GADJAH MADA JOURNAL OF PSYCHOLOGY
VOLUME 6, NO. 1, 2020: 12-28
DOI: 10.22146/gamajop.51852

ISSN 2407-7798 (Online)
https://jurnal.ugm.ac.id/gamajop

Perbandingan Properti Psikometri antara Tes PAPs Berbentuk *Computer-Based* dan *Paper and Pencil Test*

Comparison of Psychometric Properties Between Computer Based and Paper and Pencil Based PAPs Tests

Ariana Marastuti[1], Wahyu Jati Anggoro[2], Ramadhan Dwi Marvianto[3], Abdullah Azzam Al Afghani[4]
[1,2,3,4]Unit Pengembangan Alat Psikodiagnostika, Fakultas Psikologi, Universitas Gadjah Mada

Submitted 25 November 2019 **Accepted** 22 January 2020 **Published** 23 May 2020

Abstract. Current technological developments have offered various more straightforward methods in test administration, one of which is Computer-Based Testing (CBT). CBT was developed as an alternative to the Paper and Pencil Test (PPT). In practice, CBT offers multiple benefits compared to PPT. However, comparisons of the psychometric properties between the two test administration methods need to be investigated further. Research on the parallelism of the two administration models has been scarce, and especially so on tests developed by the Faculty of Psychology of Universitas Gadjah Mada, like the Graduate Academic Potential Test (PAPs Test). The analysis of this research was done by including item difficulty and item discrimination index using the Item Response Theory (IRT) approach, item fit index, and internal structure as evidence of construct validity. The findings show that the PAPs test is categorized as being equal on item difficulty, item discrimination index, and item fit index when administered in either CBT or PPT form. Therefore, in the future, the PAPs test can be administered in both forms alternatively.

Keywords: computer-based test; paper and pencil test; PAPs; psychometric properties

Abstrak. Perkembangan zaman yang diikuti dengan perkembangan teknologi telah menawarkan berbagai kemudahan dalam hal administrasi tes. Salah satunya adalah administrasi tes berbasis komputer atau yang lazim dinamakan dengan *Computer Based Test* (CBT). CBT dikembangkan untuk menjadi alternatif penyelenggaraan tes dengan menggunakan *Paper and Pencil Test* (PPT). Secara praktis CBT memiliki banyak keuntungan dibanding dengan PPT namun perbandingan mengenai properti psikometris pada kedua bentuk tes ini masih perlu ditelaah lebih lanjut. Penelitian mengenai paralelisme kedua model administrasi tes ini belum banyak dilakukan, terutama pada tes-tes yang dikembangkan oleh Fakultas Psikologi UGM seperti Tes Potensi Akademik Pascasarjana (PAPs). Analisis pada penelitian ini dilakukan dengan mengikutsertakan taraf kesukaran dan daya diskriminasi menggunakan pendekatan *Item Response Theory* (IRT), indeks ketepatan model dan struktur pengukuran sebagai bukti validitas konstruk. Temuan dari penelitian ini menunjukkan bahwa Tes PAPs secara umum tergolong memiliki kesetaraan pada parameter taraf kesukaran butir, daya diskriminasi butir dan tingkat ketepatan butir ketika disajikan dalam bentuk CBT dan PPT. Sehingga, Tes PAPs ke depannya dapat disajikan dalam kedua bentuk tersebut secara bergantian.

Kata Kunci**:** computer-based test; paper and pencil test; PAPs; properti psikometri

[1]Korespondensi mengenai artikel ini dapat dilakukan melalui amarastuti@ugm.ac.id
[2]atau ramadhan.dwi.m@mail.ugm.ac.id

Unit Pengembangan Alat Psikodiagnostika (UPAP) selaku pengembang alat tes di Fakultas Psikologi UGM berkewajiban memberikan informasi mengenai tes yang dikembangkan kepada klien, pengguna (*user*) atau masyarakat. Selain properti psikometris tes yang dikembangkan, salah satu informasi yang perlu diberikan adalah prosedur administrasi tes tersebut. Terkait dengan prosedur administrasi ini, UPAP di tahun 2017 sudah melakukan administrasi tes dengan menggunakan *Computer Based Test* (CBT) sebagai alternatif penyelenggaraan tes yang berbasis *Paper and Pencil Test* (PPT). Sebagai upaya untuk memberikan informasi bagi klien dan masyarakat, maka informasi terkait dengan CBT yang dilakukan juga perlu disosialisasikan.

Administrasi tes dengan menggunakan komputer (CBT) mulai jamak dilakukan dalam penyelenggaraan tes. Model administrasi ini memiliki berbagai macam keuntungan secara praktis dibanding dengan model administrasi menggunakan model PPT. Kelebihannya antara lain mudah dalam penyusunan dan administrasinya, pemberian skor butir dan properti butir dapat diketahui secara langsung, serta memungkinkan penambahan *innovative items* (seperti gambar, audio, dan video) (Parshall, Harmes, Davey, & Pashley, 2010).

Meskipun memiliki kelebihan secara praktis, penggunaan model CBT menyisakan banyak pertanyaan dari para peneliti mengenai paralelismenya dengan PPT (Choi, Kim, & Boo, 2003). Paralelisme diperlukan jika suatu tes dipakai sebagai prosedur administrasi tes yang saling melengkapi. Misalnya, CBT dan PPT dipakai secara bergantian yang dilakukan pada populasi individu yang sama dan tujuan yang sama. Jika paralelisme kedua tes berbeda, maka kedua administrasi tes tersebut tidak dapat dilihat sebagai dua metode yang substitutif. Tester sebuah tes harus memberikan prosedur administrasi sesuai dengan tujuan yang diinginkan. Misalnya dalam keperluan seleksi, CBT memberikan kesempatan bagi peserta tes untuk mendapatkan skor yang lebih tinggi pada tes yang melibatkan unsur kecepatan dibanding dengan PPT. Jika ini yang terjadi maka kedua prosedur administrasi tidak dapat dilakukan secara bergantian sebelum ada penyetaraan lebih lanjut (Luecht, 2005).

Beberapa penelitian sudah melakukan perbandingan properti psikometris antara tes yang diadministrasikan melalui CBT dan PPT. Piaw (2012) membandingkan hasil penyekoran antara tes yang berbasis CBT dan PPT. Hasil analisis menunjukkan bahwa tidak ada efek perbedaan skor hasil pengetesan yang signifikan yang ditemukan untuk menguji kinerja dua model pengetesan. Skor tes yang didapatkan oleh sebagian individu konsisten dari administrasi melalui CBT dan PPT. Penelitian ini menunjukkan bahwa seorang peserta di tes dengan menggunakan prosedur administrasi melalui CBT dan PBT kemungkinan besar akan menghasilkan skor yang setara.

