JP3I (Jurnal Pengukuran Psikologi dan Pendidikan Indonesia), 10(1), 2021, 1-11
DOI: http://dx.doi.org/10.15408/jp3i.v10i1.20551
http://journal.uinjkt.ac.id/index.php/jp3i

Klasika: Program Analisis *Item* dan Tes dengan Pendekatan Klasik

Bahrul Hayat

Fakultas Psikologi, UIN Syarif Hidayatullah Jakarta, Indonesia

bahrulhayat@uinjkt.ac.id

Abstract

This article introduces the Klasika software developed to run item and test analysis using the Classical Test Theory approach. Classical Test Theory is one of the specialized competencies and skills that undergraduate students of psychology must possess. Classical Test Theory becomes a mandatory course for all schools or departments of psychology in Indonesia. This article also provides a theoretical foundation of Classical Test Theory's essential concepts and statistical methods, specifically related to items and test statistics. The item analysis and test reliability procedures using Klasika, starting from the data preparation untill data interpretation, are explained with an empirical illustration. Finally, the analysis results using Klasika are compared with the results from Quest software to test the accuracy of the estimation results.

Keywords*: classical test theory, item analysis, test reliability, Klasika, scoring, psychometrics software.*

Abstrak

Artikel ini berisi pengenalan terhadap perangkat lunak Klasika yang dikembangkan untuk proses analisis *item* dan tes dengan pendekatan teori tes klasik. Teori tes klasik merupakan salah satu kompetensi dan keterampilan khusus yang wajib dimiliki sarjana psikologi berdasarkan Kurikulum Inti Psikologi Indonesia. Artikel ini juga memuat landasan teoretis tentang konsep dan statistik penting dalam teori tes klasik, khususnya terkait *item* dan tes. Prosedur aplikasi analisis *item* dan reliabilitas tes dengan program Klasika dimulai dari proses penyiapan data hingga contoh penafsiran hasil analisis yang dilengkapi dengan ilustrasi gambar untuk masing-masing tahapan analisis juga dibahas. Selamjutnya, hasil analisis Klasika dibandingkan dengan hasil analisis dari program Quest untuk mengecek akurasi hasil estimasi.

Kata Kunci: teori tes klasik, analisis *item*, tes reliabilitas, Klasika, skoring, *software* psikometri.

Pendahuluan

Suatu tes yang digunakan dalam bidang psikologi dan pendidikan pada dasarnya merupakan alat untuk memperoleh sampel dari perilaku. Biasanya perilaku dimaksud dikuantifikasi sedemikian rupa untuk menghasilkan suatu skor (Lord, 1980). Secara umum, dalam pengukuran psikologi dan pendidikan, terdapat dua klasifikasi umum jenis tes, yaitu: *maximum performance test* yang mengukur kemampuan peserta tes (*ability*) dan melibatkan penyekoran benar dan salah dan *typical performance test* yang mengukur aspek non-kemampuan (*non-ability*) dan tidak melibatkan penyekoran benar dan salah (Cronbach, 1960). Sebelum digunakan untuk pengambilan keputusan tentang peserta tes, skor yang diperoleh melalui dua jenis tes tersebut harus terlebih dahulu dilakukan pengujian secara empiris tentang kualitas dan kehandalan tes tersebut. Pengujian dimaksud merupakan analisis psikometrik, seperti analisis *item*, validitas tes, dan reliabilitas tes. Pengujian dan evaluasi karakteristik psikometrik terhadap suatu tes atau instrumen psikologi telah diwajibkan oleh *American Psychological Association* (Wilkinson & Task Force on Statistical Inference, 1999). Oleh karena itu, pemahaman dan penguasaan secara praktis tentang ilmu psikometrika sangat penting bagi para peneliti, pengembang, dan pengguna tes untuk dapat melakukan pengujian aspek psikometrik dari suatu tes agar dapat dipeorleh tes yang baik dan baku.

Dalam ilmu psikometrika, secara umum terdapat tiga aliran teori tes yang telah dikembangkan, yaitu: (1) Teori tes klasik; (2) *Item response theory* (IRT) yang seringkali dikenal sebagai teori tes modern; dan (3) Rasch measurement theory yang dikenal sebagai "*the simplest modern test theory*" (Andrich, 2004; Andrich & Marais, 2019). Traub (1997) menjelaskan bahwa teori tes klasik atau yang dalam Bahasa Inggris disebut *classical test theory* (CTT) muncul pada awal abad ke-20. CTT lahir dari kristalisasi materi yang mencakup tiga pencapaian luar biasa, yaitu: (1) kesadaran atas adanya kesalahan dalam pengukuran, (2) konsep kesalahan pengukuran tersebut merupakan variabel random, dan (3) konsep tentang korelasi dan bagaimana mengembangkan indeksnya. Kemudian pada tahun 1904 Charles Spearman menunjukkan kepada kita bagaimana cara untuk mengkoreksi koefisien korelasi saat terjadi atenuasi (hasil korelasi lebih rendah dari yang semestinya) akibat kesalahan pengukuran dan bagaimana cara untuk mendapatkan indeks reliabilitas dari suatu tes agar dapat dilakukan koreksi. Karya dari Spearman tersebut menandai awal dari CTT. Selanjutnya, kerangka teoretis tentang CTT diuraikan dan disempurnakan oleh Spearman, George Udny Yule, Truman Lee Kelley, serta tokoh lainnya sekitar seperempat abad kemudian setelah tahun 1904. Tonggak sejarah lainnya diletakkan pada tahun 1937 dengan munculnya formula Kuder-Richardson yang tak lama kemudian diikuti oleh ide dasar tentang batas bawah (*lower bound*) dari koefisien reliabilitas. Puncak dari perkembangan CTT diwujudkan dalam penjelasan sistematis yang dilakukan oleh Melvin Novick (1966).

