusia mulai dari 4 (empat) tahun hingga dewasa. Dengan tersedia dalam tiga versi (skala) yang masing-masing terdiri dari dua bentuk yang setara (A dan B), terdiri dari empat subtes yang berbeda (Colom & García-López, 2003; Ruiz, 2009). Adapun ketiga skala tersebut adalah skala 1 untuk subjek tes yang berusia 4 (empat) sampai 8 (delapan) tahun, skala 2 untuk usia 8 (delapan) sampai 12 tahun, dan skala 3 untuk usia 12 tahun ke atas (Marquart & Bailey, 1955). Instruksinya mudah dipahami serta memiliki kualitas soal yang dianggap lebih baik karena sifatnya yang tidak terikat oleh budaya tertentu sehingga dipercaya terhindar dari bias dan kesalahpahaman dalam menjawab pertanyaannya. Sedangkan batas waktu pengerjaan tes ini bervariasi (Ruiz, 2009; Suwartono dkk., 2016).

Dalam pelaksanaannya, partisipan tes yang mengerjakan CFIT berusaha memahami arti dari pola-pola yang memuat informasi visual dan mengenali hubungan antarkonsep dengan penalaran berbasis bahasa yang diinternalisasi melalui gambar visual tersebut. Permasalahan yang muncul kemudian adalah bentuk tes berupa pilihan ganda seperti pada tes CFIT dapat meningkatkan peluang munculnya perilaku menebak yang dilakukan subjek tes. Pada tes ini, partisipan didorong untuk menjawab pertanyaan sebanyak mungkin, terlepas dari apakah mereka memahami persoalan pada tes atau tidak. Pada akhirnya, hal ini juga mendorong subjek melakukan perilaku

menebak (*guessing behavior*) pola-pola yang dihadapinya. Menebak (*guessing*) berarti memberikan jawaban atau membuat penilaian tentang sesuatu tanpa memastikan semua fakta (Obinne, 2012). Terkait dalam pengerjaan tes, dijelaskan oleh Chiu dan Camilli (2013) bahwa *guessing behavior* akan meningkat ketika subjek terdorong untuk mengerjakan sebanyak mungkin persoalan yang belum pasti apakah mereka mengetahui jawaban persoalan tersebut. Dengan adanya *guessing behavior* ini maka perilaku menebak yang meningkat dapat memunculkan jenis kesalahan acak dan bias positif (Rowley & Traub, 1977). Pada akhirnya, perilaku menebak akan mempengaruhi nilai yang diperoleh subjek dalam sebuah tes (Espinosa & Gardeazabal, 2010; Ha, 2017). Atau dengan kata lain, mempengaruhi validitas hasil tes dalam membedakan subjek tes yang berkualitas dan yang kurang berkualitas.

Dalam mendeteksi peluang *guessing,* analisis *Item Response Theory* (IRT) mampu mengestimasi peluang sebuah soal dapat ditebak jawabannya dengan benar. IRT merupakan sebuah pendekatan dalam pengukuran modern yang membahas pengukuran konstruk laten seperti kemampuan atau sikap. Laten disini diartikan sebagai sifat yang tidak dapat diukur secara langsung pada individu dan harus diukur melalui respon-respon yang diberikan individu terhadap *item* atau pertanyaan yang terdapat pada tes. IRT banyak digunakan dalam pengembangan instrumen tes karena modelnya yang memanfaatkan informasi karakteristik *item* untuk kemudian dapat digunakan mengevaluasi dan menyempurnakan instrumen tersebut.

IRT sebagian besar digunakan dalam pendidikan untuk mengkalibrasi dan mengevaluasi *item* dalam tes, kuesioner, dan instrumen lainnya serta untuk menilai subjek pada kemampuan, sikap, atau sifat laten lainnya (An & Yung, 2014). IRT sebagian besar digunakan untuk melihat efektivitas *item* dan respon (Zanon dkk., 2016). Dengan menggunakan model IRT yang tepat, tingkat kemampuan subjek tes dapat diperkirakan secara akurat melalui setiap (sub) set *item* sebagai instrumen yang mengukur kemampuan terkait (Adedoyin & Mokobi, 2013). IRT mendasarkan diri pada sifat-sifat atau kemampuan laten yang mendasari kinerja atau performansi partisipan terhadap *item* tes tertentu. Metode ini bersandar pada 2 (dua) postulat dasar (Hambleton, 1990), yaitu: pertama, performansi partisipan dalam suatu tes dapat diprediksi dengan

sekumpulan faktor yang disebut trait, ciri laten, atau kemampuan; kedua, hubungan antara performansi partisipan dengan sekumpulan trait yang mendasarinya dapat digambarkan dengan fungsi yang meningkat secara monoton yang disebut *Item Characteristic Curve* (*ICC*). Fungsi ini menunjukkan bahwa bila terjadi peningkatan trait, probabilitas jawaban benar juga meningkat. IRT dapat digunakan dengan data skala biner, kategori terurut, dan skala likert, atau nominal, tetapi tidak ordinal. Hasil yang diperoleh dapat mengklasifikasikan orang kedalam kategori kemampuan tertentu seperti master dan non-master atau pakar, perantara, dan pemula (McCowan & McCowan, 1999).