Sementara itu, Hosseini, Abidin, dan Baghdarnia (2014) melakukan penelitian yang bertujuan untuk menguji komparabilitas skor tes pemahaman membaca yang menggunakan CBT dan PPT pada mahasiswa bahasa Inggris pertama Iran di Azad University of Tehran, Iran. Tujuan penelitian mereka adalah untuk

memeriksa dampak pengujian berbasis CBT pada hasil skor tes sekaligus mengeksplorasi hubungan antara karakteristik peserta tes tertentu seperti keakraban komputer dan sikap komputer. Dua tes yang setara diberikan kepada peserta pada dua kesempatan yang berbeda. Hasil analisis tidak menemukan keakraban dan sikap terhadap komputer berpengaruh secara signifikan terhadap skor siswa pada CBT dan PPT. Hasil yang sama juga ditemukan pada beberapa penelitian (Flowers, Kim, Lewis, & Davis, 2011; Hardcastle, Herrmann-Abell, & DeBoer, 2017; Pommerich, 2004).

Meskipun banyak penelitian yang juga menemukan perbedaan hasil yang signifikan antara administrasi tes dengan menggunakan CBT dan PPT. Aşkar *et al.* (2012) menemukan hasil yang berbeda. Mereka melakukan penelitian untuk menilai apakah baterai tes yang mengukur konstruk terkait dengan neuropsikologi melalui CBT dapat menghasilkan skor yang relatif sama dengan bentuk PPT. Hasil analisis yang mereka lakukan menunjukkan bahwa skor tes yang dihasilkan dari CBT tidak berkorelasi dengan skor hasil tes dengan menggunakan PPT. Skor dari administrasi dengan menggunakan model PPT lebih tinggi pada CBT. Perbedaan skor total antara modalitas secara statistik signifikan untuk kedua tes-tes yang diujikan. Mereka beranggapan bahwa penyebabnya adalah jenis peng-administrasian yang berbeda memberikan waktu pengerjaan yang berbeda pula yang berdampak pada skor perolehan individu. Hal yang sama juga terjadi pada penelitian yang lain (Vrabel, 2004).

Diskursus yang dikemukakan di atas menunjukkan bahwa masih terjadi diskrepansi hasil analisis antara tes yang diadministrasikan melalui CBT dan PPT. Perbedaan dan kesamaan skor hasil pengukurannya tergantung dari faktor-faktor yang kontekstual. Hal ini dikarenakan ada banyak hal-hal yang memengaruhi, misalnya konstruk yang diukur, fitur CBT yang dipakai hingga prosedur penyekoran. Berangkat dari permasalahan ini maka identifikasi terhadap penyelenggaraan PAPs berbasis pada CBT dan PPT merupakan penelitian yang harus dilakukan. Pertimbangannya adalah PAPs merupakan tes yang dipakai untuk keperluan *high stake* yang menyangkut masa depan individu. Di sisi lain, hasil dari identifikasi ini sangat diperlukan untuk memberikan rekomen-dasi pada pengembang Tes PAPs untuk meningkatkan kualitas tes yang dikembangkannya.

Pengadministrasian tes dengan menggunakan komputer adalah *state of the art* penyelenggaraan tes psikologi saat ini. Model pengadministrasian tes ini merupakan salah satu upaya untuk memperbaiki kelemahan tes yang dilakukan dengan menggunakan tes berbasis kertas dan pensil (PPT). Model pengadministrasian ini memiliki beberapa tipe antara lain *Computer-Assisted Assessment* atau *Computer-Aided Assessment* (CAA), *Computer-Mediated Assessment* (CMA), *Computer-Based Assessment* (CBA), *online assessment*, *Computer Based Test* (CBT). Meskipun istilah-istilah ini umumnya digunakan secara bergantian, mereka memiliki arti yang berbeda. *Computer Assisted / Mediated Assessment* mengacu

pada aplikasi komputer apa pun dalam proses administrasi tes. Peran komputer dapat bersifat ekstrinsik atau intrinsik sehingga padanan kata untuk term *e-assessment* juga menggambarkan berbagai kegiatan yang berhubungan dengan komputer. Dalam definisi ini komputer sering tidak berperan dalam penilaian terhadap jawaban atau respons yang diberikan oleh individu yang sebenarnya, akan tetapi hanya memfasilitasi penangkapan (*capture*) dan transfer tanggapan antara peserta tes dengan penyelenggara tes.

State of the art dari properti psikometri sebuah tes adalah teori tes modern atau yang lebih dikenal dengan teori respons butir (*Item Response Theory*/IRT). IRT melengkapi keterbatasan tes teori klasik (*Classical Test Theory*/CTT) yang banyak memiliki keterbatasan yang terkait dengan konsistensi parameter yang dihasilkan. Parameter butir (misalnya daya beda butir) yang dihasilkan oleh CTT sangat bergantung kepada karakteristik sampel sehingga generalisasi parameter butir yang dihasilkan oleh analisis berbasis pada teori ini memiliki banyak keterbatasan dibanding dengan analisis berbasis pada IRT. IRT saat ini perkembangannya sangat cepat karena keberadaan teknologi telah membantu banyak peneliti menggunakan pendekatan ini yang sebelumnya sempat terhenti (di tahun 1950-an) karena tidak adanya komputer yang cukup mampu untuk melakukan proses estimasi butir maupun kemampuan individu yang bersifat iteratif.

Properti psikometri tes

Properti psikometri adalah karakteristik tes yang dapat menjelaskan atribut sebuah tes baik dari sisi pengukuran, skor maupun fitur-fiturnya. Properti psikometri memberikan informasi mengenai kualitas sebuah tes, baik di level butir maupun di level tes (Leong, 2008). Contoh properti tes adalah reliabilitas dan validitas. Sebuah tes dikatakan bagus dan bisa dipakai apabila skor hasil pengukuran yang dilakukan memiliki konsistensi yang tinggi dan dapat menggambarkan apa yang diukur oleh tes tersebut (Ginty, 2013). Reliabel berarti pengulangan pada subjek yang sama akan menghasilkan skor yang konsisten sementara valid berarti tes benar mengukur variabel yang ingin diukur. Properti psikometri merupakan kualitas tes secara kuantitatif, yang dapat dihitung menggunakan metode statistik. Ada banyak jenis properti psikometri sebuah alat tes. Properti psikometri sebuah alat tes ditinjau dari reliabilitas dan validitas (Ginty, 2013), dengan parameter butir yang di dalamnya memuat tingkat kesulitan dan daya beda butir dan parameter tes yang di dalamnya memuat distribusi skor dan struktur faktor. Properti psikometri di level butir menjelaskan kualitas masing-masing butir dalam mengukur variabel yang telah ditetapkan.