Beberapa keterbatasan dari CTT telah mendorong lahir dan berkembangnya teori tes modern (lihat, Wright, 1997), meskipun CTT tidak sepenuhnya ditinggalkan dan masih terus digunakan sampai saat ini (misal, Martins et al., 2020; Raykov & Marcoulides, 2016). CTT telah menjadi materi yang wajib dikuasai oleh sarjana psikologi di Indonesia berdasarkan Kurikulum Inti Psikologi Indonesia (AP2TPI, 2018). Mata kuliah yang membahas materi CTT ini ditawarkan pada jenjang sarjana di semua jurusan atau fakultas psikologi. Meskipun tidak selalu ditawarkan dengan nama mata kuliah "Teori Tes Klasik", materi tentang hal ini seringkali dimasukkan ke dalam mata kuliah lain seperti, pengembangan instrumen psikologi, psikometrika, ataupun mata kuliah lainnya yang sejenis. Melalui mata kuliah ini, mahasiswa psikologi dituntut tidak hanya untuk mengetahui dan memahami konsep dasar CTT, namun juga dapat menerapkan konsep dasar CTT tersebut dalam melakukan analisis *item* dan tes. Kini telah tersedia berbagai program komputer yang dapat digunakan oleh dosen dan mahasiswa untuk melakukan analisis *item* dan tes (misal, Iteman dan Quest). Program tersebut hampir semua merupakan program berbahasa Inggris yang membutuhkan syntax dan bukan berbasis GUI (Graphical User Interface).

Berbeda dengan program tersebut, Klasika adalah program berbasis GUI yang dikembangkan oleh penulis untuk memenuhi kebutuhan pengguna CTT di Indonesia. Semula program ini dikembangkan untuk mengatasi keterbatasan pada penggunaan perangkat lunak Iteman yang terkadang bermasalah karena tidak kompatibel dengan perkembangan sistem operasi komputer yang semakin canggih dari

http://journal.uinjkt.ac.id/index.php/jp3i

waktu ke waktu. Program ini juga dapat digunakan untuk mengatasi kesulitan pelaporan hasil analisis dari perangkat lunak lain (seperti Iteman atau Quest) yang membutuhkan pemindahan output hasil analisis secara manual ke dalam dokumen laporan atau manuskrip (lihat, Adams & Khoo, 1993; Berk & Griesemer, 1976). Klasika dapat menghasilkan laporan hasil analisis secara otomatis dalam format Microsoft Excel yang mudah dipindahkan ke berbagai program aplikasi.

Tujuan utama program Klasika adalah untuk melakukan analisis *item* dan tes. Analisis *item* merupakan prosedur analisis psikometrik untuk menguji kualitas *item* sehingga dapat dipilih *item-item* berkualitas baik yang dapat digunakan untuk merakit suatu tes yang baku. Analisis *item* juga dapat memberikan informasi mengenai *item* mana yang kurang baik dan perlu diperbaiki (Bazaldua et al., 2017). Di samping analisis *item*, Klasika juga dapat menguji reliabilitas suatu alat tes yang memberi informasi tentang kehandalan suatu tes dan konsistensi skor yang diperoleh dari tes tersebut.

Analisis yang dapat dilakukan program Klasika mencakup empat tema yang menjadi bahasan penting dalam CTT yaitu: (1) indeks kesukaran *item*; (2) indeks daya pembeda (diskriminasi) *item*; (3) informasi tentang konsistensi internal dalam bentuk koefisien reliabilitas; dan (4) estimasi kesalahan pengukuran. Artikel ini membahas landasan teoretis tentang statistik dimaksud serta ilustrasi empiris analisis *item* dan tes dengan program Klasika. Hasil analisis program Klasika kemudian dibandingkan dengan hasil analisis dari program Quest untuk melihat apakah terdapat perbedaan hasil komputasi dari Klasika dan Quest utuk data yang sama.

Beberapa Statistik Penting dalam CTT

Pembahasan tentang berbagai statistik penting dalam pendekatan CTT kini dapat dengan mudah dijumpai dalam literatur. Statistik penting yang dilaporkan dalam Klasika mencakup analisis psikometrik klasik dan tidak jauh berbeda dengan hasil yang dilaporkan dari program analisis yang lain seperti Quest dan Iteman. Statistik dimaksud meliputi, antara lain, tingkat kesukaran *item*, daya pembeda *item*, koefisien reliabilitas, dan kesalahan pengukuran. Adapun penjelasan masing-masing konsep tersebut dapat dilihat pada masing-masing subbab berikut ini.

Indeks Tingkat Kesukaran *Item*

Dalam mengembangkan suatu tes yang baku, pengembang tes perlu meyakinkan bahwa *item* yang digunakan dalam tes tersebut memiliki karakteristik psikometrik yang baik. Oleh karena itu, analisis setiap *item* perlu dilakukan untuk menguji apakah *item-item* tersebut telah memenuhi persyaratan psikometrik. Parameter psikometrik yang umum digunakan untuk menguji kualitas *item* adalah tingkat kesukaran soal dan daya pembeda soal.