Ringkasnya, kinerja (*performance*) partisipan ujian diposisikan pada setiap tingkat kemampuan. Estimasi parameter *item* dan estimasi kemampuan partisipan ujian direvisi terus menerus hingga diperoleh kesepakatan maksimum yang dimungkinkan antara prediksi berdasarkan kemampuan dan estimasi parameter serta data pengujian aktual (Hambleton, 1990). Ada beberapa asumsi yang harus dipenuhi untuk menentukan apakah IRT merupakan teknik yang tepat untuk digunakan (Sijtsma & Junker, 2006). Asumsi yang harus dipenuhi adalah unidimensi, independensi lokal, dan *item characteristic curve* (ICC) (Rahmawati, 2014; Yang & Kao, 2014). Di samping itu, berdasarkan jumlah parameter yang diestimasi, model IRT dikelompokkan menjadi beberapa jenis, yaitu model logistik 1, 2, dan 3 parameter (Hambleton dkk., 1991). Dalam perhitungannya, IRT melakukan estimasi parameter yaitu parameter diskriminasi (dilambangkan 'a'), tingkat kesulitan (dilambangkan 'b'), dan peluang *guessing* (dilambangkan 'c').

Peluang *guessing* menjadi permasalahan yang sering dipertimbangkan dalam lingkup analisis *item*. Adapun dalam pendekatan klasik, untuk mengatasi peluang *guessing*, skor tes dihitung dengan suatu rumus koreksi yang didalamnya memuat *penalty*, yaitu skor tes sama dengan banyaknya *item* yang benar dikurangi banyaknya *item* yang salah setelah dibagi dengan banyaknya pilihan jawaban dikurangi 1 (satu) (Soekadji, 1999). Sementara dalam IRT, evaluasi tes dan skor *item* didasarkan pada hubungan matematis antara kemampuan dan respon *item* (Chiu & Camilli, 2013; Espinosa & Gardeazabal, 2010). Selain itu, juga dilakukan uji dengan memanfaatkan

pemodelan matematis yang memuat sejumlah parameter. Salah satu parameternya adalah parameter *guessing*. Dengan mengetahui peluang *guessing* diharapkan dapat diketahui kualitas dari sebuah alat ukur.

Sementara itu, hingga kini penelitian terkait analisis *item* CFIT yang menggunakan teknik IRT masih terbatas sehingga belum dapat dipastikan apakah *item - item* yang ada pada CFIT memang memiliki peluang untuk memunculkan *guessing* atau tidak. Dengan mengetahui hal tersebut maka dapat dilihat kualitas alat ukur inteligensi CFIT yang diduga memiliki kualitas yang baik berdasarkan hasil pengukuran kualitas setiap *item* nya. Adapun sebelumnya telah dilaporkan bahwa reliabilitas konsistensi internal CFIT adalah 0.74 dan reliabilitas *retest* (untuk waktu hingga satu minggu) adalah 0.69 (IPAT, 1973). Meskipun nilai-nilai ini lebih kecil dari skor 0.80 yang sering digunakan sebagai standar reliabilitas, nilai ini dapat diterima untuk tujuan penelitian (Gregory, 2011). Validitas CFIT juga pernah diukur dengan menggunakan teknik korelasi yang menghasilkan nilai korelasi sebesar 0.85 dengan faktor "g" dan dengan koefisien validitas kriteria rata-rata 0.66 dengan tes lainnya (Sigal & Mckelvie, 2012).

Selanjutnya, Nurhardini (2017) menganalisis validitas konstruk CFIT dengan menggunakan teknik *Confirmatory Factor Analysis* (CFA). Hasil penelitian tersebut menunjukkan bahwa subtes ke-2 dan subtes ke-4 CFIT tidak mengukur faktor yang diukur atau dengan kata lain model CFIT yang digunakan tidak *fit* sehingga analisis dilanjutkan dengan rekomendasi untuk memodifikasi model. Namun, Nurhardini juga melakukan tinjauan ulang pada data penelitian untuk menjawab permasalahan yang ditemukan. Akhirnya, diperoleh kesimpulan bahwa kedua subtes tersebut tidak mengukur faktor yang seharusnya diukur disebabkan oleh sampel penelitian yang digunakan kurang memiliki keterampilan, bukan terkait pada konsep yang diujikan oleh subtes-subtes terkait.

Penelitian ini bertujuan untuk menguji dugaan bahwa alat ukur CFIT 3A memiliki tingkat *guessing behavior* tinggi dengan melakukan analisis menggunakan metode *Item Response Theory* (IRT). Selain itu, metode *Confirmatory Factor Analysis* (CFA) juga dilakukan untuk melihat validitas konstruk dan menelaah kembali hasil penelitian sebelumnya terkait perlu atau tidaknya melakukan pemodelan ulang dari subtes-subtes

CFIT yang ada. Dengan demikian, diharapkan hasil penelitian dapat memberikan informasi terkini mengenai kualitas *item* CFIT 3A dan menjadi acuan bagi penelitian lain yang relevan di masa yang akan datang.