Properti psikometri di level tes merupakan properti komposit dari butir-butir yang terdapat dalam tes tersebut. Properti psikometri tes merupakan kualitas tes secara kuantitatif yang dapat dihitung dengan menggunakan metode statistik. Properti psikometri tes hanya menjadi salah satu dari banyak properti tes yang lain.

Terdapat misalnya properti tes dari segi validitas alat ukur yang didapatkan melalui *expert judgment*. Properti seperti ini tidak didapatkan melalui metode statistika.

Berdasarkan literatur yang ada properti psikometri yang perlu untuk dikaji ketika mengevaluasi kesetaraan antara pengadministrasian tes melalui CBT maupun PPT adalah: a) tingkat kesulitan butir, b) daya diskriminasi butir, c) indeks ketepatan butir pada model, dan d) struktur pengukuran/validitas konstruk (Chen & Lei, 2005; Piaw, 2012; Williamson, Mislevy, & Bejar, 2006).

Pertama, tingkat kesulitan butir menunjukkan seberapa jauh butir tersebut dapat dijawab dengan benar oleh peserta tes. Dari perspektif CTT, butir yang mudah adalah butir yang banyak dijawab benar oleh individu peserta tes sedangkan dari sisi IRT butir yang mudah adalah butir yang memiliki probabilitas tinggi untuk dijawab benar pada individu yang memiliki kemampuan rendah. Oleh karena individu yang memiliki kemampuan rendah memiliki probabilitas yang tinggi dalam menjawab benar, individu yang memiliki kemampuan yang tinggi tentunya akan memiliki probabilitas yang lebih besar. Isu mengenai kesetaraan tingkat kesulitan butir ini merupakan isu pertama yang paling fundamental dalam studi mengenai kesetaraan pengukuran antara CBT dan PPT (Mead & Drasgow, 1993). Penyebabnya dapat bervariasi namun intinya adalah seberapa jauh perbedaan fitur dan fasilitas di dalam CBT akan membantu individu untuk menjawab butir yang diberikan dengan benar (Keller, Swaminathan, & Sireci, 2003). Misalnya, penyajian gambar sangat jelas pada CBT dapat mempermudah butir karena individu dapat memahami soal dengan tepat.

Kedua, parameter yang kedua dalam penelitian adalah daya diskriminasi butir. Daya diskriminasi butir menunjukkan seberapa besar sebuah butir mampu membedakan individu berdasarkan kemampuan yang diukur (Crocker & Algina, 1986). Meskipun dalam teori tes yang menggunakan Model Rasch, parameter ini tidak mendapatkan perhatian yang berarti, daya diskriminasi memberikan dampak langsung pada skor peserta tes. Dalam penelitian yang menelaah CBT, parameter daya beda banyak dilibatkan sebagai objek yang diteliti (The National Assessment of Educational Progress, 2014). Penyebab sebuah butir akan memiliki daya diskriminasi yang berbeda ketika dihasilkan dari CBT dan PPT adalah faktor-faktor individu sekaligus tes yang dipakai. Dari sisi tes yang dipakai, faktor kejelasan instruksi atau tugas yang diberikan kepada peserta tes untuk diselesaikan sangat memengaruhi daya diskriminasi tes. Instruksi tes yang jelas akan meningkatkan bahwa individu yang memiliki kemampuan tinggi akan dapat menjawab soal tersebut dan individu yang memiliki kemampuan rendah tidak akan dapat menjawab soal tersebut. Hal ini berbeda dengan ketika sebuah butir disajikan dalam bentuk yang tidak familier. Orang yang memiliki kemampuan tinggi dan rendah akan menjawab soal dengan salah karena keduanya kurang memahami maksud dari soal yang diberikan. Akibatnya, daya diskriminasi butir terhadap individu berdasarkan kemampuan yang diukur menjadi rendah.

Salah satu properti yang menggabungkan informasi mengenai kedua parameter di atas secara simultan adalah tingkat ketepatan butir dengan model (*model fit*). Ketepatan model dengan butir menunjukkan performansi butir. Model adalah kondisi ideal yang menunjukkan bagaimana kinerja sebuah butir. Model tersebut secara tidak langsung menunjukkan beberapa hal yang terkait dengan tingkat kesulitan, daya beda dan munculnya tebakan semu (Janssen, Meier, & Trace, 2014).

Properti keempat yang biasa dikaji adalah struktur pengukuran atau yang lebih sering dinamakan dengan dimensionalitas pengukuran oleh sebuah tes. Struktur pengukuran adalah informasi yang menunjukkan berapa dimensi atau aspek yang ada di dalam pengukuran sebuah tes dan bagaimana hubungan antar dimensi atau aspek tersebut (Bacharach & Furr, 2007; Tate, 2003). Evaluasi terhadap struktur pengukuran harus menjadi bagian penting dari pengembangan, evaluasi, dan pemeliharaan tes dalam skala besar. Evaluasi tersebut dapat memberikan dukungan empiris terhadap konten dan proses kognitif yang melandasi skor yang didapatkan oleh peserta tes. Evaluasi terhadap struktur ini dilakukan dengan beberapa tujuan, misalnya untuk menguji kekhawatiran tentang kemungkinan adanya pelanggaran asumsi unidimensionalitas. Penyelenggaraan dengan menggunakan model administrasi yang berbeda memungkinkan untuk mengubah struktur pengukuran, misalnya dari struktur unidimensi menjadi multidimensi atau sebaliknya. Meskipun ada keyakinan bahwa dimensi yang terkait dengan konten ukur tetap kuat ketika tes diadministrasikan dengan cara yang berbeda, namun beberapa penelitian menemukan kemungkinan munculnya dimensi baru yang tidak direncanakan atau disengaja untuk diukur. Konsekuensi pelanggaran ini salah satunya adalah ancaman (*threat*) terhadap aspek konten validitas jika dimensi yang muncul dari analisis faktor tidak relevan atribut yang diukur dan reliabilitas yang dilaporkan.

Penelitian ini bertujuan untuk membandingkan properti psikometri Tes PAPs antara yang diadministrasikan melalui CBT dan PPT. Tes PAPs yang dianalisis adalah seri E1 dan E2 Tes PAPs yang dimiliki oleh Fakultas Psikologi UGM karena Tes PAPs seri tersebut sudah diadministrasikan dalam bentuk CBT dan PPT. Properti psikometri butir dan tes yang dikaji adalah semua properti psikometri yang berbasis pada teori tes modern (*Item Response Theory*).

Penelitian ini mengajukan beberapa pertanyaan terkait dengan konsistensi (invariansi) parameter butir dan struktur dimensionalitas tes ketika Tes PAPs diadministrasikan dalam bentuk CBT dan PPT. Pertanyaan yang diajukan adalah seberapa konsisten parameter butir dan struktur pengukuran oleh Tes PAPs ketika diadministrasikan melalui CBT dan PPT. Berdasarkan beberapa fitur yang ada dalam Tes PAPs beberapa jawaban diharapkan muncul. Tingkat kesulitan butir, daya beda butir maupun ketepatan butir dengan model IRT diharapkan tidak menghasilkan perbedaan yang signifikan. Struktur tes yang dihasilkan pengukuran dengan

menggunakan CBT maupun PPT tidak banyak berubah. Invariansi struktur tersebut mencakup invariansi parameter fit yang dimiliki pada bentuk CBT dan PBT.