Untuk butir soal yang bersifat dikotomi,yaitu diberi skor 0 jika salah dan 1 jika benar, tingkat kesukaran soal didefinisikan sebagai proporsi peserta tes yang menjawab benar suatu *item*. Proporsi untuk *item* ke-i biasanya dilambangkan dengan p_i. Nilai p_i berkisar dari 0,00 hingga 1,00. Proporsi untuk suatu *item* merupakan rata-rata skor untuk *item* tersebut. Rata-rata skor dari suatu tes sama dengan jumlah tingkat kesukaran *item* dari tes tersebut (Crocker & Algina, 2008), dan dapat dihitung dengan rumus berikut:

$$\mu_X = \sum_i p_i$$

Dimana μ_X merupakan rata-rata skor tes, p_i adalah tingkat kesukaran *item*. Lebih lanjut rata-rata tingkat kesukaran *item* dapat diperoleh dengan membagi rata-rata skor tes dengan banyaknya *item* dalam suatu tes, seperti dapat dilihat dalam rumus:

$$\mu_p = \frac{(\mu_X)}{k}$$

Dimana μ_p = rata-rata tingkat kesukaran *item*, μ_X = rata-rata skor tes, k = banyaknya *item* pada suatu tes.

Tingkat kesukaran *item*, *pi*, merupakan informasi penting dalam analisis *item*. Jika p_i mendekati 0 atau 1, *item* tersebut harus dipertimbangkan kembali untuk dimasukan ke dalam tes, karena *item* tersebut tidak memberikan informasi mengenai perbedaan tingkat kemampuan individu peserta tes. Jika $p_i = 0$, berarti tidak ada satu orang pun yang menjawab benar *item* tersebut dan item tersebut terlalu sulit serta tidak berguna. Sebaliknya, jika $p_i = 1$, berarti seluruh peserta tes menjawab benar *item* tersebut dan tidak ada informasi mengenai perbedaan individu yang didapat dari *item* tersebut.

Tingkat kesukaran *item* menentukan varians dari suatu *item* karena $\sigma_i^2 = p_i q_i$ di mana q_i adalah (1 - p_i). Varian skor total dari suatu tes akan maksimal apabila p_i = 0,50 (Crocker & Algina, 2008). Hal ini berarti tes akan dengan optimal membedakan kemampuan peserta tes jika *item-item* dalam tes memiliki tingkat kesukaran 0.50. Namun demikian, nilai optimal sangat tergantung pada bentuk soal yang digunakan. Sebagai contoh, untuk bentuk soal pilihan ganda dengan 4 (empat) pilihan jawaban, tingkat kesukaran akan optimal pada p_i = 0,625 setelah diperhitungkan peluang tebakan. Sebagai pedoman umum, sebagaimana dinyatakan oleh Allen & Yen (1979) rentang tingkat kesukaran antara 0,30 sampai 0,70 memberikan informasi perbedaan antar individu peserta tes dengan baik.

Indeks Daya Pembeda (Diskriminasi) *Item*

Indeks daya pembeda *item* merupakan indeks tentang kemampuan suatu *item* dalam membedakan kelompok peserta tes yang memiliki kemampuan (skor) tinggi dalam tes dengan mereka yang memiliki kemampuan (skor) rendah. Dengan kata lain, daya pembeda soal memberi informasi tentang seberapa efektif sebuah *item* membedakan peserta tes yang memiliki kemampuan baik dan kurang baik.

Terdapat beberapa teknik untuk menghitung indeks daya pembeda soal. Teknik yang paling sederhana adalah dihitung dengan rumus:

$$D = p_u - p_i$$

Di mana p_u adalah proporsi peserta tes dari kelompok atas yang menjawab benar dan p_i adalah proporsi peserta tes dari kelompok bawah yang menjawab benar. Kriteria yang banyak digunakan untuk menentukan kelompok atas dan kelompok bawah adalah 27% dari peserta tes terbaik (teratas) dan 27% dari peserta terjelek (terbawah).

Nilai D berkisar dari -1.00 sampai 1.00. Nilai yang positif menunjukkan bahwa *item* berfungsi seperti yang diharapkan di mana kelompok atas lebih banyak yang menjawab benar *item* dibanding kelompok bawah. Nilai yang negatif menunjukkan bahwa *item* berfungsi sebaliknya di mana kelompok bawah lebih banyak yang menjawab benar *item* dibanding kelompok atas. Berdasarkan pengalaman praktis, Ebel & Frisbie (1991) membuat panduan interpretasi nilai D dengan kriteria: Jika D ≥ 0,40, maka *item* berfungsi dengan sangat baik. Jika 0,30 ≤ D ≤ 0,39, *item* membutuhkan sedikit ataupun tanpa revisi. Jika 0,20 ≤ D ≤ 0,29, *item* berada pada ambang batas dan membutuhkan revisi. Jika D ≤ 0,19, maka *item* harus dieliminasi atau direvisi secara keseluruhan.

Teknik lain untuk menghitung indeks daya pembeda soal untuk soal yang bersifat dikotomi adalah dengan menghitung korelasi antara skor suatu *item* dengan skor total tes. Dengan kata lain, indeks daya pembeda dihitung dengn melihat seberapa dekat *performance* terhadap suatu *item* berhubungan dengan *performance* dari skor total tes (Crocker & Algina, 2008). Rumus korelasi untuk menghitung indeks daya pembeda untuk *item* dikotomi adalah dengan *item/total-test-score-point-biserial correlation*, r_{pbis}, sebagai berikut:

$$r_{pbis} = \frac{(\mu_+ - \mu_X)}{\sigma_X}\sqrt{p/q}$$

http://journal.uinjkt.ac.id/index.php/jp3i

Di mana μ_X adalah rata-rata skor dari orang yang menjawab benar *item*, μ_X adalah rata-rata skor dari seluruh peserta tes dan σ_X adalah standar deviasinya, p adalah tingkat kesukaran soal, dan q adalah $(1 - p)$.

Crocker & Algina (2008) menyampaikan bahwa korelasi lainnya yang mirip dengan point biserial adalah korelasi biserial. Letak perbedaan antara keduanya adalah dalam proses estimasi terhadap koefisien korelasi tersebut, dimana dalam proses komputasi korelasi biserial, variabel yang berbentuk dikotomi dirubah menjadi variabel yang diasumsikan berdistribusi normal.