Metode

Partisipan Penelitian

Subjek yang terlibat dalam penelitian ini berasal dari berbagai kalangan baik akademisi maupun umum yang berusia mulai 16 sampai 40 tahun yang berdomisili di berbagi kota di Indonesia, yaitu Bandung, Yogyakarta, Jakarta, Medan, Padang, Pekanbaru. Adapun pengambilan data dilakukan selama tahun 2017 hingga awal tahun 2020 dan diperoleh subjek penelitian 1995 partisipan. Dengan rata-rata usia partisipan 21.66 tahun dan standar deviasi 5.20 tahun, berikut ini adalah ringkasan sebaran populasi penelitian berdasarkan kota domisilinya.

Tabel 1.
Demografi Subjek Penelitian (dalam %)

Kota	Jumlah%
Bandung	47.06
Bogor	6.91
Cirebon	5.93
Depok	5.47
Jakarta	2.35
Lampung	0.51
Malang	3.73
Medan	3.02
Padang	1.38
Pekanbaru	1.59
Semarang	5.12
Sukabumi	2.25
Tasikmalaya	3.84
Yogyakarta	1.13

Instrumen Penelitian

Instrumen yang digunakan dalam penelitian ini adalah alat ukur inteligensi *Culture Fair Intelligence Test* skala 3 bentuk A (CFIT 3A). CFIT format A ini terdiri dari empat subtes dengan jumlah *item* dan durasi pengerjaan yang berbeda-beda. Informasi mengenai jumlah *item* dan durasi pengerjaan ditampilkan pada Tabel 2.

Tabel 2.
Rincian Subtes CFIT Skala 3

Subtes	Kemampuan yang Diukur	Jumlah Item	Waktu Pengerjaan (menit)
Series	Sistematika berpikir, yaitu kemampuan berpikir runtut untuk memahami rangkaian suatu permasalahan yang berkesinambungan.	13	3
Classification	Ketajaman diferensiasi, yaitu kemampuan untuk mengamati hal-hal yang detil secara tajam dan berpikir dengan kritis untuk mengidentifikasi permasalahan.	14	4
Matrices	Asosiasi, yaitu kemampuan analisis-sintesis untuk menghubungkan dua atau lebih permasalahan yang serupa.	13	3
Topology	Pemahaman konsep, yaitu kemampuan memahami suatu prinsip untuk diterapkan ke dalam situasi yang berbeda.	10	2.5

Adapun, setiap *item* dalam setiap subtes merupakan soal pilihan ganda yang memuat persoalan figural dengan tingkat kesulitan yang semakin meningkat dari satu subtes ke subtes selanjutnya.

Prosedur Penelitian

Pengambilan data dilakukan administrator yang merupakan seorang psikolog dengan memberikan petunjuk administrasi tes sesuai dengan yang terdapat pada manual tes. Tes diberikan secara klasikal dengan batasan tidak lebih dari 30 orang setiap pengambilan data. Data yang diperoleh dari pengetesan kemudian diskor dengan prosedur skoring CFIT 3A, yaitu skor 1 (satu) untuk setiap *item* yang dijawab benar dan skor 0 (nol) untuk *item* yang dijawab salah ataupun tidak dijawab (Wulan, 2010). Skor mentah *(raw score*) adalah skor total yang diperoleh dengan menjumlahkan seluruh jawaban yang benar pada setiap subtes. Selanjutnya skor mentah ini dapat dikonversi menjadi skor IQ berdasarkan norma CFIT (Naderi dkk., 2010).

Analisis Data

Penelitian analisis *item* CFIT ini menerapkan metode penelitian kuantitatif, yakni penelitian yang melibatkan prosedur statistik sebagai pemegang peran sentral dalam mengukur faktor kecerdasan yang diperoleh melalui suatu prosedur tes (Fischer

dkk., 2014). Data skor CFIT 3A berjenis dikotomi (benar dan salah) sehingga dapat dilakukan analisis *item* dengan metode *Item Response theory* (IRT). Analisis *item* dilakukan guna mengidentifikasi kualitas setiap *item* dalam membedakan subjek dengan kemampuan inteligensi tinggi dan rendah. Serta untuk menganalisis parameter *guessing* setiap *item*.

Adapun analisis dilakukan dengan tahapan sebagai berikut: analisis deskriptif; uji kelayakan *item* dengan metode IRT; analisis parameter *guessing*; dan *Confirmatory Factor Analysis* (CFA) sebagai pendukung uji kelayakan *item* dengan metode IRT. Uji multivariat *Confirmatory Factor Analysis* dilakukan untuk melihat validitas konstruk CFIT 3A ini dan juga sebagai bukti empiris asumsi IRT, yaitu unidimensionalitas. Rangkaian proses ini dilakukan untuk mendapatkan bukti empiris terkait *item* mana saja yang masih layak dan tidak layak digunakan berdasar pada kecocokan *item* dengan model serta besar peluang menebak (*guessing*) *item*. Adapun, analisis IRT pada penelitian ini dilakukan dengan bantuan *software jMetrik* versi 4.1.1 dan CFA menggunakan JASP.