Keyakinan peneliti untuk menemukan konsistensi dan invariansi parameter butir dan struktur pengukuran Tes PAPs dilakukan karena beberapa pertimbangan. Pertama, fitur-fitur Tes PAPs versi CBT tidak jauh berbeda dengan PPT. Bentuk dan ukuran huruf tidak memilik perbedaan yang cukup besar. Penyediaan nomor-nomor butir di sebelah soal yang tengah dikerjakan akan mempermudah individu untuk berpindah ke soal-soal tersebut. Adanya fitur ini menjadikan prosedur pengerjaan di CBT tidak berbeda jauh dengan PPT yang memungkinkan peserta untuk mengerjakan soal tidak berdasarkan urutan nomor butir. Kedua, prosedur administrasi pada CBT memilik kemiripan dengan PPT, misalnya waktu pengerjaan, jumlah butir, serta urutan dan jenis sub tes yang dilibatkan. Adanya kesetaraan ini diasumsikan akan mendukung invariansi struktur Tes PAPs karena efek kontekstual (peletakan urutan butir yang biasa memunculkan dimensi ukur baru) tidak memberikan dampak pada struktur pengukuran.

Metode

Partisipan penelitian

Partisipan penelitian ini adalah calon mahasiswa pascasarjana yang mengikuti Tes PAPs pada tahun 2016 dan 2018. Pemilihan partisipan pada periode ini dilakukan karena pelaksanaan CBT pada administrasi Tes PAPs dilakukan sejak tahun 2016.

Variabel penelitian

Variabel penelitian ini adalah properti psikometri butir Tes PAPs dan struktur pengukuran Tes PAPs. Properti psikometri yang dikaji adalah tingkat kesulitan butir, daya diskriminasi butir dan tingkat ketepatan model dengan butir. Struktur pengukuran yang dikaji adalah korelasi skor pada setiap dimensi dan tingkat kesesuaian model antara Tes PAPs dengan bentuk CBT dan PPT.

Instrumen penelitian

Instrumen yang dipakai dalam penelitian ini adalah seri E1 dan E2 Tes PAPs. Seri ini dipilih karena telah diadministrasikan secara CBT dan PPT. Jumlah total soal pada kedua seri tersebut 150 butir yang terdiri dari 3 sub tes, yaitu verbal, kuantitatif dan figural.

Prosedur penelitian

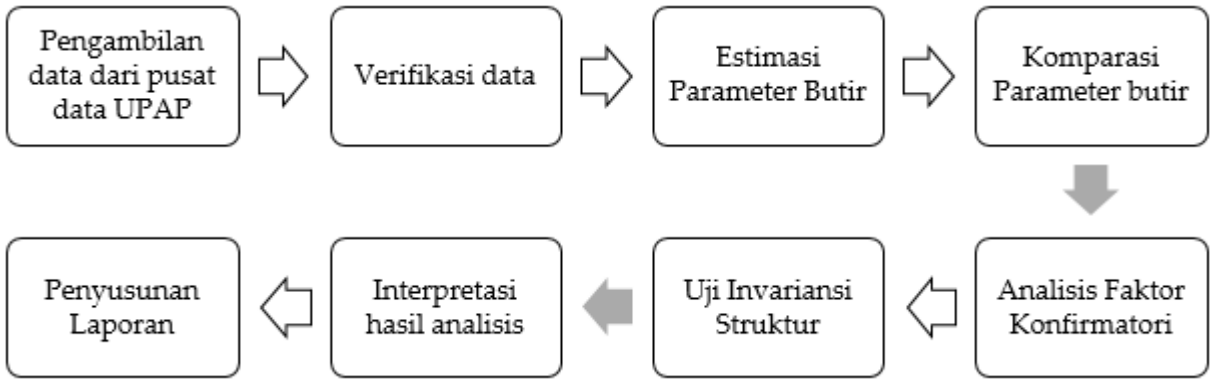
Penelitian ini dilakukan dalam beberapa tahap. Tahap pertama adalah pengumpulan *database* yang memuat kegiatan mengumpulkan semua data-data pendukung yang akan dipakai untuk penelitian. Tahap kedua adalah analisis data yang melibatkan kegiatan pemberian skor dengan menggunakan teori Rasch untuk mendapatkan parameter butir. Proses ini dilakukan dengan bantuan program analisis berbasis *R* dengan menggunakan paket analisis *ltm* (Rizopoulos, 2007) dan *irtoys* (Partchev, 2009).

Tahap selanjutnya adalah melakukan analisis faktor untuk mendapatkan struktur pengukuran Tes PAPs melalui CBT dan melalui PPT. Hasil dari identifikasi tersebut kemudian dibandingkan untuk

mengidentifikasi invariansi struktur pengukuran dengan menggunakan CBT dan PPT. Pada tahap akhir penulis merangkum dan menganalisis keseluruhan hasil analisis yang dilakukan. Prosedur analisis selengkapnya dapat dilihat pada gambar 1.

Analisis data

Penelitian ini dilakukan dalam dua tahap. Tahap pertama adalah perbandingkan parameter butir antara Tes PAPs yang diadministrasikan dari CBT dan PPT. Analisis data dilakukan dengan menggunakan teori tes modern yaitu Model Rasch untuk mengidentifikasi properti psikometri masing-masing Tes PAPs.



Gambar 1. Diagram alur pelaksanaan penelitian

Sementara parameter butir yang akan dicari yaitu tingkat kesulitan butir, daya beda butir dan tingkat ketepatan model butir. Selanjutnya, tahap kedua adalah menguji konsistensi struktur pengukuran antara pengadministrasian Tes PAPs melalui CBT dan PPT dengan analisis faktor konfirmatori (CFA)

Formulasi pengukuran Model Rasch adalah menggunakan matriks data yang berisi jawaban dari responden *n* (dinotasikan sebagai *n*=1,2,...,*N*) dan satu set butir i (butir *i*=1,2,...,*L*) yang mengukur *latent trait* yang sama (Wright & Mok, 2004). Untuk data yang berbentuk dikotomi, permodelan Rasch menggabungkan suatu algoritma yang menyatakan hasil ekspektasi probabilistik dari butir *i* dan responden *n* yang secara sistematis dinyatakan sebagai berikut. (Bond & Fox, dalam Sumintono & Widhiarso, 2013).

$$P_{ni}=(x_{ni}=1 \mid \beta_n,\delta_i) = \frac{e^{\beta n-\delta i}}{1+e^{\beta n-\delta i}}$$

$P_{ni}=(x_{ni}=1 \mid \beta_n,\delta_i)$ adalah probabilitas dari responden *n* dalam butir *i* untuk menghasilkan jawaban betul (x_{ni}=1) dengan kemampuan responden β_n dan tingkat kesulitan butir δ_i. Persamaan tersebut dapat disederhanakan dengan memasukkan fungsi logaritma dan menjadikannya:

$$\text{Log}(P_{ni}(x_{ni}=1 \mid \beta_n,\delta_i))= \beta_n-\delta_i$$

Nilai inilah yang disebut logit atau *W-score* atau nilai *measure*. Nilai logit pada butir soal dapat disebut juga sebagai taraf kesukaran (Bond & Fox, 2015).