Besaran nilai korelasi point biserial antara skor *item* dan skor total terkadang merupakan nilai yang perlu ditafsirkan secara hati-hati karena skor *item* berkontribusi pada skor total peserta tes (Crocker & Algina, 2008). Jika jumlah *item* besar (mungkin 25 atau lebih), maka hal tersebut bukan merupakan masalah. Namun, dengan jumlah *item* yang sedikit, permasalahan ini perlu dikoreksi dengan rumus:

$$\rho_{i(X-i)} = \frac{\rho_{Xi}\sigma_X - \sigma_i}{\sqrt{\sigma_i^2 + \sigma_X^2 - 2\rho_{Xi}\sigma_X\sigma_i}}$$

Dimana $\rho_{i(X-i)}$ adalah korelasi antara skor *item* dan skor total dengan *item* mengeluarkan *item* tersebut dari perhitungan, dan $\sigma_X\sigma_i$ adalah standar deviasi total dan standar deviasi dari *item*. Dengan koreksi tersebut, hasil komputasi korelasi point-biserial akan tetap akurat meskipun jumlah *item* berkurang. Penggunaan prosedur koreksi seperti ini telah dilakukan secara otomatis dalam perangkat lunak standar yang digunakan untuk analisis *item* dengan pendekatan CTT (Adams & Khoo, 1993).

Koefisien Reliabilitas

Konsep reliabilitas tes merupakan konsep sentral dalam CTT. Sebuah tes dinyatakan telah memenuhi persyaratan sebagai alat ukur yang baik dan mencapai tingkat kepercayaan yang memadai, apabila telah memiliki nilai koefisien reliabilitas yang baik (Adams, 2005). Gagasan tentang reliabilitas pertama kali diperkenalkan oleh Spearman tahun 1904 ketika dia menemukan bahwa terdapat kesalahan pengukuran (*measurement error*) yang berdampak pada melemahnya korelasi antar skor yang teramati (*observed score*) dari dua variabel. Estimasi korelasi antar variabel menjadi lebih rendah dari yang seharusnya akibat adanya kesalahan pengukuran dan perlu dilakukan koreksi agar diperoleh estimasi yang lebih akurat dengan memperhitungkan kesalahan pengukuran. Pada tahun 1910 Spearman memperkenalkan istilah koefisien reliabilitas dan mendefinisikannya sebagai korelasi antara setengah bagian dengan setengah bagian lainnya dari suatu tes yang mengukur hal yang sama.

Secara sederhana reliabilitas dimaknai sebagai seberapa konsisten skor seseorang apabila dilakukan pengukuran berulang dengan tes yang sama atau yang dianggap paralel. Dalam konteks ini, reliabilitas dapat didefinisikan secara statistik sebagai korelasi antara skor yang diperoleh seseorang (*observed score*) dari pengukuran berulang dari suatu tes yang sama atau tes yang paralel. Dengan kata lain, jika peserta tes memperoleh skor yang sama dari dua pelaksanaan tes yang sama atau dua tes yang paralel maka tes tersebut memiliki reliabilitas sempurna ($\rho_{XX'} = 1$). Sebaliknya, jika peserta tes memperoleh skor dari suatu tes yang tidak berhubungan sama sekali dengan skor yang diperoleh dari tes lain yang diasumsikan paralel ($\rho_{XX'} = 0$) , maka kedua tes dimaksud sama sekali tidak reliabel (Allen & Yen, 1979). Oleh karena itu, secara teoritik, koefisien reliabilitas berkisar antara 0 sampai 1, namun secara empirik koefisien reliabilitas tidak pernah mencapai 1. Artinya secara praktis tidak pernah didapat tes yang memiliki reliabilitas sempurna. Ketidakkonsistenan skor antara dua tes yang paralel disebabkan oleh kesalahan pengukuran yang mempengaruhi performa peserta tes dalam menempuh tes.

Terdapat beberapa metoda dan prosedur untuk mengestimasi reliabilitas suatu tes. Perbedaan metoda yang digunakan untuk mengestimasi reliabilitas berimplikasi pada interpretasi dan makna reliabiltas yang sedikit banyak berbeda. Koefisien reliabiltas yang diestimasi dengan memberikan suatu tes secara berulang kepada sekelompok peserta tes (*test-retest*) dimaknai sebagai koefisien stabilitas tes. Sedangkan

koefisien reliabilitas yang diperoleh dari korelasi antar subtes atau paket tes bermakna koefisien ekuivalensi (kesetaraan) tes. Sebaliknya, koefisien reliabilitas yang diperoleh dari pemberian satu paket tes kepada sekelompok orang lebih tepat dimaknai sebagai koefisien konsistensi internal tes (Crocker & Algina, 2008).

Salah satu prosedur estimasi koefisien reliabilitas konsistensi internal yang paling banyak digunakan adalah Crobach Alpha (α) yang dikembangkan oleh Lee J. Cronbach tahun 1951. Koefisien alpha (α) dapat digunakan untuk mengestimasi reliabilitas dari suatu tes dengan hanya satu kali pelaksanaan tes kepada sekelompok peserta tes (Allen & Yen, 1979). Koefisien alpha (α) juga digunakan untuk mengatasi kondisi pemilahan suatu tes ke dalam dua bagian yang tidak paralel (Allen & Yen, 1979). Rumus dari koefisien alpha (α) adalah (Cronbach, 1951):

$$\alpha = \left(\frac{n}{n-1}\right)\left(1 - \frac{\sum \sigma_i^2}{\sigma_X^2}\right)$$

Dimana n adalah jumlah *item* pada tes, σ_X^2 adalah varians skor tes, σ_i^2 adalah varians *item* ke-i, dan $\sum \sigma_i^2$ merupakan penjumlahan varians keseluruhan *item*. Untuk data dikotomi, rumus koefisien alpha (α) identik dengan rumus Kuder Richardson 20 (KR 20) di mana $\sigma_i^2 = p_i q_i$. KR-20 adalah kasus khusus dari koefisien alpha (α) yang dapat dihitung dengan rumus yang sama (Loewenthal, 2001).