Hasil

Analisis Statistik Deskriptif

Analisis kelayakan *item* CFIT 3A yang terdiri dari empat subtes dengan jumlah *item* keseluruhan 50 *item* ini dilakukan dengan subjek penelitian berukuran 1,955 subjek. Adapun hasil ini ditunjukkan sebagaimana yang diringkas pada Tabel 3.

Tabel 3.
Demografi Sampel Analisis Item

Demografi		Ukuran
Ukuran Sampel		1,955 partisipan
Jenis Kelamin	Laki-laki	865 (44%)
	Perempuan	1090 (56%)
Usia	Rentang	16-62 tahun
	Rata-rata	21.66 tahun
	Standar deviasi	5.20 tahun

Berdasarkan hasil yang ditampilkan pada Tabel 4, diperoleh informasi bahwa subtes *Series* adalah subtes yang memiliki rata-rata raw score tertinggi dan subtes *Classification* adalah subtes dengan rata-rata *raw score* terendah.

Tabel 4.
Ringkasan Skor Subtes CFIT

	1 Series	*2 Classification*	*3 Matrices*	*4-Topology*
Rata-rata	7.56	3.72	6.00	4.58
Standar Deviasi	1.74	2.34	2.06	1.81
Min.	0	0	0	0
Max.	13	13	12	10

Selanjutnya, Tabel 5 memberikan gambaran data berupa rata-rata jumlah subjek yang menjawab benar dan salah serta persentase untuk setiap subtes CFIT 3A. Terlihat bahwa partisipan tes terbanyak menjawab benar pada subtes *Series*, yaitu 58% subjek dari keseluruhan. Sedangkan subtes yang dijawab dengan benar paling sedikit adalah subtes C*lassification*, yaitu 27%.

Tabel 5.
Rata-rata Jumlah Subjek Menjawab Benar & Salah

Subtes	Jumlah		Persentase	
	Benar	Salah	Benar	Salah
Series	1137	818	58%	42%
Classification	520	1435	27%	73%
Matrices	902	1053	46%	54%
Topology	896	1059	46%	54%

Berdasarkan hal ini diperoleh hasil yang sejalan pada Tabel 4 dan Tabel 5, yaitu subtes *Series* sebagai subtes dengan rata-rata *raw score* tertinggi dan subtes C*lassification* dengan rata-rata skor terendah, sejalan dengan hasil bahwa rata-rata jumlah partisipan yang dapat menjawab benar pada subtes *Series* adalah sebesar 58% sedangkan pada subtes *Classification* hanya 27%.

Analisis *Item Response Theory*

Setelah menganalisis data skor CFIT 3A, selanjutnya dalam penelitian ini disajikan hasil pengujian kesesuaian *item* dengan model IRT-3PL. Pemilihan jenis model

IRT-3PL didasarkan atas pertimbangan bahwa model ini menghasilkan estimasi parameter *guessing*, yaitu besaran yang menyatakan peluang menebak jawaban yang dilakukan subjek tes terhadap *item-item* CFIT 3A. Pengujian mengandalkan nilai estimasi *chi-square* dan p-value sebagai kriteria pentuan *item* yang fit (signifikan) atau tidak fit (tidak signifikan).

Adapun kriteria tersebut adalah *item* yang memiliki p-value ≥ 0.05 adalah *item* yang diindikasikan sebagai *item* fit dan berlaku sebaliknya. Diperoleh hasil bahwa model IRT 3PL yang menggunakan metode estimator maximum likelihood dan taraf signifikansi 0.05 ini menyatakan bahwa sebanyak 34 *item* tergolong kepada *item* fit. Jumlah ini setara dengan 72% dari jumlah seluruh *item*. Berikut disajikan rekapitulasi hasil analisis.

Tabel 6.
Ringkasan Analisis Item CFIT 3A

	1 Series	*2 Classification*	*3 Matrices*	*4 Topology*
Jumlah *Item*	13	14	13	10
Jumlah *Item* Fit	9 (69%)	14 (100%)	12 (92%)	9 (90%)
No. *Item* Fit	1, 6, 7, 8, 9, 10, 11, 12 13	1 - 14	1, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13	1, 3, 4, 5, 6, 7, 8, 9, 10

Penelitian ini menunjukkan subtes dengan jumlah *item* fit paling banyak dan paling sedikit, berturut-turut adalah subtes *Classification* dan subtes *Series* (subtes pertama). Adapun persentase jumlah *item* fit dari kedua subtes sebagaimana yang ditunjukkan pada Tabel 6 berturut-turut adalah 100% dan 69%. Nilai ini diperoleh dari persentase jumlah *item* fit terhadap jumlah seluruh *item*.

Selanjutnya, nilai estimasi parameter IRT dengan metode estimasi *maximum likelihood* pada *item* CFIT 3A disajikan dalam Tabel 7 berikut. Penggolongan parameter tingkat kesulitan (b) dan daya beda (a) pada penelitian ini mengacu pada kriteria yang dirumuskan oleh Hambleton dkk. (1991), yaitu tingkat kesulitan sedang (tidak terlalu mudah ataupun terlalu sulit) yaitu berada pada range -2 sampai 2, serta daya beda *item* yang baik (normal) berada dalam range 0 sampai 2.