Berikutnya daya beda atau daya diskriminasi. Daya diskriminasi dalam

Tabel 1.
Taraf Kesukaran Butir Tes PAPs E1 dan E2

Bentuk Tes	Bentuk Penalaran	Taraf Kesukaran E1	Taraf Kesukaran E2
CBT	Figural	-0,06 – 5,87	-0,15 – 3,69
	Kuantitatif	-8,35 – 8,10	-2,12 – 9,76
	Verbal	-4,09 – 6,41	-10,72 – 7,31
PPT	Figural	-17,98 – 18,65	-10,17 – 8,93
	Kuantitatif	-9,18 – 8,63	-4,77 – 11,21
	Verbal	-10,17 – 13,27	-6,25 – 4,56

Rasch pada dasarnya sama dengan korelasi poin biserial (*rpbis*) pada teori klasik, kecuali pada Rasch menggunakan nilai *measure* sementara pada teori klasik menggunakan skor total. Interpretasi bebas mengenai rentang nilai daya diskriminasi menurut Alagumalai dan Curtis (2005) adalah sangat bagus (>0,40), bagus (<0,39, >0,30), cukup (<0,29, >0,20), tidak mampu mendiskriminasi (<0,19, >0,00), dan membutuhkan pemeriksaan terhadap butir (<0,00). Daya diskriminasi yang optimal adalah mendekati 0,5 (Linacre, 2011)

Parameter butir berikutnya adalah ketepatan butir dengan model. Hal yang biasa digunakan dalam menentukan apakah butir fit atau tidak dengan model didasarkan pada dua aspek, yaitu *infit* (*information-weighted fit*) dan *outfit* (*outlier-sensitive fit*). Melalui *infit* dan *outfit* dapat diketahui *mean square* tidak terstandar (MNSQ) dan bentuk terstandar (ZSTD) (Bond & Fox, 2015). Setiap *outfit* menunjukkan nilai *mean square* (MNSQ) dan *z-score* (ZSTD). Nilai MNSQ yang ditoleransi berada pada rentang 0,5 s.d. 1,5 sementara nilai ZSTD berada pada rentang -2 s.d. +2. Butir yang memiliki nilai *outfit* yang berada di luar batas toleransi tersebut adalah butir yang *misfit* (Bond & Fox, 2015).

Hasil

Tingkat kesulitan butir

Taraf kesukaran menunjukkan seberapa jauh butir dapat dijawab dengan benar oleh peserta tes. Berdasar Tabel 1, sub tes figural Tes PAPs dalam bentuk PPT memiliki taraf kesukaran yang lebih bervariasi (seri E1 -17,98 hingga 18,65; seri E2 -10,17 hingga 8,93) dibandingkan bentuk CBT (seri E1 -0,06 hingga 5,87; seri E2 -0,15 hingga 3,69) pada semua seri. Sub tes kuantitatif menunjukkan hasil yang cukup berbeda. Pada seri E1, bentuk CBT dan PBT tergolong memiliki taraf kesukaran yang setara yaitu -8,35 hingga 8,10 dan -9,18 hingga 8,63 secara berturut-turut.

Berbeda dengan dua sub tes sebelumnya, sub tes verbal pada seri E1 menunjukkan taraf kesukaran yang lebih bervariasi pada bentuk PPT (-10,17 hingga 13,27) dibanding pada bentuk CBT (-4,09 hingga 6,41). Di sisi lain, pada seri E2 bentuk CBT memiliki taraf kesukaran yang lebih bervariasi (-10,72 hingga 7,31). Secara umum, ketiga sub tes di atas dapat dikatakan setara pada dua bentuk tes dan dua seri yang telah disajikan.

Daya diskriminasi butir

Daya diskriminasi butir adalah kemampuan butir dalam membedakan antara subjek yang memiliki kemampuan tinggi dengan subjek yang memiliki kemampuan rendah

Tabel 2.
Daya Diskriminasi Butir Tes PAPs E1 dan E2

Bentuk Tes	Bentuk Penalaran	Daya Diskriminasi E1	Daya Diskriminasi E2
CBT	Figural	0,40 – 5,82	0,31 – 8,12
	Kuantitatif	0,11 – 6,82	0,06 – 4,62
	Verbal	0,16 – 1,60	0,12 – 1,57
PPT	Figural	0,05 – 4,37	0,04 – 4,87
	Kuantitatif	0,08 – 4,80	0,04 – 4,31
	Verbal	0,03 – 4,03	0,05 – 4,74

Tabel 3.
Tingkat Ketepatan Model Tes PAPs E1 dan E2

Bentuk Tes	Bentuk Penalaran	*Item Fit* E1	*Item Fit* E2
CBT	Figural	0,55 – 2,38	0,51 – 2,2
	Kuantitatif	0,54 – 1,81	0,62 – 2,09
	Verbal	0,73 – 1,25	0,74 – 1,45
PPT	Figural	0,55 – 2,31	0,56 – 1,95
	Kuantitatif	0,58 – 1,67	0,66 – 1,76
	Verbal	0,51 – 2,44	0,47 – 2,12

(Azwar, 2013). Berdasarkan Tabel 2, daya diskriminasi pada sub tes figural seri E1 dan E2 menunjukkan bahwa daya diskriminasi CBT (0,40 – 5,82 untuk seri E1; 0,31 – 8,12 untuk seri E2) tergolong sedikit lebih baik dibandingkan PPT (0,05 – 4,37 untuk seri E1; 0,04 – 4,87 untuk seri E2. Bentuk CBT memiliki nilai terendah di atas 0,3 dibandingkan PPT yang hanya berada di atas 0,04.