Dengan mengetahui koefisien reliabilitas, pengembang dan pengguna tes dapat melakukan estimasi seberapa besar kesalahan pengukuran dari suatu tes. Hal ini akan sangat membantu dalam interpretasi skor yang diperoleh dari suatu tes. Walaupun kita tidak dapat menentukan dengan pasti besaran kesalahan pengukuran, CTT memberikan metoda untuk mengestimasi kesalahan pengukuran dengan *standard error of measurement* (SEM). SEM yang disimbulkan dengan σ_E merupakan estimasi variasi skor yang diperoleh (*observed score*) terhadap skor sesungguhnya (*true score*) jika dilakukan tes berulangkali. SEM dihitung dengan rumus:

$$\sigma_E = \sigma_X \sqrt{(1 - \rho_{XX'})}$$

Di mana σ_X adalah standar deviasi skor tes dan $\rho_{XX'}$ adalah reliabilitas tes.

Terlihat sangat jelas bahwa kesalahan pengukuran yang diindikasikan dengan SEM sangat berhubungan dengan reliabilitas suatu tes. Semakin tinggi koefisien reliabilitas suatu tes, maka kesalahan pengukuran akan semakin kecil. SEM dapat digunakan untuk membuat interval keyakinan (*confidence interval*) di sekitar skor yang diperoleh seseorang dari suatu tes (*observed score*). Dengan mengacu pada tabel distribusi normal kita dapat memiliki keyakinan sebesar 68% bahwa skor sesungguhnya (*true score*) berada di antara plus-minus satu SEM ($X \pm 1$ SEM) dari skor yang diperoleh (*observed score*). Demikian pula, kita meyakini bahwa skor sesungguhnya (*true score*) terletak plus-minus dua SEM ($X \pm 2$ SEM) dari skor yang diperoleh (*observed score*) dengan tingkat keyakinan 95%.

Ilustrasi Empiris: Klasika vs Quest

Ilustrasi empiris dilakukan untuk membandingkan hasil komputasi berbagai statistik CTT yang dihasilkan melalui Klasika dan Quest. Adapun tampilan awal GUI (*Graphical User Interface*) program Klasika dapat dilihat dalam Appendix. Lebih lanjut, ilustrasi empiris dilakukan terhadap data pola respon jawaban soal pilihan ganda 4-opsi jawaban (A, B, C, dan D) dengan skor *item* bersifat dikotomi (1= jawaban benar; 0= jawaban salah) untuk 35 *item* dan 279 responden. Adapun perbandingan yang dilakukan meliputi statistik untuk tingkat kesukaran *item*, daya pembeda *item*, koefisien reliabilitas, dan standard error of measurement. Korelasi antara hasil komputasi analisis *item* dengan Klasika dengan hasil komputasi Quest dapat dilihat pada Tabel 1.

http://journal.uinjkt.ac.id/index.php/jp3i

Tabel 1. Korelasi antara Klasika dan Quest

Statistik	Tingkat kesukaran Quest	Point biserial Quest	SEm Quest	Reliabilitas Quest
Tingkat kesukaran Klasika	1,00			
Korelasi Point biserial Klasika		1,00		
SEm Klasika			1,00	
Reliabilitas Klasika				1,00[a]

Catatan: [a]makna 1,00 menunjukkan nilai reliabilitas yang sama antar kedua *software*

Berdasarkan Tabel 1, diketahui bahwa seluruh korelasi hasil analisis Klasika dengan hasil analisis Quest bernilai sempurna (1.00). Hal ini dapat ditafsirkan bahwa Klasika dan Quest menghasilkan estimasi yang sama, baik untuk statistik tingkat kesukaran *item*, daya pembeda *item* dalam bentuk point biserial, dan statistik biserial yang dalam Quest tidak dihasilkan namun dapat dikomputasi melalui korelasi point-biserial. Selain itu, *standard error of measurement* dan koefisien reliabilitas yang dihasilkan oleh kedua *software* juga sama. Untuk data dikotomi, koefisien reliabilitas yang dihasilkan adalah Kuder-Richardson 20 (KR-20). Sedangkan untuk data politomi, koefisien reliabilitas yang dihasilkan adalah Cronbach's alpha (α).

Gambar 1 menyajikan statistik *item* dari hasil analisis Klasika. Sebagai contoh *Item* 2, memiliki tingkat kesulitan *item* sebesar 0.814, yang artinya bahwa sebanyak 81,4% dari 279 orang yang menempuh tes dapat menjawab *Item* 2 dengan benar. Output Quest juga menghasilkan angka yang sama dalam "*Percent* (%)" sebesar 81,4. Daya pembeda dalam bentuk korelasi point-biserial yang dihasilkan Klasika sebesar 0,455 sedikit berbeda dengan Quest sebesar 0,46. Hal ini disebabkan karena Quest hanya melaporkan angka dalam dua desimal sehingga angka 0,455 dibulatkan menjadi 0,46.