Tabel 7.
Ringkasan Statistik Keseluruhan Item IRT-3PL

		1 Series	*2 Classification*	*3 Matrices*	*4 Topology*
Jumlah *Item*		13	14	13	10
Jumlah Subjek		1,955 partisipan			
Rata-rata		7.56	3.72	6.00	4.58
Rata-rata	a-par	0.87	0.93	0.91	1.45
	b-par	0.18	1.49	0.68	0.87
	c-par	0.25	0.05	0.16	0.16

Adapun terkait estimasi parameter *guessing*, dengan dugaan bahwa tes inteligensi non-verbal dan berjenis pilihan ganda memiliki tingkat guessing behavior lebih tinggi, maka berikut ini diuraikan analisis estimasi parameter guessing alat ukur intelegensi CFIT 3A. *Item* berkualitas baik adalah yang memiliki nilai parameter *guessing* mendekati 0 (Adedoyin & Mokobi, 2013). Dengan kata lain, *item* yang berkualitas baik dipandang dari parameter guessing-nya adalah yang memiliki nilai peluang menebak yang rendah, yaitu *item* dengan guessing yang dapat diterima. Menurut Baker (2001), *item* yang baik memiliki nilai estimasi parameter guessing tinggi sebagaimana yang telah diuraikan pada bagian pendahuluan penelitian ini, yaitu yang berada dalam rentang 0 sampai 0.35. Berikut ini disajikan secara ringkas distribusi *item* berdasarkan kategori parameter guessing menurut Baker untuk setiap subtes CFIT 3A pada Tabel 8.

Tabel 8.
Analisis Parameter Guessing CFIT 3A

Subtes	Dapat Diterima	Tidak Dapat Diterima
Series	9	4
Classification	14	0
Matrices	12	1
Topology	9	1
Total	44 (88%)	6 (12%)

Berdasarkan Tabel 8, diperoleh hasil bahwa sebanyak 44 *item* CFIT 3A termasuk ke dalam kategori *item* dengan parameter guessing yang masih dapat diterima dan 6 (enam) *item* lainnya adalah *item* dengan nilai estimasi parameter guessing yang tidak dapat diterima karena nilainya yang lebih dari 0.35.

Confirmatory Factor Analysis (CFA) CFIT

Tabel 9 menampilkan rangkuman hasil uji kecocokan model data skor CFIT 3A. Berdasarkan hasil uji model fit yang telah dilakukan pada data hasil tes diperoleh bahwa secara keseluruhan hasil uji model adalah *good fit.*

Tabel 9.
Uji Kecocokan Model Secara Keseluruhan

GOF	Tingkat Kecocokan yang dapat diterima	Indeks Model	Keterangan
Chi-square	Semakin kecil semakin baik (p-value ≥ 0.05)	3.453	Good Fit
GFI	GFI ≥ 0.90 good fit	0.999	Good Fit
	0.80 ≤ 0.90 marginal fit		
RMSEA	RMSEA ≤ 0.05 good fit	0.020	Good Fit
NNFI	NNFI ≥ 0.90 good fit	0.995	Good Fit
	0.80 ≤ NNFI ≤ 0.90 marginal fit		
NFI	NFI ≥ 0.90 good fit	0.996	Good Fit
	0.80 ≤ NFI ≤ 0.90 marginal fit		
RFI	RFI ≥ 0,90 good fit	0.987	Good Fit
	0.80 ≤ RFI ≤ 0.90 marginal fit		
IFI	IFI ≥ 0.90 good fit	0.998	Good Fit
	0.80 ≤ IFI ≤ 0.90 marginal fit		
CFI	CFI ≥ 0.90 good fit	0.998	Good Fit

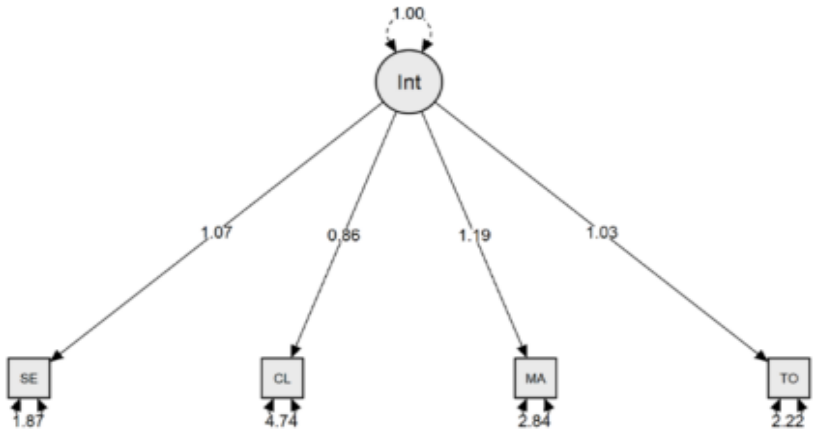
Dengan kriteria bahwa suatu variabel dikatakan mempunyai validitas yang baik terhadap variabel latennya jika nilai t muatan faktornya (*loading factors*) lebih besar dari nilai kritis (atau ≥ 1.96 atau praktisnya ≥ 2) dan muatan faktor standarnya (*standardized loading factor*) ≥ 0.50 dan idealnya 0.7 atau lebih tinggi. Sementara itu, Gambar 1 menunjukkan plot hasil analisis CFA alat ukur intelegensi CFIT 3A.