Sub tes kuantitatif pada kedua bentuk tes, CBT dan PPT, memiliki daya diskriminasi yang setara di mana nilai terendah berada di atas 0,04 dan nilai tertinggi selalu di atas 4. Sedangkan pada sub tes verbal, bentuk CBT memiliki nilai terendah yang sedikit lebih besar yaitu di atas 0,12 namun demikian nilai tertinggi dari daya diskriminasi hanya di bawah 1,57. Nilai daya diskriminasi pada sub tes verbal dalam bentuk PBT di semua seri memiliki nilai terendah yang di bawah nilai CBT namun memiliki nilai tertinggi jauh di atas bentuk CBT yaitu di atas 4. Namun demikian, secara keseluruhan butir dalam Tes PAPs memiliki daya diskriminasi yang memuaskan.

Tingkat ketepatan model

Tabel 3 menunjukkan indeks ketepatan butir dengan model (*item-fit*) Tes PAPs seri E1 dan E2. Pada sub tes Figural, tampak nilai *item-fit* pada seri E1 dan E1 pada tes berbentuk CBT (0,55 – 2,38 untuk E1; 0,51 – 2,2 untuk E2) dan PPT (0,55 – 2,38 untuk E1; 0,51 – 2,2 untuk E2) tergolong setara atau tidak jauh berbeda. Hal ini diikuti sub tes kuantitatif di mana bentuk CBT pada seri E1 (0,54 – 1,81) dan E2 (0,62 – 2,09) tidak

terlalu berbeda jauh dengan PPT di seri E1 (0,58 – 1,67) dan E2 (0,66 – 1,76).

Cukup berbeda hasil yang didapat dalam sub tes verbal. Pada bentuk CBT, *item-fit* pada seri E1 dan E2 berada di 0,7 hingga 1,4 yang di mana kedua nilai tersebut cukup berbeda dengan bentuk PPT di E1 dan E2 yang memiliki nilai terendah 0,51 dan 0,47 masing-masing serta tertinggi 2,44 dan 2,12 secara berurutan. Akan tetapi, butir-butir yang ada dalam setiap subtes memiliki *item-fit* yang tidak jauh berbeda berdasarkan bentuk tesnya.

Sub tes kuantitatif pada seri E1 menunjukkan hubungan yang kuat dengan nilai korelasi 0,969 ($p < 0,01$) , demikian juga dengan sub tes kuantitatif seri E2 dengan signifikasi 0.878 dan sub tes figural dengan nilai korelasi 0,864 ($p < 0,01$). Sementara itu hasil sub tes figural pada seri E2 menunjukkan hasil moderat dengan nilai signifikansi 0,757 ($p < 0,01$). Hasil pada sub tes verbal meski menunjukkan adanya signifikasi pada kedua seri, yaitu 0,579 ($p < 0,01$) pada seri E1 dan 0,659 ($p < 0,01$) pada seri E2. Maknanya, kedua bentuk tes tersebut dapat dikatakan sebagai konstruk yang serupa karena memiliki korelasi positif yang tinggi. Dengan kata lain, hal ini menunjukkan bukti validitas konvergen kedua bentuk tes tersebut.

Tabel 4.

Hasil Uji Korelasi Antara Seri E1 dan E2 Tes PAPs *Computer Based Test* dan *Paper and Pencil Test*

			Paper and Pencil Test					
			Seri E1			**Seri E2**		
			Verbal	**Penalaran**	**Kuantitatif**	**Verbal**	**Penalaran**	**Kuantitatif**
Computer Based Test	**Seri E1**	**Verbal**	0,579**					
		Penalaran		0,864**				
		Kuantitatif			0,969**			
	Seri E2	**Verbal**				0,659**		
		Penalaran					0,757**	
		Kuantitatif						0,878**

** Korelasi signifikan di bawah 0,01 (2-*tailed*)

Struktur pengukuran

Selain melihat parameter butir, korelasi antar bentuk tes berdasar sub tesnya dilakukan untuk membuktikan bukti validitas konvergen yang muncul dalam Tes PAPs (dapat dilihat pada tabel 4). Hasil Korelasi antara hasil tes berbentuk CBT dan PPT dengan menggunakan dua seri PAPs (seri E1 dan E2) menunjukkan hasil yang signifikan antara semua seri antara CBT dan PPT.

Tabel 5.

Indeks Ketepatan Model (*Goodness Of Fit Indices*) Pengukuran antara Seri E1 dan E2 Tes PAPs *Computer Based Test* dan *Paper and Pencil Test*

Parameter	E1 *Computer Based Test*	E1 *Paper and Pencil Test*	E2 *Computer Based Test*	E2 *Paper and Pencil Test*
Chi-Square	599,410	429,157	621,864	593,899
p-value	0,000	0,000	0,000	0,000
CFI	0,879	0,968	0,861	0,950
TLI	0,847	0,960	0,825	0,938
RMSEA	0,107	0,066	0,109	0,079
SRMR	0,078	0,031	0,076	0,045

Tabel 5 meringkas hasil pengujian struktur tes pada PAPs seri E1 dan E2 dalam bentuk CBT dan PPT. Tes PAPs dalam bentuk CBT, baik itu seri E1 dan E2 memiliki nilai CFI dan TLI yang berada di bawah 0,9 yang artinya model struktur tes tersebut kurang sesuai dengan data. Di tambah lagi nilai RMSEA yang berada di atas 0,08 menunjukkan bahwa eror yang terjadi berada di atas batas toleransi walaupun residu yang muncul berada di bawah 0,08 (SRMR < 0,08).

Di sisi lain, Tes PAPs yang dalam bentuk PPT memiliki CFI dan TLI yang baik (> 0,90) yang artinya bahwa struktur tes ini didukung oleh data yang ada. Serta, nilai eror yang terjadi masih dalam taraf yang dapat ditoleransi (RSMEA < 0,08; SRMR < 0,08)

Diskusi

Perbandingan properti psikometri dari sebuah tes berbentuk CBT dan PPT belum banyak dilakukan. Di tambah lagi, semakin maraknya penggunaan tes berbentuk CBT berdampak pada kebutuhan mengenai properti psikometri tes tersebut. Penelitian ini bertujuan untuk membandingkan properti psikometri tes berbentuk CBT dan PPT menggunakan Tes PAPs seri E1 dan E2.

Analisis dilakukan pada parameter taraf kesukaran butir menunjukkan nilai yang bergerak dari negatif yang artinya mudah ke positif yang bermakna semakin sulit. Standar yang dikemukakan oleh Hambleton, Swaminathan, dan Rogers (1991) berdasarkan pendekatan IRT, butir dapat dikatakan sangat mudah apabila memiliki taraf kesukaran di bawah -2. Berikutnya butir soal dikatakan mudah apabila taraf kesukaran bergerak antara -0,2 sampai -0,5. Butir dapat dikatakan memiliki taraf kesukaran medium apabila taraf kesukarannya antara -0,5 sampai 0,5. Butir sulit antara 0,5 sampai 2, sementara butir yang sangat sulit memiliki taraf kesukaran di atas 2. Hasil analisis menunjukkan bahwa baik Tes PAPs seri E1 maupun E2 baik berbentuk CBT maupun PPT memiliki taraf kesukaran butir yang bergerak dari sangat mudah hingga sangat sulit. Sementara itu dalam penyajiannya menggunakan format CBT dan PPT

hasilnya menunjukkan bahwa secara umum tes dalam bentuk PPT memiliki rentang taraf kesukaran lebih luas dibandingkan tes dalam format CBT. Meskipun demikian, hal ini tergolong dapat menggambarkan bahwa kedua bentuk tes ini memiliki tingkat kesukaran butir yang tidak jauh berbeda.