No. Soal	Scale Item	Statistik Butir			Rincian Statistik per Opsi				
		Tingkat Kesulitan	Biserial	Point Biserial	Opsi	N (Proporsi)	Biserial	Point Biserial	Kunci
1	0-1	0.935	0.694	0.356	A	0.039	-0.595	-0.260	
					B	0.011	-0.398	-0.109	
					C	0.935	0.694	0.356	*
					D	0.014	-0.700	-0.214	
					Lain	0.000	-9.000	-9.000	
2	0-2	0.814	0.660	0.455	A	0.115	-0.376	-0.229	
					B	0.814	0.660	0.455	*
					C	0.047	-0.633	-0.293	
					D	0.025	-0.726	-0.272	
					Lain	0.000	-9.000	-9.000	

```
Item    2: item 2                                  Infit MNSQ = 0.92
                                                     Disc = 0.46

Categories          A [0]      B [1]      C [0]      D [0]     missing

Count                  32        227         13          7           0
Percent (%)          11.5       81.4        4.7        2.5
Pt-Biserial         -0.23       0.46      -0.29      -0.27
Mean Ability         0.36       1.27      -0.33      -0.62          NA
StDev Ability        0.88       1.00       1.03       0.51          NA

Step Labels                       1

Thresholds                    -0.73
Error                          0.17
```

Gambar 1. Perbandingan Statistik Butir Klasika vs Quest

Gambar 1 juga memuat informasi mengenai distribusi pilihan jawaban peserta tes terhadap empat pilihan jawaban yaitu A, B, C, D, serta lainnya untuk "*nonresponse*" atau jawabannya kosong yang dalam Quest masuk disebut "missing". Pada kolom N (Proporsi) berisi informasi tentang banyaknya peserta tes yang memilih setiap pilihan jawaban. Sebagai contoh pada *Item* 2 terdapat sebanyak 11,5% orang yang memilih opsi A. Kolom kunci yang berisi tanda bintang (asterisk) menunjukkan bahwa kunci jawaban dari *Item* 2 adalah opsi B. Kolom N (proporsi) pada opsi B sebagai kunci menunjukkan angka 0,814 dan sama dengan tingkat kesulitan *item*. Dalam Quest, kunci jawaban ditandai dengan [1]. Gambar 1 secara umum menunjukkan bahwa Klasika dan Quest menghasilkan "solusi" yang sama dalam menyediakan informasi statistik *item* berdasarkan pendekatan CTT. Hal yang berbeda antara Klasika dan Quest adalah dalam bentuk pelaporan output seperti terlihat pada Gambar 1.

Selain itu, terdapat informasi lain yang dapat digali pada Gambar 1. Informasi dimaksud berkaitan dengan kolom proporsi, korelasi biserial, dan point-biserial untuk masing-masing opsi. Sebagai contoh, untuk *Item* 1 dengan kunci jawaban C, terdapat 93,5% responden yang memilih opsi tersebut, sedangkan opsi A dipilih oleh 3,9%, B dipilih oleh 1,1% dan D dipilih oleh 1,4% responden. Berdasarkan data tersebut, dapat disimpulkan bahwa seluruh opsi berfungsi cukup baik; dalam arti tidak ada satupun opsi yang tidak dipilih oleh peserta tes (proporsi = 0).

Lebih lanjut, secara hipotetik, diteorikan bahwa kunci jawaban harus memiliki korelasi biserial ataupun point-biserial yang positif. Setidaknya ada dua makna mengenai arah dan besaran korelasi tersebut. Makna pertama adalah apabila korelasinya positif berarti peserta tes yang memperoleh skor total tinggi cenderung memilih opsi kunci jawaban (jawaban benar). Makna kedua dengan menghitung skor rata-rata peserta tes yang memilih setiap opsi dihipotesisakan bahwa semakin tinggi rata-rata skor total untuk suatu opsi menunjukkan bahwa opsi tersebut dipilih oleh responden yang memiliki kemampuan tinggi. Sebagai contoh, dari hasil perhitungan diketahui bahwa rata-rata skor total peserta tes yang memilih opsi A adalah 20, opsi B sebesar 30, opsi C yang merupakan jawaban benar sebesar 35, sedangkan opsi C sebesar 25. Dari data tersebut, terlihat dengan jelas bahwa rata-rata skor total peserta tes yang memilih opsi C sebagai jawaban benar adalah 35 yang merupakan rata-rata skor total tertinggi dibanding rata-rata skor total peserta tes yang memilih opsi A, B, dan D. Berlawanan dengan hal tersebut, korelasi point-biserial yang negatif menunjukkan bahwa suatu opsi dipilih oleh peserta tes yang memiliki kemampuan lebih rendah dari rata-rata kemampuan peserta tes yang memilih opsi jawaban benar. Dengan kata lain, opsi pengecoh (distraktor) dipilih oleh peserta tes yang diteorikan memiliki kemampuan rendah. Pembahasan secara komprehensif tentang analisis distraktor dapat dilihat dalam literatur (lihat, Testa et al., 2018; Wang, 1998).