Tabel 10.
Ringkasan Uji Kecocokan Model Pengukuran

Variabel lain	Subtes	*Std.loading factor*	t-value	p	Keterangan
Intelegensi	Series	0.616	21.356	<.001	Valid
	Classification	0.368	13.719	<.001	Valid
	Matrics	0.577	20.669	<.001	Valid
	Topology	0.569	22.249	<.001	Valid

Gambar 1.

Hasil Uji Kecocokan Model Skor CFIT 3A



Diskusi

Untuk sebuah tes pilihan ganda, masuk akal untuk mengasumsikan bahwa responden menebak ketika mereka yakin bahwa mereka tidak tahu jawaban yang benar (Chiu & Camilli, 2013). Solusi untuk meminimalisir permasalahan ini adalah dengan menguji *item* tersebut, salah satunya dengan menggunakan IRT. Diketahui pada analisis deskriptif statistik sebelumnya, rata-rata skor untuk subtes *Series* dan *Classification*, berturut-turut adalah 7.56 dan 3.72, yang merupakan rata-rata tertinggi dan terendah dari keempat subtes. Juga dapat dilihat dari hasil analisis di bagian sebelumnya bahwa sebanyak 58% partisipan yang mampu menjawab benar subtest *Series*, sementara hanya 27% partisipan yang mampu menjawab benar pada subtest *Classification*. Dengan kata lain, partisipan lebih banyak yang dapat menjawab benar pada subtest *Series* daripada subtest *Classification*. Apabila merujuk pada tabel 4 hasil perolehan data, subtes *Series* adalah subtes yang memiliki skor rata-rata tertinggi dan subtes *Classification* adalah subtes dengan rata-rata skor terendah. Hal ini berarti sebagian besar partisipan lebih menguasai tipe soal pada subtes *Series* daripada ketiga subtes lainnya. Serta, secara umum, partisipan kurang menguasai tipe soal *Classification*.

Berdasarkan hasil analisis data parameter guessing diperoleh bahwa dari sebanyak 44 *item* CFIT 3A dinilai dapat diterima dan layak, sementara 6 (enam) *item*

lainnya adalah *item* dengan nilai estimasi parameter guessing yang tidak dapat diterima karena nilainya yang tergolong tinggi (lebih dari 0.35). Keenam *item* ini, empat diantaranya berada pada subtest *Series*. Hal ini menunjukkan bahwa subtest *Series* yang memiliki tingkat kesukaran item lebih rendah sekaligus memiliki nilai *guessing* yang tinggi. Artinya, *item* pada subtest *Series* perlu dievaluasi kembali keefektifannya. Sebaliknya, *item* pada subtest *Classification* memiliki estimasi parameter guessing yang masih dapat diterima pada keseluruhan item. Hal ini menunjukkan bahwa tipe subtes ini tidak mudah ditebak oleh partisipan. Hal ini memunculkan dugaan bahwa tingkat kesulitan yang diwakili oleh rata-rata skor memiliki hubungan atau mungkin memengaruhi tingkat *guessing behavior* item tersebut. Untuk itu dibutuhkan penelitian lanjutan untuk menyelidiki penyebab dan dinamika tersebut.

Melihat hasil data pengujian kesesuaian item dengan model IRT-3PL dapat ilihat bahwa 72% item dinyatakan fit dengan model IRT 3PL yang menggunakan metode estimator *maximum likelihood* dan taraf signifikansi 0.05. Subtes dengan jumlah *item* fit paling banyak dan paling sedikit, berturut-turut adalah subtes *Classification* sebanyak 14 *item* (seluruh *item*) dan subtes *Series* (subtes pertama) sebanyak 9 (sembilan) *item*. Apabila dilihat secara keseluruhan, rata-rata subtes CFIT memiliki jumlah *item* fit yang cukup baik di atas 90% di tiap subtesnya. Apabila melihat hasil Uji Kecocokan Model Skor CFIT 3A didapat bahwa model tersebut mengukur satu faktor (unidimensional). Artinya, CFIT dengan keempat subtestnya baik *Series, Classification, Matrices, Topology* mengukur satu faktor, yaitu inteligensi.

Sebagaimana tujuan semula dari penelitian ini yaitu memberikan informasi terkini mengenai kualitas *item* CFIT 3A maka hasil ini tentu saja meningkatkan keyakinan bagi para praktisi maupun peneliti untuk menggunakan CFIT 3A sebagai alat ukur inteligensi. Setidaknya, dari sini dapat diketahui bahwa CFIT 3A memiliki nilai guessing yang cukup rendah dengan jumlah *item* tidak layak yang hanya 6 item. Hal ini dapat dibandingkan dengan analisis butir untuk tes IST, yang juga mengukur kognitif, dimana hanya sekitar 50% saja pada butir soal di tiap subtes yang dianggap layak (Agustin & Sirodj, 2018). Apalagi bila dibandingkan dengan kualitas butir *Wonderlic*

*Personnel Test (*WPT), yang juga banyak digunakan untuk mengetahui inteligensi pada seleksi karyawan, yang menunjukkan hanya 62% butir yang dinyatakan layak (Tarigan & Fadillah, 2019).