Pada parameter daya diskriminasi, Baker & Kim (2017) menetapkan kriteria yaitu daya diskriminasi sangat rendah bergerak antara 0,01 sampai 0,34, daya diskriminasi rendah antara 0,35 sampai 0,64, Butir dengan daya diskriminasi cukup baik antara 0,65 sampai 1,34, butir dengan diskriminasi baik memiliki skor bergerak antara 1,35 sampai 1,69 dan butir dengan daya diskriminasi di atas 1,7 merupakan butir dengan daya diskriminasi yang sangat baik.

Hasil analisis daya diskriminasi menggunakan Rasch menunjukkan bahwa terdapat beberapa butir dalam Tes PAPs baik seri E1 maupun seri E2 baik dalam CBT dan PBT yang memiliki daya diskriminasi di bawah 0,5. Nilai tersebut bermakna bahwa beberapa butir tersebut tidak dapat membedakan antara kelompok rendah dan tinggi (Azwar, 2016) dan juga hal ini dapat menandakan bahwa soal tersebut tidak memiliki variasi sehingga setiap kelompok dapat memberikan jawaban benar atau salah (Furr & Bacharach, 2013). Namun demikian, secara keseluruhan butir dalam Tes PAPs seri E1 dan E2 dalam bentuk CBT maupun PBT memiliki daya diskriminasi yang memuaskan. Dengan kata lain, kedua bentuk Tes PAPs seri E1 dan E2 memiliki kesetaraan dalam hal daya diskriminasi butir.

Dalam hal ketepatan butir, secara umum seluruh sub tes dalam PAPs baik berbentuk CBT maupun PBT memiliki rentan nilai antara 0,5 hingga 1,5 yang di mana nilai tersebut ditetapkan oleh Linacre (2011) sebagai standar sebuah butir dikatakan sebagai butir yang memiliki ketepatan model yang baik. Artinya, butir tersebut dapat dijawab oleh subjek yang benar-benar memiliki kemampuan untuk mengerjakan soal tersebut dan tidak bisa dijawab oleh subjek yang memiliki kemampuan di bawah taraf kesukaran butir tersebut. Dengan demikian, dapat dikatakan secara keseluruhan butir-butir pada Tes PAPs seri E1 dan E2 yang berbentuk CBT dan PPT memiliki ketepatan model yang baik meskipun masih terdapat beberapa butir yang tergolong tidak fit atau tidak memenuhi standar *item-fit*. Dengan kata lain, kedua bentuk ini memiliki kesetaraan dalam hal nilai ketepatan butir atau *item-fit*.

Hasil uji analisis faktor konfirmatori menunjukkan bahwa Tes PAPs yang berbentuk PPT baik seri E1 dan E2 memiliki ketepatan model yang baik yaitu CFI dan TLI di atas 0,9 dan RMSEA dan SRMR di bawah 0,08 (Furr & Bacharach, 2013; Ghazali, 2017). Namun demikian, tes dalam bentuk CBT memiliki nilai yang sedikit di bawah nilai standar tersebut. Hal ini merupakan bagian dari *method effect* yang di mana hal ini merujuk kepada dampak dari penggunaan metode yang berbeda dalam pengukuran psikologi yang mampu membuat variansi skor yang berbeda di luar dari efek yang diperhitungkan (Maul, 2013). Variansi yang kecil menimbulkan *factor loading* menjadi semakin kecil (Widhiarso & Haryanta, 2015) yang di mana

kecilnya *factor loading* akan membuat SRMR dan RMSEA semakin besar dan CFI serta TLI semakin kecil (Cole, Perkins, & Zelkowitz, 2015).

Hasil investigasi terhadap Tes PAPs berbentuk CBT dengan membandingkan dengan tes yang sama dengan kertas dan pensil (PPT) menunjukkan hasil yang ekuivalen, dan sebagaimana dalam beberapa parameter di atas telah membuktikan bahwa tes berbentuk CBT mempunyai hasil yang setara dengan PPT. Sehingga berdasarkan parameter ini dan mempertimbangkan bahwa peserta telah sepenuhnya memahami tentang pengoperasian komputer dasar dan adanya arahan serta latihan penggunaan fungsi tombol dalam tes yang dipandu oleh tester maka tes berbentuk CBT ini terbukti layak memenuhi standar ekuivalensi sebagai-mana diajukan oleh Bugbee Jr. (1996). Hasil ini sesuai dengan hasil meta-analisis atas tes kemampuan kognitif pada remaja yang menunjukkan ekuivalensi antara ke dua jenis tes yang berbeda administrasi tersebut, meski kajian tentang kecepatan pengerjaan tes perlu dikaji lebih lanjut (Mead & Drasgow, 1993).

Kesimpulan

Temuan dalam penelitian ini mendukung bahwa pengadministrasian Tes PAPs menggunakan CBT dan PPT memiliki kesetaraan dalam beberapa properti psikometrinya yang meliputi taraf kesukaran butir, daya beda butir dan tingkat ketepatan model meskipun secara struktur pengukuran tes berbentuk CBT belum terdukung secara kuat seperti tes berbentuk PPT.

Saran

Untuk penelitian lebih lanjut, analisis lebih komprehensif sebisa mungkin dilakukan untuk mendapatkan parameter kesetaraan antara CBT dan PBT yang lebih ketat, seperti menggunakan *item drift analysis*. Dalam penelitian ini, *item drift analysis* dapat memberikan gambaran secara visual mengenai perbedaan kedua bentuk tes tersebut.

Kepustakaan

Alagumalai, S., & Curtis, D. D. (2005). Classical test theory. In S. Alagumalai, D.D. Curtis, & N. Hungi (Eds.), *Applied Rasch measurement: A book of exemplars: Papers in honour of John P. Keeves* (hal.1-14). Dordrecht, The Netherlands: Springer.

Azwar, S. (2016). *Konstruksi tes kemampuan kognitif* (Edisi pertama). Yogyakarta: Pustaka Pelajar.

Aşkar, P., Altun, A., Cangöz, B., Çevik, V., Kaya, G., & Türksoy, H. (2012). A comparison of paper-and-pencil and computerized forms of line orientation and enhanced cued recall tests. *Psychological Reports, 110*(2), 383–396. doi: 10.2466/03.22.PR0.110.2.383-396

Bacharach, V. R., & Furr, R. M. (2007). *Psychometrics: An introduction*. Thousand Oaks, CA: SAGE Publications.