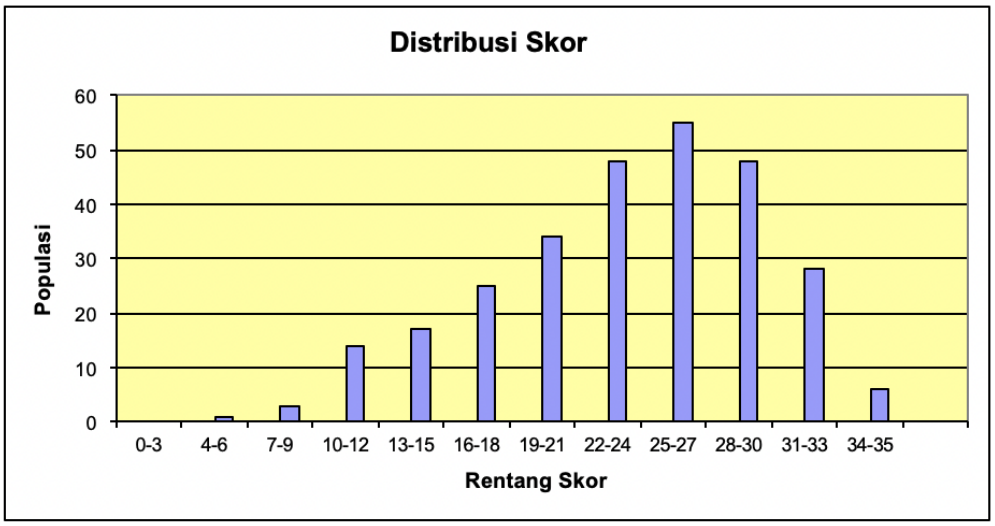
Setelah melakukan analisis *item*, peneliti umumnya akan mengklasifikasikan *item* ke dalam bentuk tinggi rendahnya tingkat kesukaran *item* dan baik tidaknya pembeda *item*. Tingkat kesukaran *item* biasanya dikategorikan ke dalam "mudah", "sedang", dan "sulit". Sedangkan daya pembeda *item* dikelompokkan ke dalam "diterima", "direvisi", dan "ditolak". Dengan menggunakan Quest, peneliti perlu melakukan kategorisasi secara manual, sedangkan dengan Klasika, tabel hasil kategorisasi dapat langsung diperoleh dalam bentuk file Microsoft Excel. Dengan rangkuman kategorisasi tingkat kesukaran *item* dan daya pembeda *item* tersebut, peneliti akan dapat langsung dengan mudah menafsirkan karakteristik *item* yang dianalisis. Tabel klasifikasi *item* dimaksud dapat dilihat seperti pada Gambar 2.

http://journal.uinjkt.ac.id/index.php/jp3i

Rangkuman Tingkat Kesulitan			Rangkuman Daya Pembeda		
No. Soal	**Tingkat Kesulitan**	**Kategori**	**No. Soal**	**Point Biserial**	**Kategori**
1	0.935	Mudah	1	0.356	Diterima
2	0.814	Mudah	2	0.455	Diterima
3	0.613	Sedang	3	0.537	Diterima
4	0.839	Mudah	4	0.342	Diterima
5	0.778	Mudah	5	0.592	Diterima
6	0.857	Mudah	6	0.376	Diterima
7	0.742	Mudah	7	0.414	Diterima
8	0.921	Mudah	8	0.223	Diterima
9	0.792	Mudah	9	0.462	Diterima
10	0.588	Sedang	10	0.365	Diterima
11	0.541	Sedang	11	0.219	Diterima
12	0.584	Sedang	12	0.490	Diterima
13	0.566	Sedang	13	0.342	Diterima
14	0.430	Sedang	14	0.513	Diterima
15	0.789	Mudah	15	0.500	Diterima
16	0.455	Sedang	16	0.276	Diterima
17	0.634	Sedang	17	0.511	Diterima
18	0.642	Sedang	18	0.381	Diterima
19	0.591	Sedang	19	0.546	Diterima
20	0.857	Mudah	20	0.466	Diterima
21	0.821	Mudah	21	0.475	Diterima
22	0.731	Mudah	22	0.536	Diterima
23	0.821	Mudah	23	0.533	Diterima
24	0.609	Sedang	24	0.491	Diterima
25	0.591	Sedang	25	0.220	Diterima
26	0.297	Sulit	26	0.163	Revisi/Cek
27	0.624	Sedang	27	0.438	Diterima
28	0.717	Mudah	28	0.505	Diterima
29	0.563	Sedang	29	0.196	Revisi/Cek
30	0.427	Sedang	30	0.291	Diterima
31	0.437	Sedang	31	0.404	Diterima
32	0.781	Mudah	32	0.512	Diterima
33	0.860	Mudah	33	0.345	Diterima
34	0.774	Mudah	34	0.359	Diterima
35	0.466	Sedang	35	0.315	Diterima

Gambar 2. *Output* Kategorisasi Statistik Butir Klasika

Di samping statistik *item* dan reliabilitas tes, informasi yang seringkali dibutuhkan oleh peneliti adalah distribusi skor peserta tes. Informasi distribusi skor memberikan gambaran tentang sebaran kemampuan peserta tes. Berbeda dari Quest, Klasika memberikan hasil analisis skor peserta tes dalam bentuk grafik pada program Excel. Contoh grafik skor peserta tes yang dihasilkan Klasika dapat dilihat pada Gambar 3.



Gambar 3. *Output* Distribusi Skor Responden

Berdasarkan Gambar 3 dapat dilihat distribusi skor dari 279 peserta tes. Grafik distribusi skor dapat memberikan informasi awal bagi peneliti untuk menilai apakah data yang dimiliki mengikuti distribusi normal atau tidak. Normalitas data sangat penting mengingat banyak uji statistik yang mengasumsikan bahwa data berdistribusi normal (lihat, Price, 2017; Raykov & Marcoulides, 2011). Dari Gambar 3, peneliti juga dapat dengan mudah mengetahui skor tertinggi dan terendah, serta skor dengan persentase tertinggi. Untuk data yang sama, Quest menghasilkan informasi dalam bentuk tabel dengan kolom sebanyak jumlah *item* dan baris sebanyak jumlah peserta tes, sehingga untuk dapat menghasilkan Gambar 3, diperlukan pemindahan data terlebih dahulu kedalam program Excel.