Kesimpulan

Berdasarkan hasil dan pembahasan penelitian maka dapat diperoleh kesimpulan penelitian bahwa *item* dalam alat ukur inteligensi CFIT masih layak digunakan, menurut model IRT 3 parameter (IRT 3PL). Akan tetapi alat ukur ini mengandung *item* dengan tingkat *guessing* yang tidak dapat diterima sebesar 12% dari jumlah *item* keseluruhan sehingga perlu dilakukan telaah ulang terhadap *item* tersebut. Hasil uji kecocokan model dengan teknik CFA menunjukkan bahwa keempat subtes CFIT mengukur satu faktor yang sama atau bersifat unidimensial yaitu inteligensi. Adapun untuk penelitian selanjutnya, peneliti menyarankan untuk melakukan analisis komparasi tingkat *guessing* antara alat ukur inteligensi berbentuk figural dan non-figural serta menyelidiki secara kualitatif penyebab dan dinamika tingkat kesulitan *item* dan tingkat *guessing item*.

Daftar Pustaka

Agustin, D., & Sirodj, N. (2018). Analisis kualitas item Intelligence Structure Test (IST) melalui metode Item Response Theory (IRT). In schema Journal of Psychological Research (Vol. 4, Issue 2).

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

Andriani, F., Hadi, C., & Paramita, P. P. (2017). Development and Validity of Fluid Intelligence Test Based on Cattle-Horn-Carrol Theory: A Pilot Project. *INSAN Jurnal Psikologi Dan Kesehatan Mental*, *1*(2), 76. https://doi.org/10.20473/jpkm.v1i22016.76-84

Brown, R. E. (2016). Hebb and cattell: The genesis of the theory of fluid and crystallized intelligence. *Frontiers in Human Neuroscience*, *10*(DEC2016). https://doi.org/10.3389/fnhum.2016.00606

Bureau of Exceptional Education and Student Services. (2005). Nonverbal Tests of Intelligence. *Florida Deparemtne of Education Technical Assisstance Paper*, *May*, 1–12.

Cattell, R. B. (2009). Reprinted from The British Journal of Psychology (1946), 36, 159-174: Personality structure and measurement II : the determination and utility of trait modality. *British Journal of Psychology (London, England : 1953)*, *100*(Pt 1A),

233–248. https://doi.org/10.1348/000712608X344807

Cattell, R. B. (1951). *Classical and Standard Score IQ Standardization of The I.P.A.T. Culture Free Intelligence Scale 2*. 154–159.

Cattell, R. B. (1963). Theory of fluid and crystallized intelligence: A critical experiment. *Journal of Educational Psychology*, *54*(1), 1–22. https://doi.org/10.1037/h0046743

Chiu, T. W., & Camilli, G. (2013). Comment on 3PL IRT Adjustment for Guessing. *Applied Psychological Measurement*, *37*(1), 76–86. https://doi.org/10.1177/0146621612459369

Colom, R., & García-López, O. (2003). Secular gains in fluid intelligence: Evidence from the culture-fair intelligence test. *Journal of Biosocial Science*, *35*(1), 33–39. https://doi.org/10.1017/S0021932003000336

America Psychological Association. (n.d.). *Nonverbal Test*. https://dictionary.apa.org/nonverbal-test

Espinosa, M. P., & Gardeazabal, J. (2010). Optimal correction for guessing in multiple-choice tests. *Journal of Mathematical Psychology*, *54*(5), 415–425. https://doi.org/10.1016/j.jmp.2010.06.001

Fischer, H. E., Boone, W. J., Fischer, H. E., & Boone, W. J. (2014). *Quantitative Research Designs and Approaches Quantitative Research Designs and Approaches University of Duisburg-Essen , Essen , Germany Knut Neumann Leibniz-Institute for Science and Mathematics Education (IPN), Kiel , Germany*. *October 2015*.

Gregory, R. J. (2011). *Psychological testing: History, principles, and applications (6th ed.)* (6th ed.). Allyn & Bacon.

Ha, D. T. (2017). *Applying Multidimensional Three-Parameter Logistic Model (M3PL) In Validating a Multiple-Choice Test*. *7*(2), 175–183.

Hambleton, R. (1990). Item response theory: introduction and bibliography. *Psicothema*, *2*(1), 97–107.

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

IPAT. (1973). *Measuring intelligence with the Culture Fair Tests: Manual for Scales 2 and 3*. Institute for Personality and Ability Testing.

Kumara, A. (1989). *Studi Validitas dan Reliabilitas Culture Fair Intellegence Test Skala 3 Sebagai Alat Ukur Inteligensi pada Para Mahasiswa*. Universitas Gadjah Mada.