Baker, F. B., & Kim, S.-H. (2017). *The basics of item response theory using R*. Springer

International Publishing.

Bond, T. G., & Fox, C. M. (2015). *Applying the Rasch model fundamental measurement in the human sciences* (Edisi ketiga). New York: Routledge.

Chen, S.-Y., & Lei, P.-W. (2005). Controlling item exposure and test overlap in computerized adaptive testing. *Applied Psychological Measurement, 29*(3), 204–217. doi: 10.1177/0146621604271495

Bugbee Jr., A. C. (1996). The equivalence of paper-and-pencil and computer-based testing. *Journal of Research on Computing in Education, 28*(3), 282-299. doi: 10.1080/08886504.1996.10782166

Choi, I.-C., Kim, K. S., & Boo, J. (2003). Comparability of a paper-based language test and a computer-based language test. *Language Testing, 20*(3), 295–320. doi: 10.1191/0265532203lt258oa

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory*. New York: Holt, Rinehart, and Winston.

Cole, D. A., Perkins, C. E., & Zelkowitz, R. L. (2015). Impact of homogeneous and heterogeneous parceling strategies when latent variables represent multidimensional constructs. *Psychological Methods, 21*(2), 164–174. doi: 10.1037/met0000047

Flowers, C., Kim, D.-H., Lewis, P., & Davis, V. C. (2011). A comparison of computer-based testing and pencil-and-paper testing for students with a read-aloud accommodation. *Journal of Special Education Technology, 26*(1), 1–12. doi: 10.1177/016264341102600102

Furr, M. R., & Bacharach, V. R. (2013). *Psychometric: An introduction* (Edisi kedua). Thousand Oaks: SAGE Publisher.

Ghazali, I. (2017). *Model persamaan struktural dengan AMOS 24: Update Bayesian SEM* (Edisi ketujuh). Semarang: Badan Penerbit Universitas Diponegoro.

Ginty, A. T. (2013). Psychometric Properties. In M. Gellman & R. J. Turner, *Encyclopedia of Behavioral Medicine*. New York, NY: Springer.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of Item response theory*. Thousand Oaks: SAGE.

Hardcastle, J., Herrmann-Abell, C. F., & DeBoer, G. E. (2017). Validating an assessment for tracking students' growth in understanding of energy from elementary school to high school Joseph. *NARST Annual International Conference* (hal. 1–10). San Antonio, TX, April.

Hosseini, M., Abidin, M. J. Z., & Baghdarnia, M. (2014). Comparability of test results of computer based tests (CBT) and paper and pencil tests (PPT) among English language learners in Iran. *Procedia - Social and Behavioral Sciences, 98*, 659–667. doi: 10.1016/j.sbspro.2014.03.465

Janssen, G., Meier, V., & Trace, J. (2014). Classical test theory and item response theory. Two understandings of one high-stakes performance exam. *Colombian Applied Linguistics Journal, 16*(2), 167–184. doi: 10.14483/udistrital.jour.calj.2014.2.a03

Keller, L. A, Swaminathan, H., & Sireci, S. G. (2003). Education evaluating scoring procedures for context-

dependent item sets 1. *Applied Measurement in Education, 16*(3), 207–222. doi: 10.1207/S15324818AME1603

Leong, F. T. (2008). *Encyclopedia of counseling*. Thousand Oaks, CA: Sage Inc.

Linacre, J. M. (2011). *A User's guide to winstep minsitep Rasch-model computer program*. Beaverton, Oregon: Winsteps.com.

Luecht, R. M. (2005). Some useful cost-benefit criteria for evaluating computer-based test delivery models and systems. *Journal of Applied Testing Technology, 7*(2), 1-35.

Maul, A. (2013). Method effects and the meaning of measurement. *Frontiers in Psychology, 4*, 1-13. doi: 10.3389/fpsyg.2013.00169

Mead, A. D., & Drasgow, F. (1993). Equivalence of computerized and paper-and-pencil cognitive ability tests: A meta-analysis. *Psychological Bulletin, 114*(3), 449–458. doi: 10.1037/0033-2909.114.3.449

Partchev, I. (2009). *irtoys: A collection of functions related to item Response Theory (IRT). R package version 0.2.1*. Diakses melalui https://CRAN.R-projects.org/package=irtoys

Parshall, C. G., Harmes, J. C., Davey, T., & Pashley, P. (2009). Innovative items for computerized testing. In W.J. van der Linden & C.A.W. Glas (Eds.) *Elements of adaptive testing. statistical for social and behavior sciences.* New York, NY: Springer

Piaw, C. Y. (2012). Replacing paper-based testing with computer-based testing in assessment: Are we doing wrong? *Procedia - Social and Behavioral Sciences, 64*, 655–664. doi: 10.1016/j.sbspro.2012.11.077

Pommerich, M. (2004). Developing computerized versions of paper-and-pencil tests: Mode effects for passage-based tests. *The Journal of Technology Learning a-nd Assessment, 2*(6), 1–44.

Rizopoulos, D. (2007). ltm: An R Package for latent variable modeling and item response analysis. *Journal of Statistical Software, 17*(5), 1-25. doi: 10.18637/jss.v017.i05

Sumintono, B., & Widhiarso, W. (2013). *Aplikasi Model Rasch untuk penelitian ilmu-ilmu sosial*. Cimahi: Trim Komunikata Publishing House.

Tate, R. (2003). A comparison of selected empirical methods for assessing the structure of responses to test items. *Applied Psychological Measurement, 27*(3), 159–203. doi: 10.1177/0146621603252327

The National Assessment of Educational Progress. (2014). *Main NAEP assessments.* Diakses pada 15 Oktober 2019 melalui https://nces.ed.gov/nationsreportcard/assessments/

Vrabel, M. (2004). Computerized versus paper-and-pencil testing methods for a nursing certification examination: A Review of the literature. *CIN: Computers, Informatics, Nursing, 22*(2), 94-98. doi: 10.1097/00024665-200403000-00010

Widhiarso, W., & Haryanta. (2015). Examining method effect of synonym and antonym test in verbal abilities measure. *Europe's Journal of Psychology, 11*(3), 419–431. doi:

10.5964/ejop.v11i3.865

Williamson, D. M., Mislevy, R. J., & Bejar, I. I. (2006). *Automated scoring of complex tasks in computer-based testing*. Mahwah, NJ: Psychology Press.

Wright, B. D., & Mok, M. M. (2004). An overview of the family of Rasch measurement models. In E. V. Smith Jr. & R. M. Smith (Eds), *Introduction to Rasch measurement: Theory, models, and applications* (hal. 1-24). Minnesota: Jam Press.