Penutup

Berdasarkan temuan yang telah dipaparkan, dapat disimpulkan bahwa Klasika merupakan aplikasi yang dapat dengan mudah digunakan untuk analisis *item* dan tes. Statistik yang dihasilkan Klasika menunjukkan proses komputasi yang terpercaya sebagaimana yang dihasilkan Quest, *software* analisis psikometrik yang telah secara luas digunakan. Statistik yang dihasilkan Klasika berbasis CTT mencakup tingkat kesukaran *item*, daya pembeda *item*, sebaran jawaban, reliabilitas tes, dan *standard error of measurement*. Di samping statistik *item* dan tes, Klasika juga memberikan hasil analisis skor pserta tes dalam bentuk grafik yang mudah difahami. Beberapa kemudahan ditawarkan Klasika dalam pelaporan hasil analisis *item* dan tes. Klasika yang merupakan program berbasis GUI memudahkan pengguna untuk melakukan analisis data tanpa *syntax* ataupun *code*. Pengembangan Klasika diharapkan dapat menginisasi perkembangan program lain, khususnya program analisis psikometrik dengan pendekatan teori tes modern.

Daftar Pustaka

Adams, R. J. (2005). Reliability as a measurement of design effect. *Studies in Educational Evaluation,* 31(2-3), 162-172. https://doi.org/10.1016/j.stueduc.2005.05.008.

Adams, R. J., & Khoo, S. T. (1993). *Quest: the interactive test analysis system*. Australian Council for Educational Research.

Allen, M. J. & Yen, W. M. (1979). *Introduction to measurement theory.* Brooks/Cole Publishing Company.

Andrich, D. (2004). Controversy and the Rasch model: A characteristic of incompatible paradigms? *Medical Care*, 42(1), 7-16. https://doi.org/10.1097/01.mlr.0000103528.48582.7c.

Andrich, D., & Marais, I. (2019). *A course in Rasch Measurement Theory: Measuring in the educational, social and health sciences.* Springer.

AP2TPI. (2018). *Kurikulum inti program studi psikologi jenjang sarjana.* Asosiasi Penyelenggara Pendidikan Tinggi Psikologi Indonesia (AP2TPI). https://ap2tpi.or.id/wp-content/uploads/2019/05/SK-AP2TPI-Perubahan-Kurikulum-Inti-Program-Studi-Sarjana-Final-sdh-ttd-22-november-2018.pdf.

Bazaldua, D. A. L., Lee, Y. S., Keller, B., & Fellers, L. (2017). Assessing the performance of classical test theory item discrimination estimators in monte carlo simulations. *Asia Pacific Education Review,* 18, 585-598. https://doi.org/10.1007/s12564-017-9507-4.

Berk, R. A., & Griesemer, H. A. (1976). Iteman: an item analysis program for tests, questionnaires, and scales. *Educational and Psychological Measurement,* 36(1), 189-191. https://doi.org/10.1177/001316447603600122.

Brennan, R. L. (2011). Generalizability theory and classical test theory. *Applied Measurement in Education,* 24(1), 1-21. https://doi.org/10.1080/08957347.2011.532417.

Crocker, L., & Algina, J. (2008). *Introduction to classical and modern test theory* (3rd edition.). Cengage Learning.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, 16(3), 297-334. https://doi.org/10.1007/BF02310555.

Cronbach, L. J. (1960). *Essentials of psychological testing* (2nd ed.). Harper & Row.

Ebel, R. L., & Frisbie, D. A. (1991). *Essentials of educational measurement* (5th edition.). Prentice Hall.

Loewenthal, K. M. (2001). *An introduction to psychological tests and scales* (2nd ed.). Psychology Press.

Lord, F. M., & Novick, M. E. (1968). *Statistical theories of mental test scores.* Addison-Wesley.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems.* Lawrence Erlbaum Associates.

Martins, P. S. R., Barbosa-Pereira, D., Valgas-Costa, M., & Mansur-Alves, M. (2020). Item analysis of the Child Neuropsychological Assessment Test (TENI): Classical test theory and item response theory. *Applied Neuropsychology: Child.* https://doi.org/10.1080/21622965.2020.1846128.

Novick, M. R. (1966). The axioms and principal results of classical test theory. *Journal of Mathematical Psychology,* 3(1), 1–18. https://doi.org/10.1016/0022-2496(66)90002-2.

Price, L. R. (2017). *Psychometric methods: Theory into practice.* The Guilford Press.

Raykov, T., & Marcoulides, G. (2011). *Introduction to psychometric theory.* Routledge.

Raykov, T., & Marcoulides, G. A. (2016). On the relationship between classical test theory and item response theory: From one to the other and back. *Educational and Psychological Measurement,* 76(2), 325–338. https://doi.org/10.1177/0013164415576958.

Testa, S., Toscano, A., & Rosato, R. (2018). Distractor efficiency in an item pool for a Statistic Classroom Exam: Assessing its relation with item cognitive level classified according to Bloom's taxonomy. *Frontiers in Psychology,* 9: 1585. https://doi.org/10.3389/fpsyg.2018.01585.

Traub, R. E. (1997). Classical test theory in historical perspective. *Educational Measurement: Issues and Practice,* 16(4), 8-14. https://doi.org/10.1111/j.1745-3992.1997.tb00603.x.

Wang, W. C. (1998). Rasch analysis of distractors in multiple-choice items. *Journal of Outcome Measurement,* 2(1), 43-65.

Wilkinson, L., & Task Force on Statistical Inference, American Psychological Association, Science Directorate. (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist*, 54(8), 594–604. https://doi.org/10.1037/0003-066X.54.8.594.

Wright, B. D. (1997). A history of social science measurement. *Educational Measurement: Issues and Practice,* 16(4), 33–45, 52. https://doi.org/10.1111/j.1745-3992.1997.tb00606.x.