Larson, S. S. (1967). *DigitalCommons @ USU A Comparison of Two Non-Verbal Intelligence Tests as Predictors of Academic Success of Navajo Students*.

Marquart, D. I., & Bailey, L. L. (1955). An evaluation of the culture free test of intelligence. *Journal of Genetic Psychology*, *86*(2), 353–358. https://doi.org/10.1080/00221325.1955.10532206

Naderi, H., Abdullah, R., Aizan, H. T., & Sharir, J. (2010). Intelligence and academic achievement: An investigation of gender differences. *Life Science Journal*, *7*(1), 83–87.

Neisser, U., Boodoo, G., Bouchard, T. J., Boykin, A. W., Brody, N., Ceci, S. J., Halpern, D. F., Loehlin, J. C., Perloff, R., Sternberg, R. J., & Urbina, S. (1996). Intelligence: Knowns and Unknowns. *American Psychologist*, *51*(2). https://doi.org/10.1037/0003-066X.51.2.77

Nenty, H. J., & Dinero, T. E. (1981). A Cross-Cultural Analysis of the Fairness of the Cattell Culture Fair Intelligence Test Using the Rasch Model. *Applied Psychological Measurement*, *5*(3), 355–368. https://doi.org/10.1177/014662168100500309

Nurcahyo, F. A., Suprapto, M. H., Boeditjahjono, Hosea, J., & Putriadi, G. E. (2019). KORELASI ANTARA CFIT, TES PEMAHAMAN, DAN TES BERHITUNG PADA SISWA KELAS XII DI KEPULAUAN MENTAWAI. *Journal of Chemical Information and Modeling*, *53*(9), 1689–1699. https://doi.org/10.1017/CBO9781107415324.004

Obinne, A. D. E. (2012). Using IRT in Determining Test Item Prone to Guessing. *World Journal of Education*, *2*(1), 91–95. https://doi.org/10.5430/wje.v2n1p91

Pham, T., & Chen, Y.-H. (2014). Cognitive models in educational assessment. In JSM 2014 Proceedings, Social Statistics Section. Alexandria, VA: American Statistical Association. 315-324

Rahmawati, E. (2014). Evaluasi Karakteristik Psikometri Intelligenz Struktur Test (IST). Proceeding Seminar Nasional Psikometri, 270–282.

Rowley, G. L., Traub, R. E. (1977). Formula scoring, number-right scoring, and test-taking strategy. *J Educ Meas, 14*(1),15–22

Ruiz, P. E. (2009). Measuring fluid intelligence on a ratio scale: Evidence from nonverbal classification problems and information entropy. *Behavior Research Methods*, *41*(2), 439–445. https://doi.org/10.3758/BRM.41.2.439

Setiyowati, N. (2018). *Hubungan Antara Kemampuan Intelegensi (Cfit) Dan Potensi Performa Kerja (Dari Hasil Kraepelin Test) Pada Calon Karyawan Bank Swasta Di Jawa Timur*. 5.

Sigal, M. J., & Mckelvie, S. J. (2012). Is Exposure to Visual Media Related to Cognitive Ability? Testing Neisser's Hypothesis for the Flynn Effect. *Journal of Articles in Support of the Null Hypothesis Journal of Articles in Support of the Null Hypothesis. JASNH*, *9*(1), 23–50.

Sijtsma, K., & Junker, B. W. (2006). Item Response Theory: Past Performance, Present Developments, and Future Expectations. *Behaviormetrika, 33*(1), 75–102. https://doi.org/10.2333/bhmk.33.75

Soekadji, S. (1999). *Guessing : Instructed or Discouraged Penalized or Unpenalized ? 2*, 93–105.

Suwartono, C., Amiseso, C. P., & Handoyo, R. T. (2017). Uji Reliabilitas dan Validitas Eksternal The Raven's Standard Progressive Matrices. *Humanitas*, *14*(1), 1. https://doi.org/10.26555/humanitas.v14i1.5772

Suwartono, C., Hidajat, L. L., Halim, M. S., Hendriks, M. P. H., & Kessels, R. P. C. (2016). External Validity of the Indonesian Wechsler Adult Intelligence Scale – Fourth Edition (WAIS-IV-ID). *ANIMA Indonesian Psychological Journal*, *32*(1), 16–28. https://doi.org/10.24123/aipj.v32i1.581

Yang, F. M., & Kao, S. T. (2014). Item response theory for measurement validity. *Shanghai Archives of Psychiatry*, *26*(3), 171–177. https://doi.org/10.3969/j.issn.1002-0829.2014.03.010

Wulan, R. (2010). Peranan Inteligensi, Penguasaan Kosakata, Sikap, dan Minat terhadap Kemampuan Membaca pada Anak. *Jurnal Penelitian Dan Evaluasi Pendidikan*, *14*(2), 166–185. https://doi.org/10.21831/pep.v14i2.1077

Zanon, C., Hutz, C. S., Yoo, H., & Hambleton, R. K. (2016). An application of item response theory to psychological test development. *Psicologia: Reflexao e Critica*, *29*(1). https://doi.org/10.1186/s41155-016-0040-x